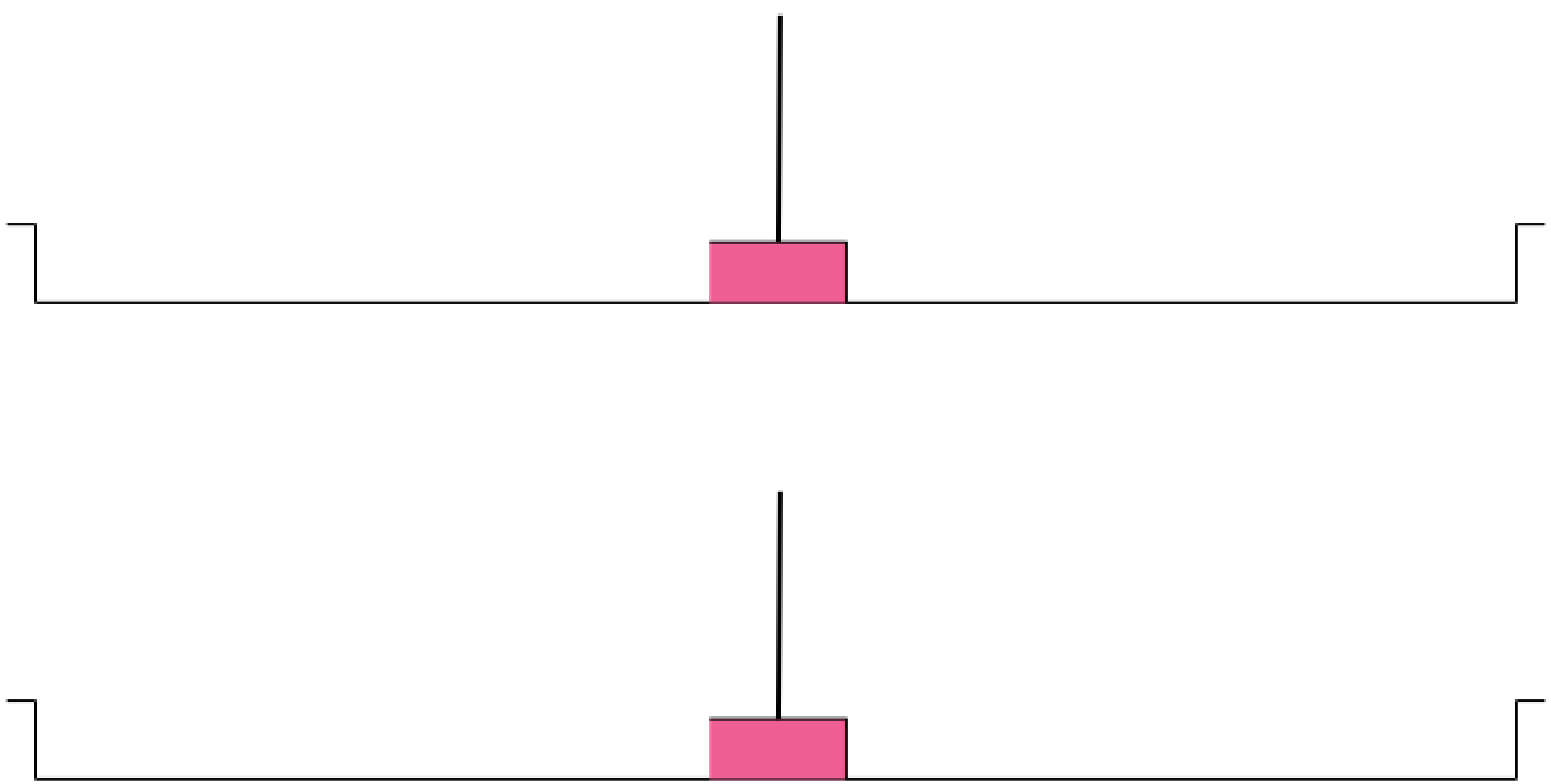
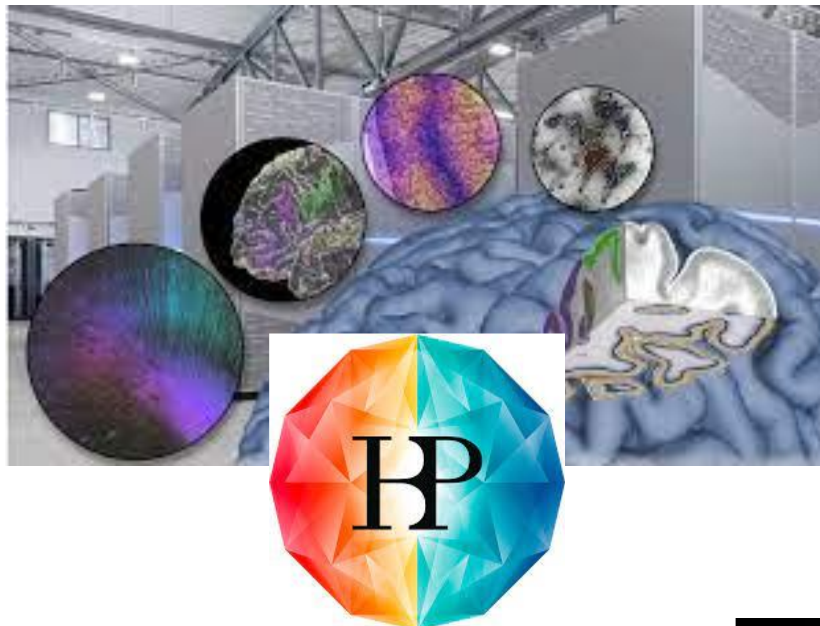
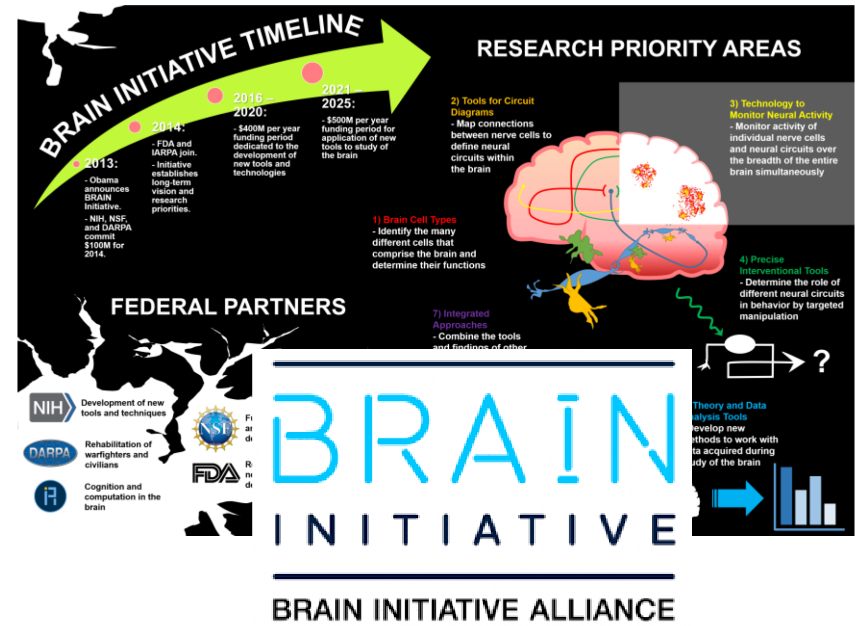
Seeking entropy

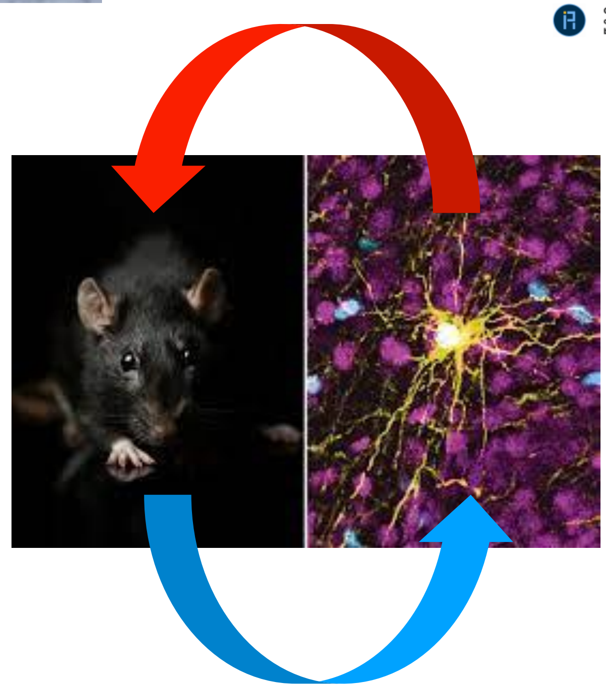
Complex behavior from intrinsic motivation to occupy action-state path space



Approaches for Brain and Behavior







Getty Images

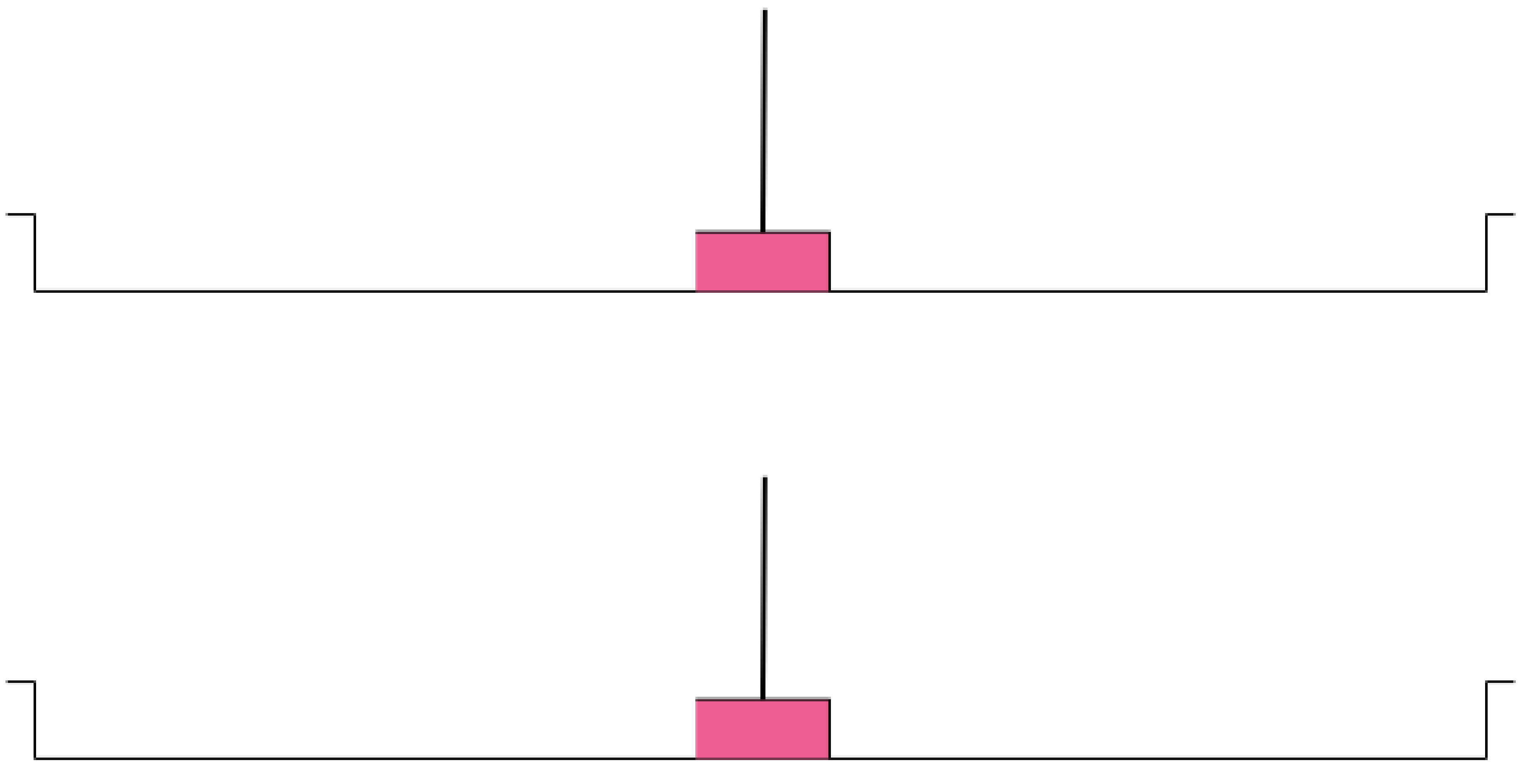
Bottom-Up Approach: from synapses neurons and circuits to emerging behaviors

- emphasis on data collection and simulation, but not on theory
- no emphasis on behavior

Proposal. **Top-Down approach**: from behavior to synapses, neurons and circuits

Seeking entropy

Complex behavior from intrinsic motivation to occupy action-state path space



Life is (in) motion

- Natural tendency to move, explore, and interact with the environment with curiosity
- 7-12m babies babble and motor-babble
- Infants explore with curiosity

Why?

- Movement and curiosity → learning
- Learning → higher future rewards

Standard Hypothesis: Animals are reward/utility maximizers (von Neumann, Sutton & Barto, Kahneman)

Are we utility maximizers?

Reward function?

(Getty Images)

BBC News: a robot escapes the hotel room where it was cleaning

w_1

w_2

w_3

The goal: occupy action-state path space

- We abandon the idea that maximizing utility is the goal and that moving is the mean to achieve the goal

- We adopt the opposite view: moving around is the goal, and external rewards are just means

The goal: occupy action-state path space

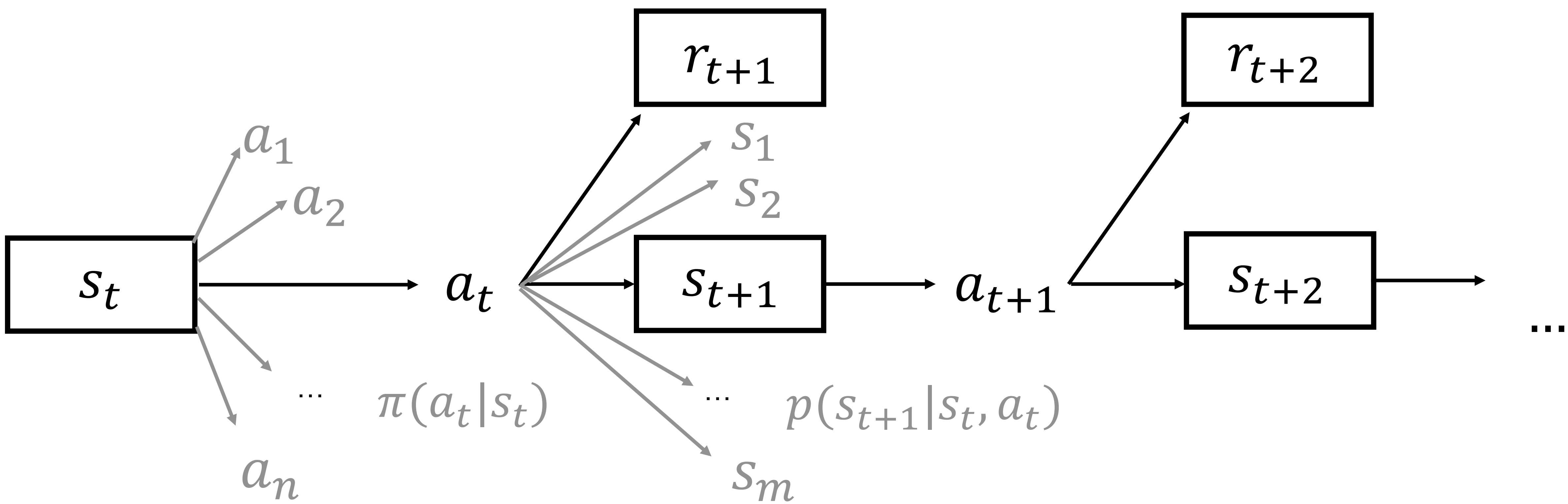
Principle: *agents maximize occupancy of action-state path space*

Max Occupancy Principle (MOP)

Ramírez-Ruiz, Grytskyy, Moreno-Bote, arXiv, 2022

- These agents will be naturally “curious” and “explorative”
- They will seek reward only to occupy more space
- Survival instinct (will avoid terminal states with no actions available)
- Preference of freedom
- They will occupy internal states → variability in neural activity

Modeling behavior with MDPs



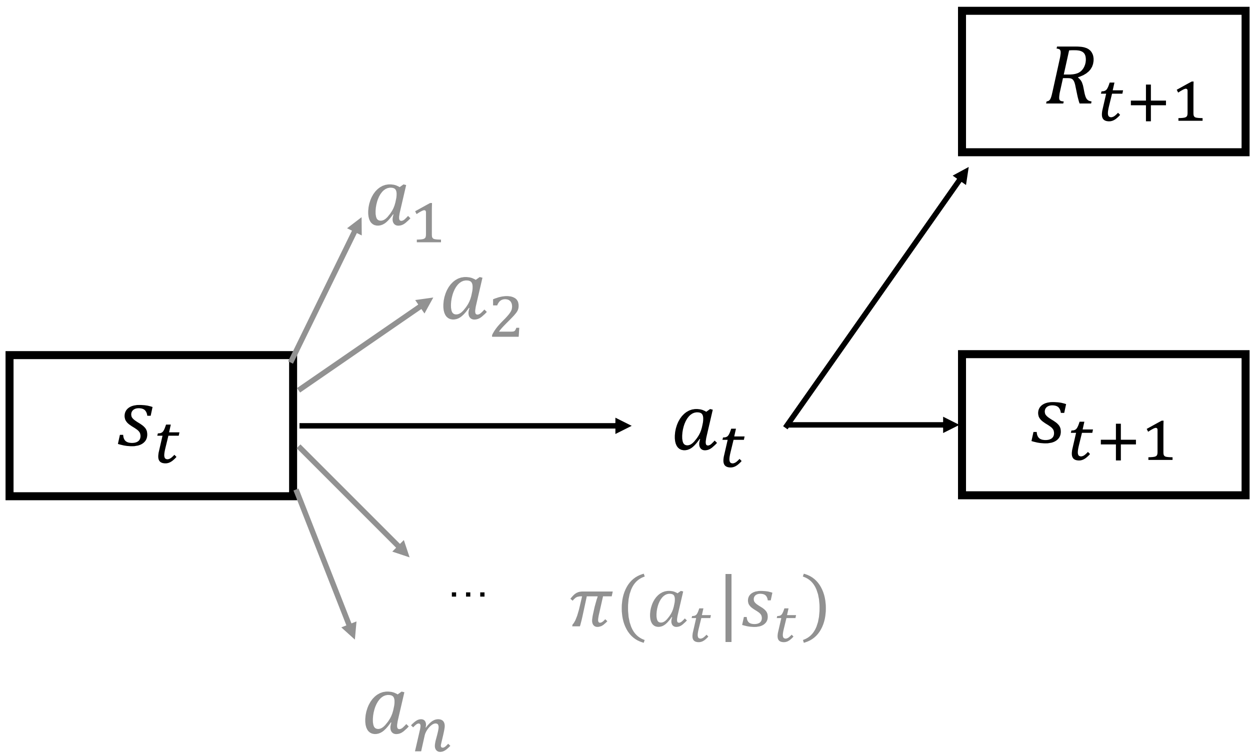
$\pi(a|s_t)$ is the policy: probability of performing an action given current state

$p(s_{t+1}|s_t, a_t)$ is the world model: a (stochastic) mapping between states, given actions

r_{t+1} is the reward: a policy-independent, action-state signal, $r(s, a)$

$V_\pi(s) \equiv \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right]$ is the state value under the policy

Entropy as a measure of action-state occupancy



Deterministic policy: $\pi(a|s_t) = 1$ for only one action a

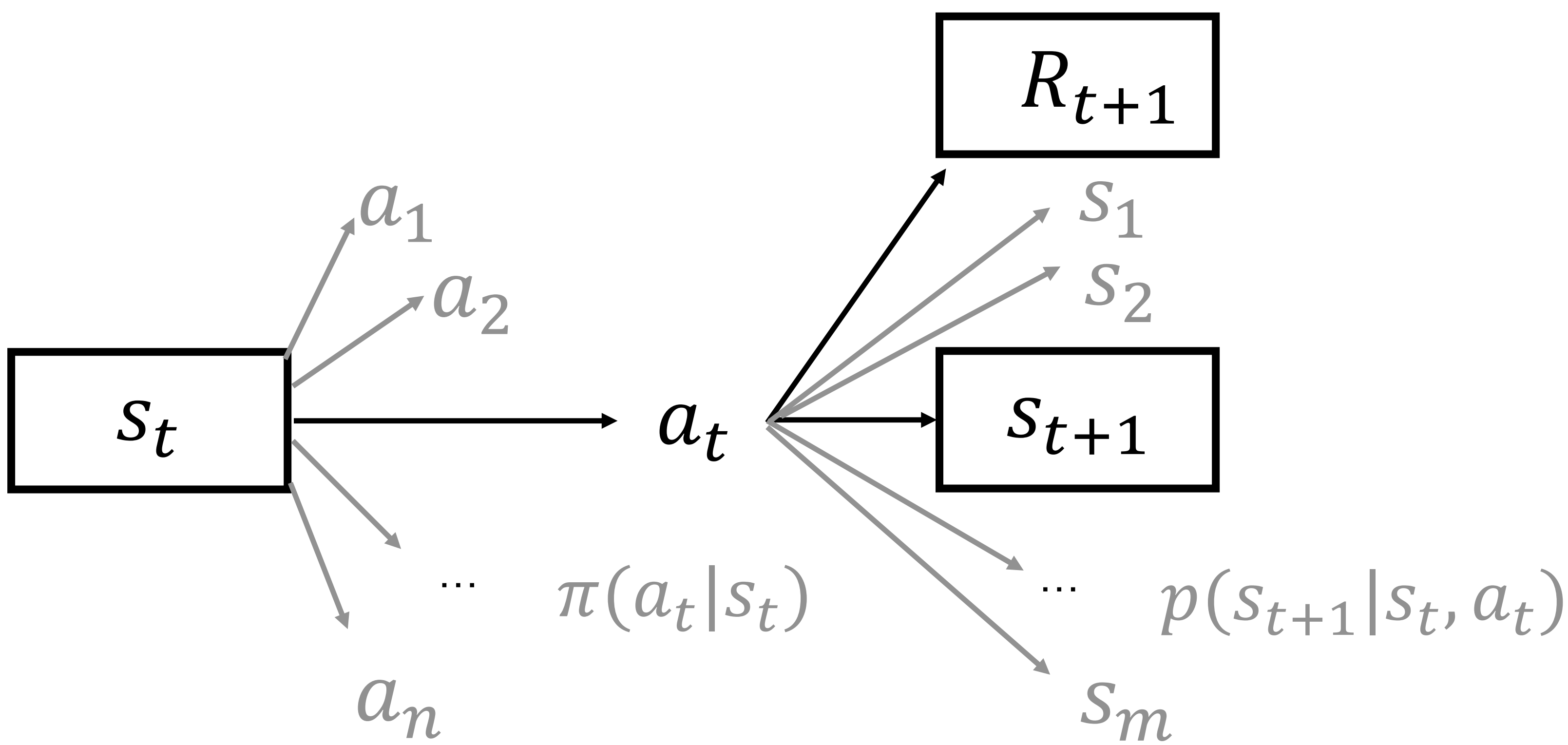
Occupancy gain is $R_{t+1} = 0$

Occupancy gain: $R_{t+1} = -\ln\big(\pi(a_t|s_t)\big)$, a form of intrinsic reward; policy-dependent

Action occupancy is its expectation (= policy entropy),

$$\mathbb{E}_\pi[R_{t+1}|s_0 = s_t] = H(A|s_t) = -\sum_a \pi(a|s_t)\ln(\pi(a|s_t))$$

Entropy as a measure of action-state occupancy



The joint probability of an action-state (a_t, s_{t+1}) is $\pi(a_t|s_t)p(s_{t+1}|s_t, a_t)$

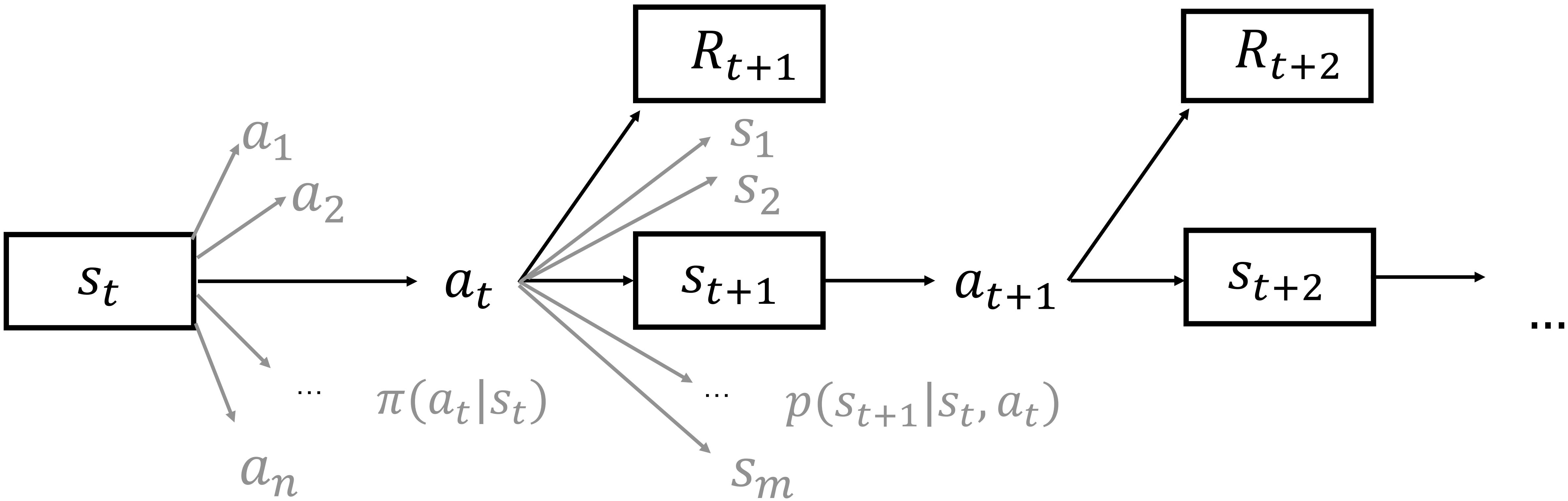
For deterministic policy and environments, only one action-state (a_t, s_{t+1}) is available

Thus, occupancy gain of that action-state is $R_{t+1} = 0$

Action-state occupancy gain: $R_{t+1} = -\ln(\pi(a_t|s_t)p(s_{t+1}|s_t, a_t))$

Action-state occupancy: $\mathbb{E}_\pi[R_{t+1}|s_0 = s_t] = H(A|s_t) + \mathbb{E}_{s',a_t|\pi}[H(S'|s_t, a_t)|s_0 = s_t]$

Cumulative entropy measures action-state path occupancy



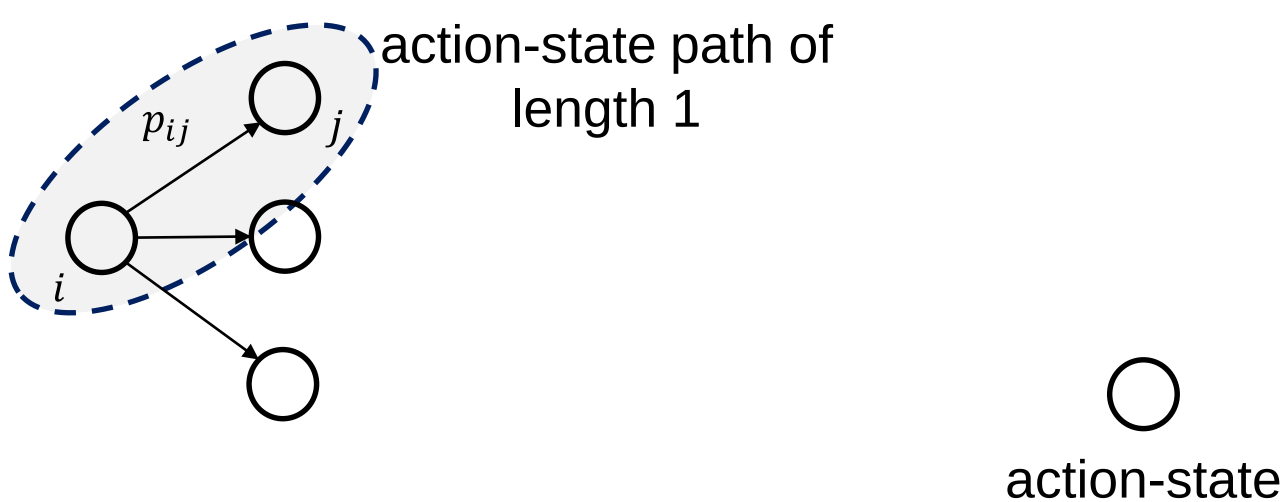
$$V_\pi(s) \equiv \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left(\underbrace{H(A|s_t)}_{\text{Action Occupancy}} + \underbrace{H(S'|s_t, a_t)}_{\text{State Occupancy}}\right) | s_0 = s\right]$$

Cumulative future action-state entropy is the only measure with the *additive property:* "occupancy of a path of any length is the sum of expected occupancies of any of its sub-paths"

Desired properties of action-state path occupancy

1. Occupancy gain by performing a transition from action-state i to j is a function $C(p_{ij})$

2. Performing a low probability transition leads to a higher occupancy gain

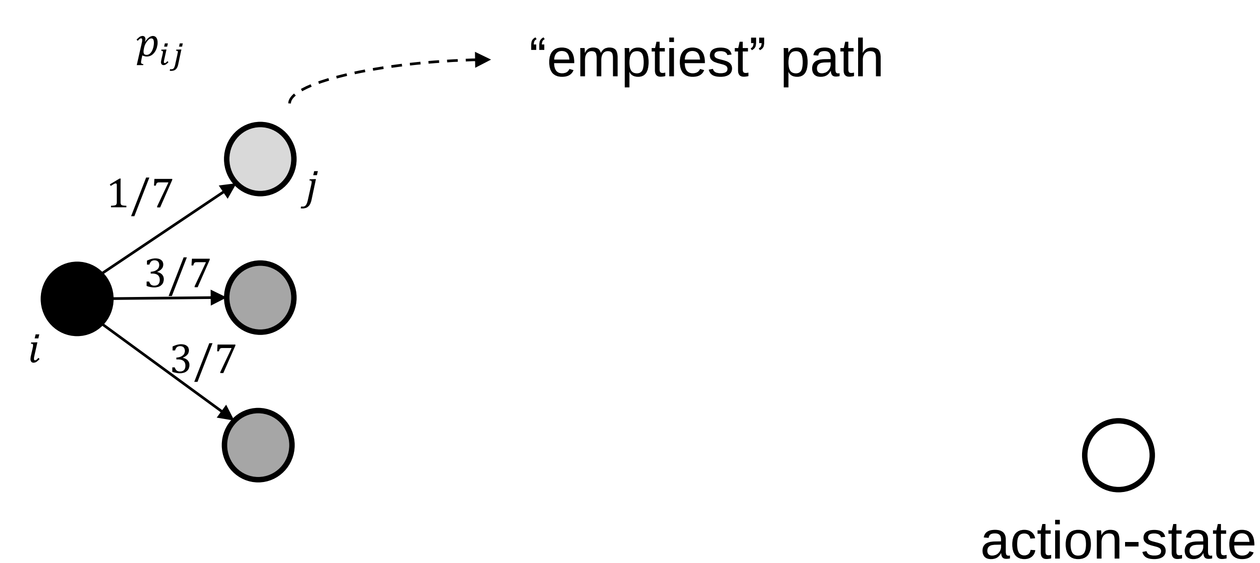


Desired properties of action-state path occupancy

1. Occupancy gain by performing a transition from action-state i to j is a function $C(p_{ij})$

2. Performing a low probability transition leads to a higher occupancy gain

3. $C(p)$ is a smooth function



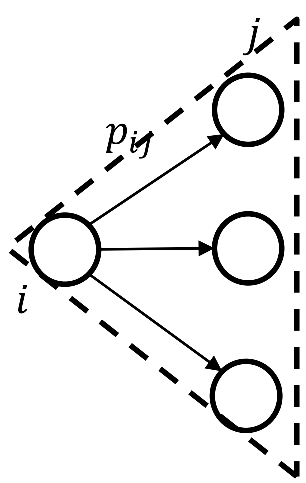
Desired properties of action-state path occupancy

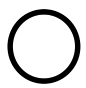
1. Occupancy gain by performing a transition from action-state i to j is a function $C(p_{ij})$

2. Performing a low probability transition leads to a higher occupancy gain

3. $C(p)$ is a smooth function

Definition: occupancy of one-step paths is $C_i^{(1)} \equiv \sum_j p_{ij} C(p_{ij})$





action-state

Desired properties of action-state path occupancy

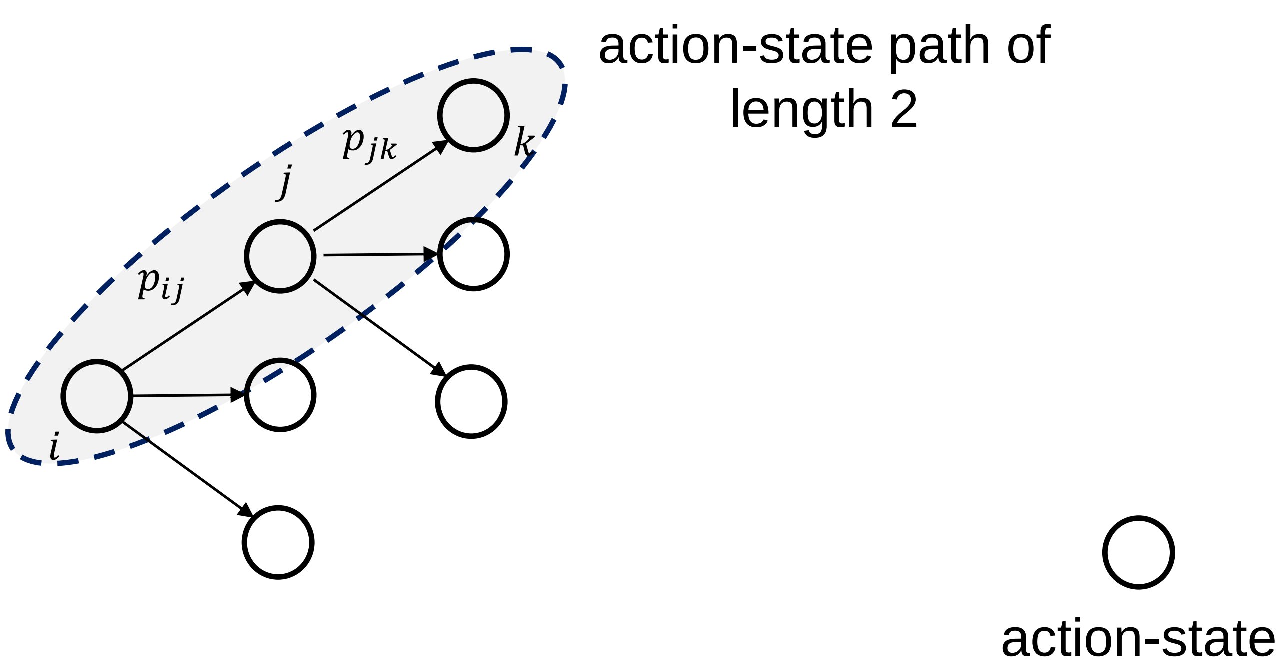
1. Occupancy gain by performing a transition from action-state i to j is a function $C(p_{ij})$

2. Performing a low probability transition leads to a higher occupancy gain

3. $C(p)$ is a smooth function

Definition: occupancy of one-step paths is $C_i^{(1)} \equiv \sum_j p_{ij} C(p_{ij})$

4. Additive property: $C_i^{(2)} \equiv \sum_{jk} p_{ij} p_{jk} C\left(p_{ij} p_{jk}\right) = C_i^{(1)} + \sum_j p_{ij} C_j^{(1)}$



Desired properties of action-state path occupancy

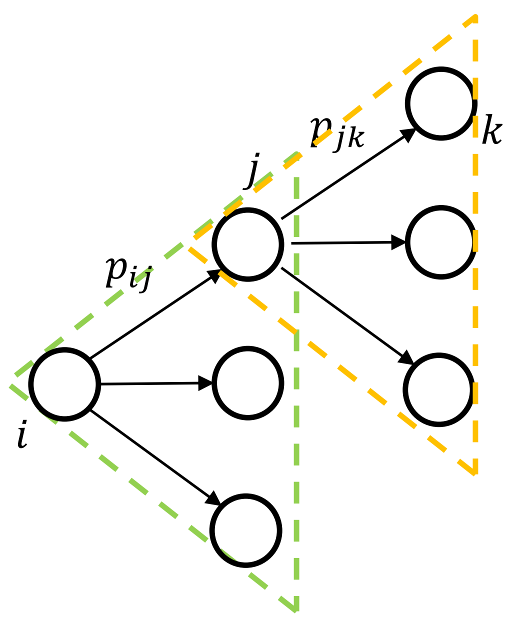
1. Occupancy gain by performing a transition from action-state i to j is a function $C(p_{ij})$

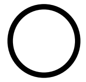
2. Performing a low probability transition leads to a higher occupancy gain

3. $C(p)$ is a smooth function

Definition: occupancy of one-step paths is $C_i^{(1)} \equiv \sum_j p_{ij} C(p_{ij})$

4. Additive property: $C_i^{(2)} \equiv \sum_{jk} p_{ij} p_{jk} C\left(p_{ij} p_{jk}\right) = C_i^{(1)} + \sum_j p_{ij} C_j^{(1)}$



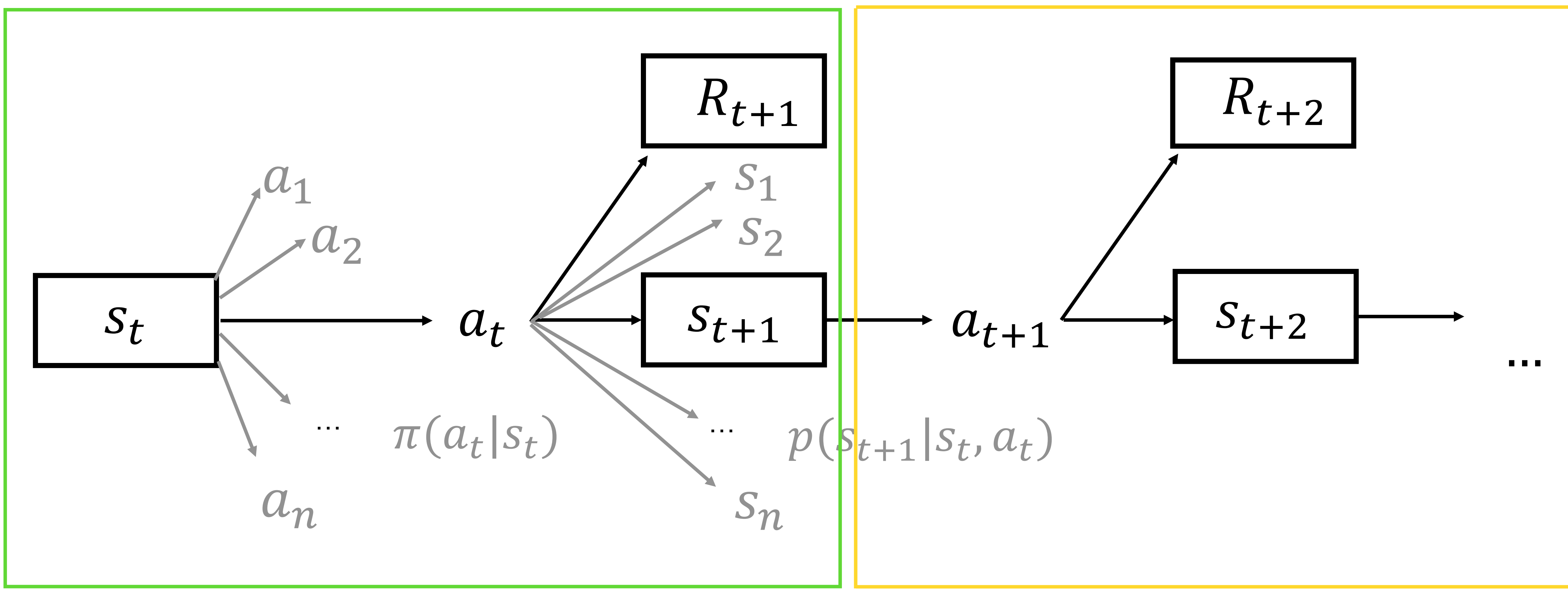


$C(p)$ must be $-\log p$

Additive property: $C_i^{(2)} \equiv \sum_{jk} p_{ij} p_{jk} C(p_{ij} p_{jk}) = C_i^{(1)} + \sum_j p_{ij} C_j^{(1)}$

$$
\begin{aligned}
C_i^{(2)} &\equiv \sum_{jk} p_{ij} p_{jk} C(p_{ij} p_{jk}) \\
&= -\sum_{jk} p_{ij} p_{jk} \log(p_{ij} p_{jk}) \\
&= -\sum_{jk} p_{ij} p_{jk} \log(p_{ij}) - \sum_{jk} p_{ij} p_{jk} \log(p_{jk}) \\
&= -\sum_j p_{ij} \log(p_{ij}) - \sum_j p_{ij} \sum_k p_{jk} \log(p_{jk}) \\
&= C_i^{(1)} + \sum_j p_{ij} C_j^{(1)}
\end{aligned}
$$

Cumulative entropy measures of action-state occupancy



$$V_\pi(s) \equiv \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \, (\alpha H(A|s_t) + \beta H(S'|s_t, a_t))|s_0 = s\right]$$

Bellman eq: $V_\pi(s) = \alpha H(A|s_t) + \beta \mathbb{E}_{a,s'|s,\pi}[H(S'|s,a)|s] + \gamma \, \mathbb{E}_{s'|s,\pi}[V_\pi(s')\,]$

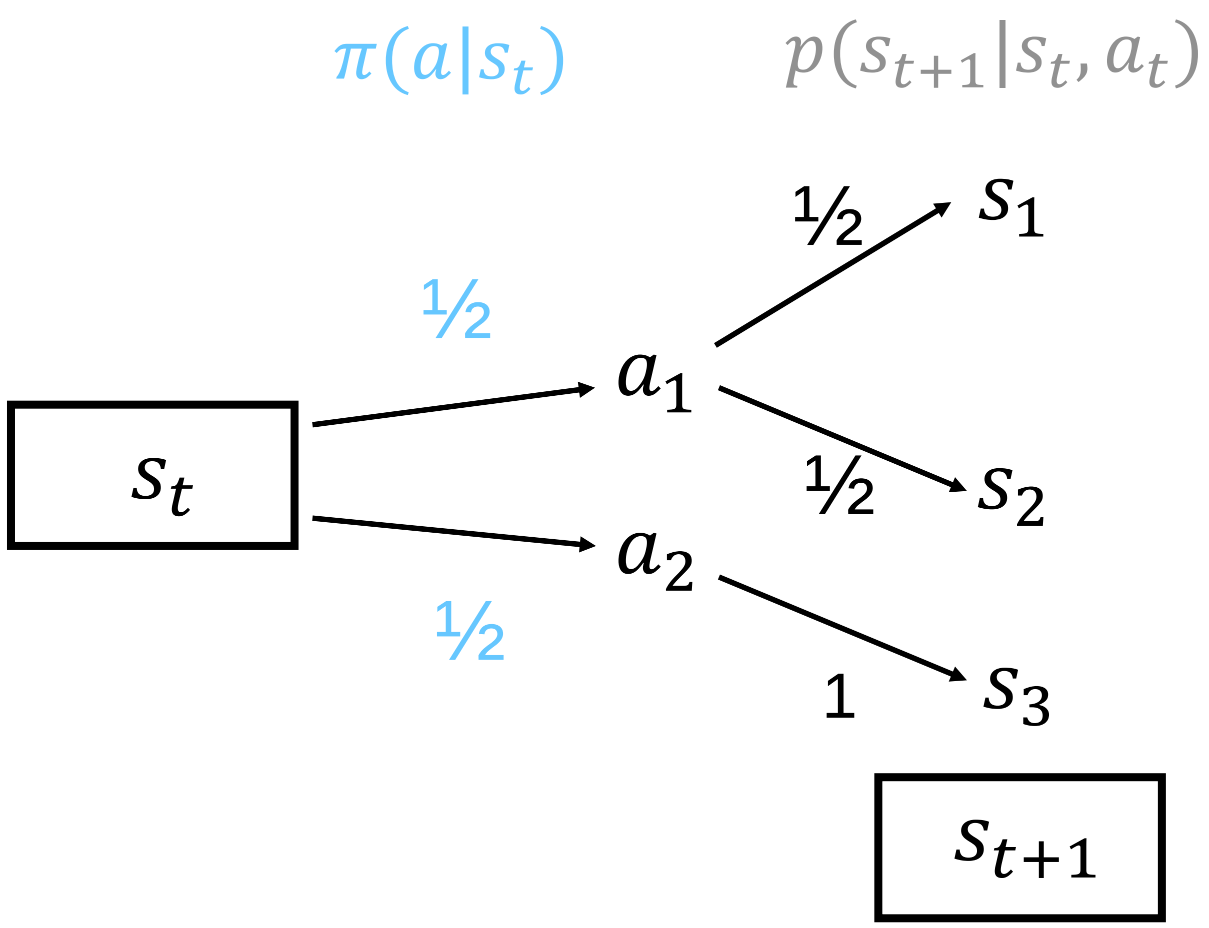
Immediate Occupancy

Future Occupancy

Optimal value: $V^*(s) = \ln(\sum_a \exp\left(\alpha^{-1}\beta H(S'|s,a) + \gamma \, \alpha^{-1}\mathbb{E}_{s'|s,a}[V^*(s')]\right)\,)$

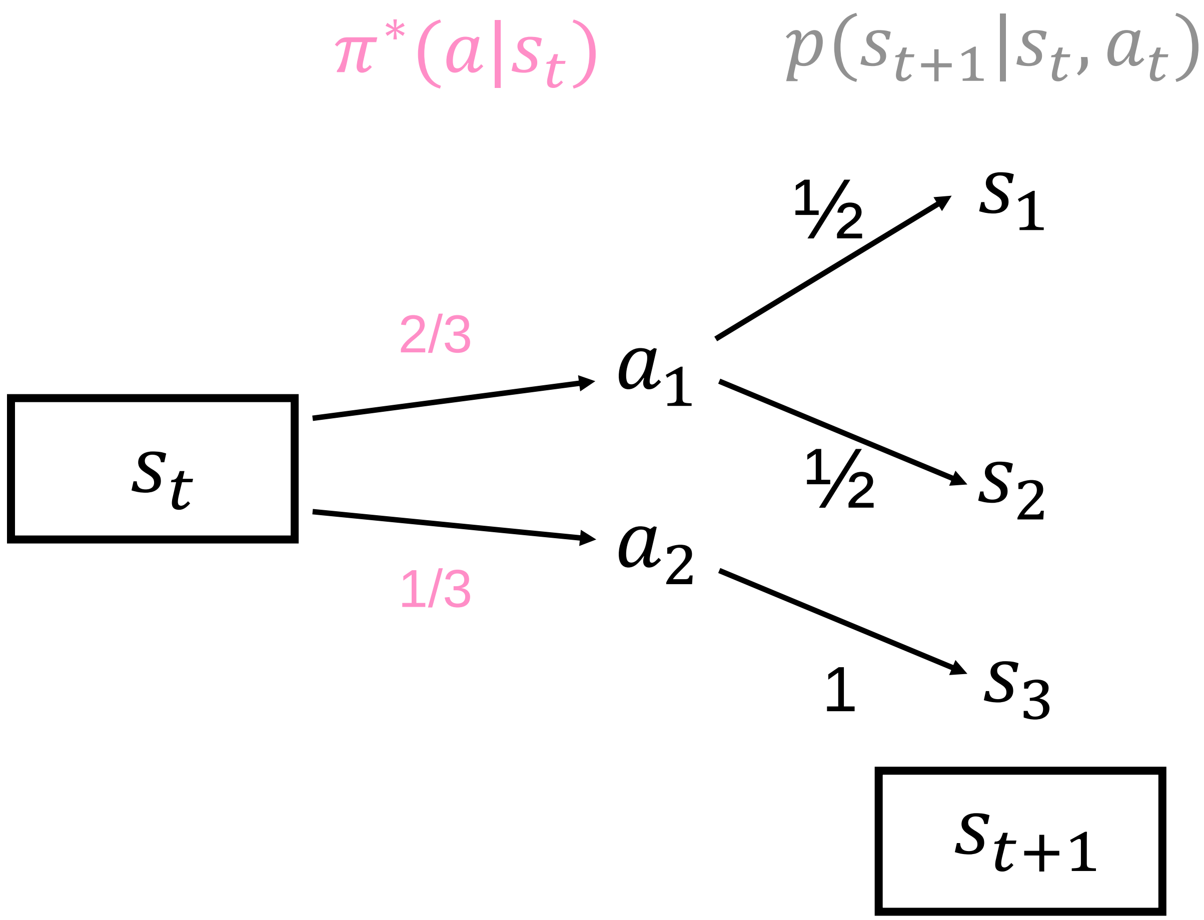
Optimal policy: $\pi^*(a|s) \propto \exp(\alpha^{-1}\beta H(S'\,|s,a) + \gamma \, \alpha^{-1}\mathbb{E}_{s'|s,a}[V^*(s')])$

Example (1 step forward)



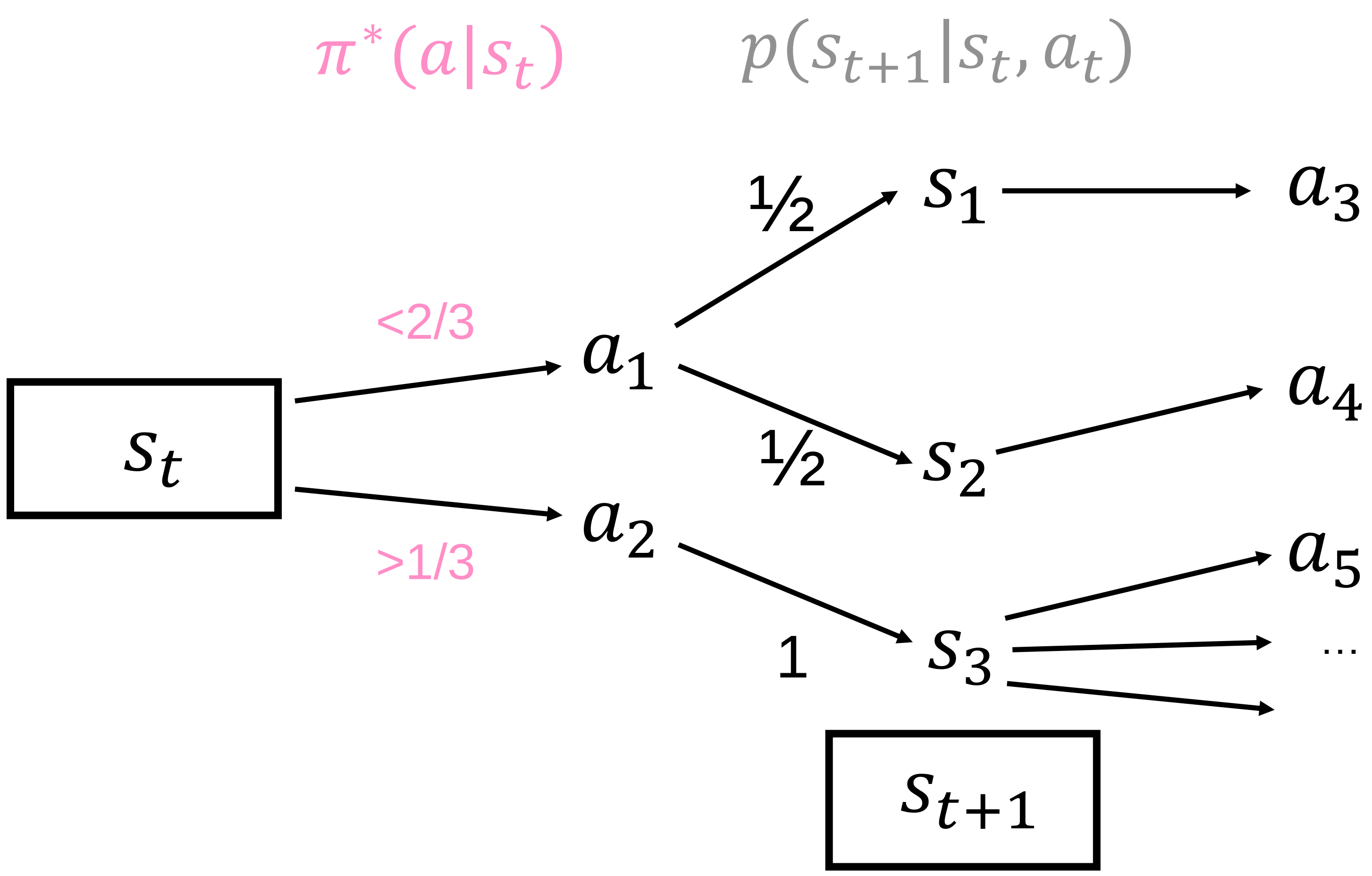
$\pi(a|s_t)$ is suboptimal!

Example (1 step forward)

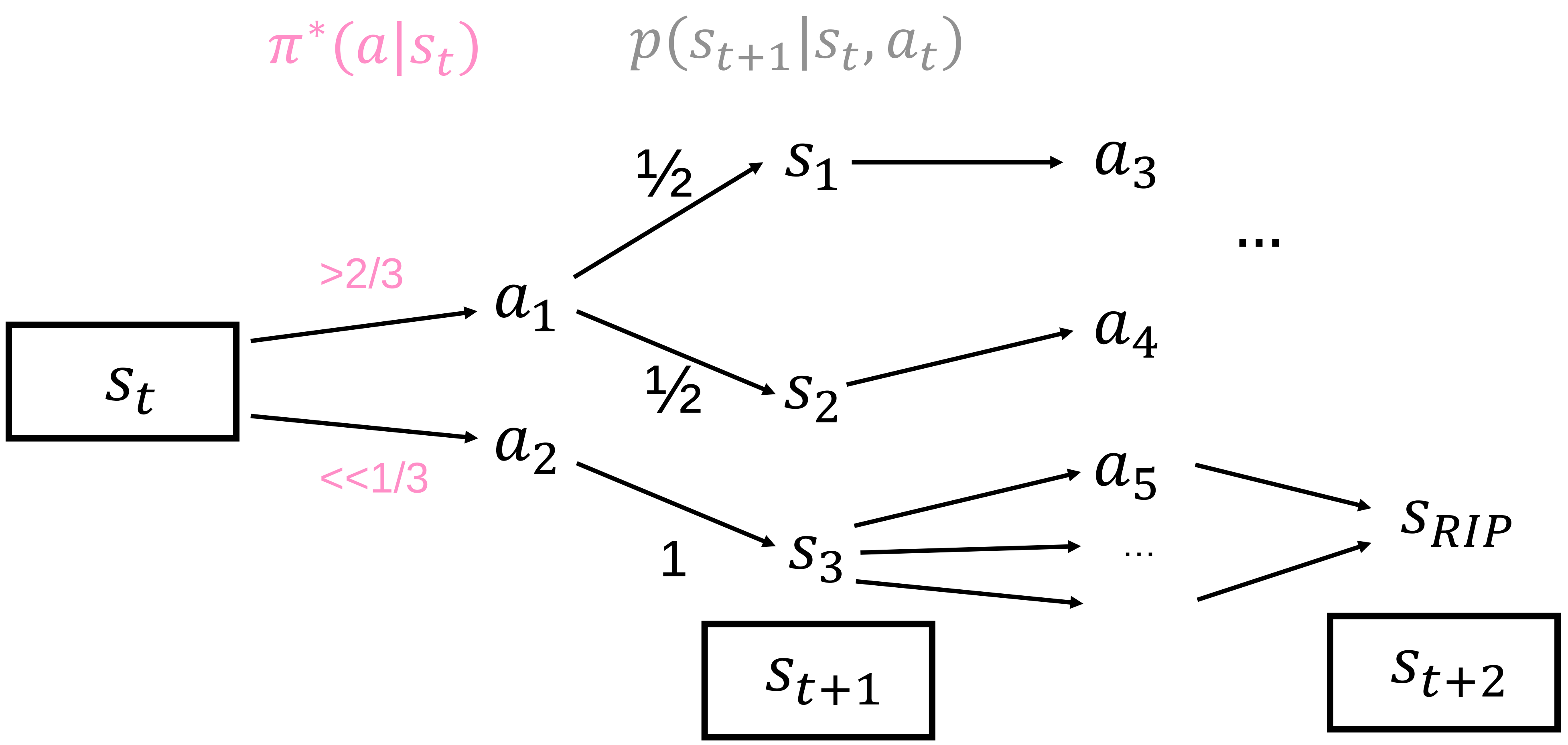


$\pi^*(a|s_t)$ is optimal!

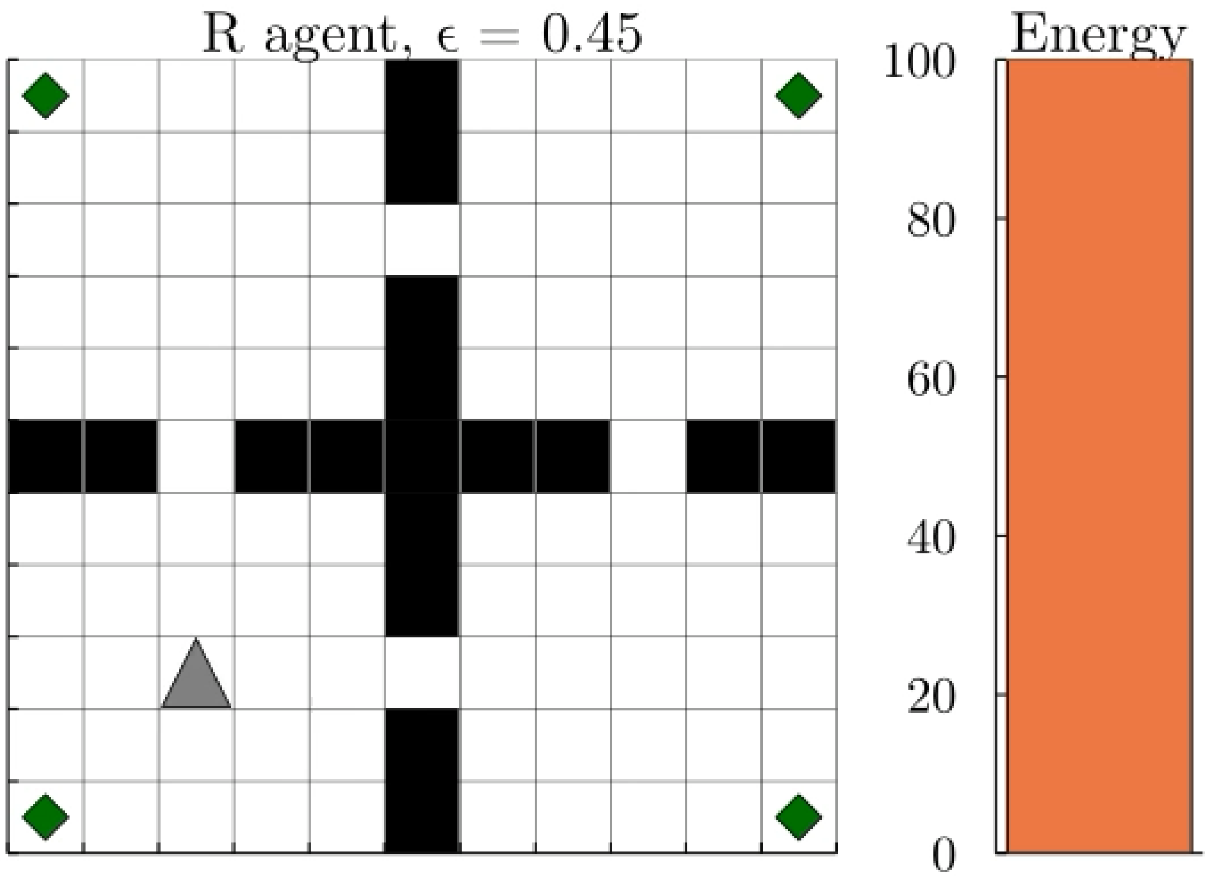
Example (2 steps forward)

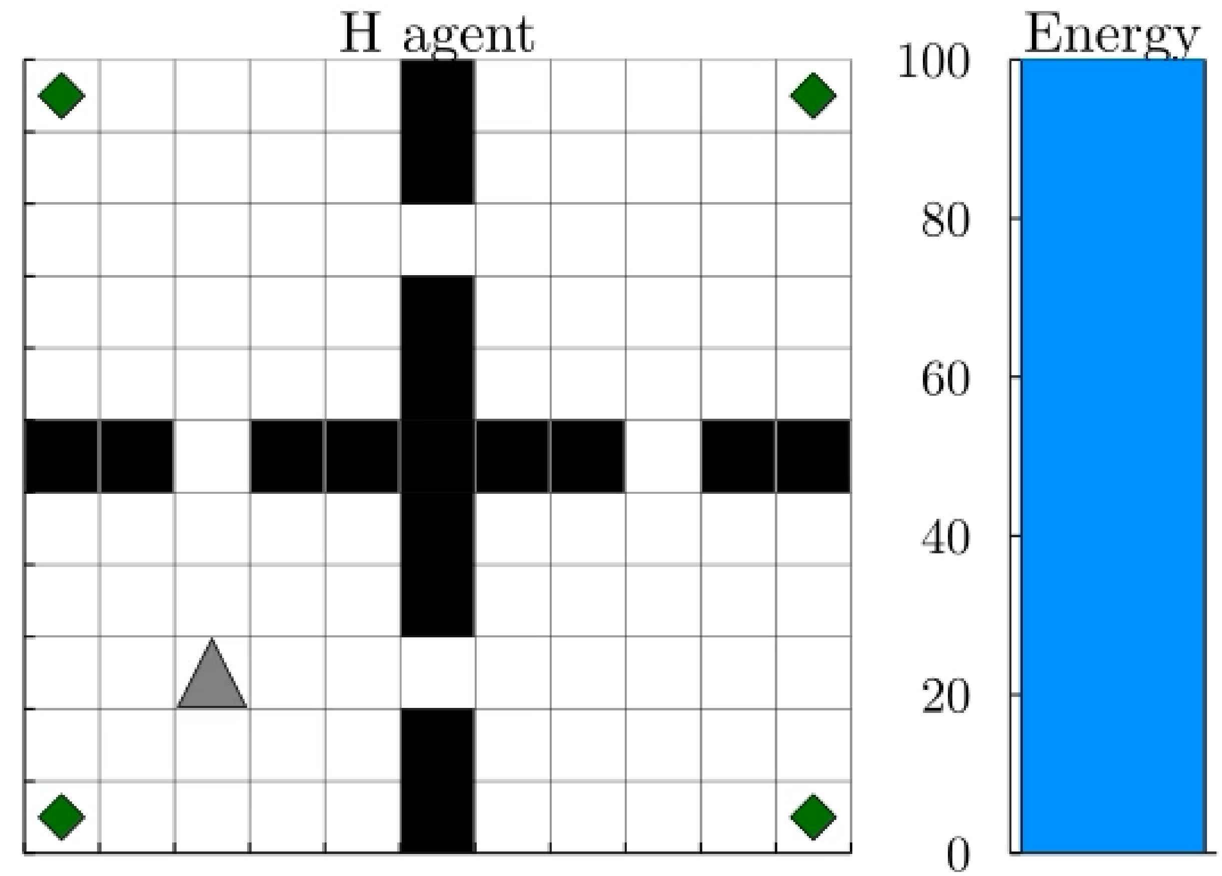


Example (2 steps forward)



Occupancy vs reward maximization





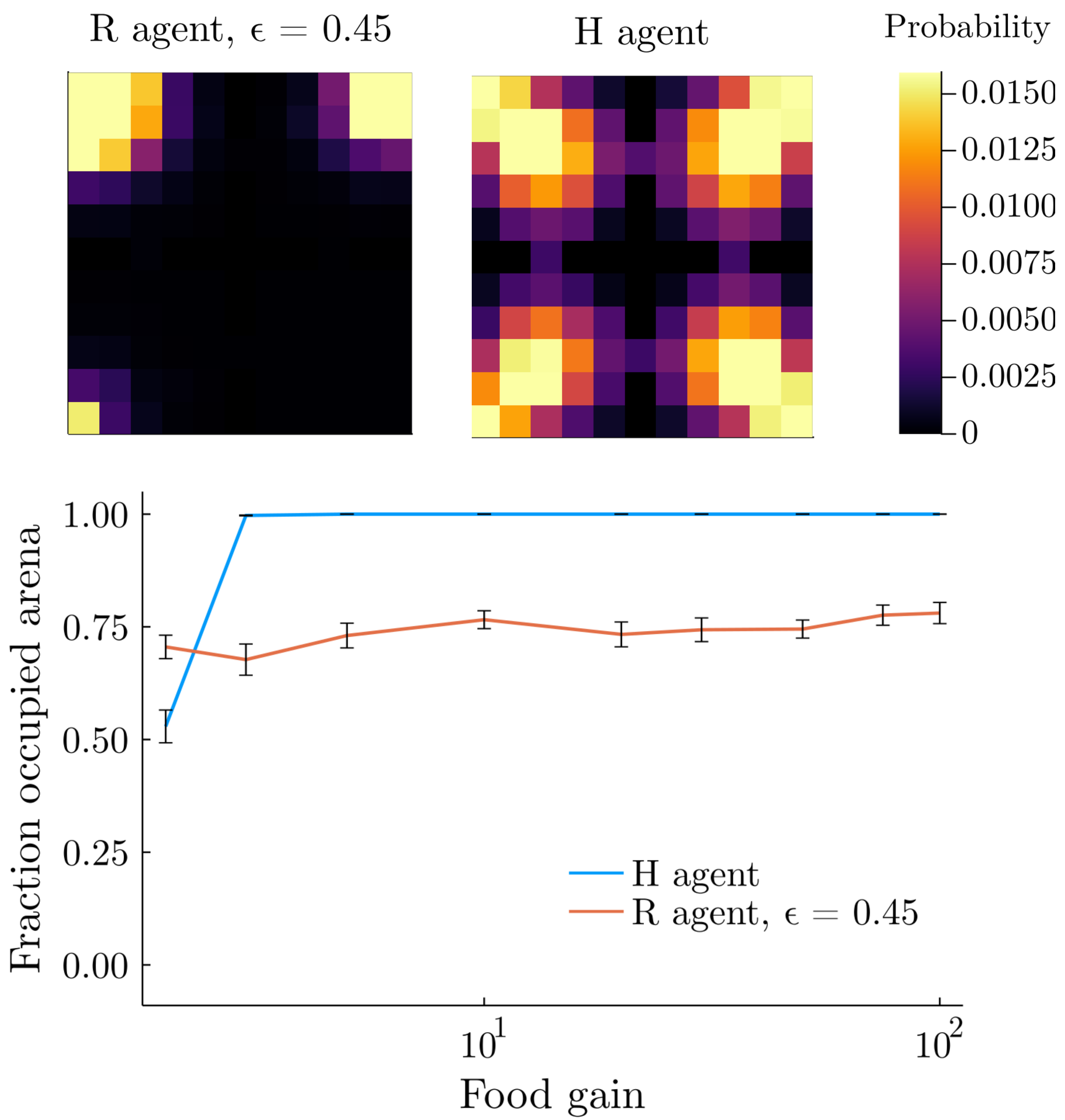
$R_\pi(s,a) = r(s,a)$

$R_\pi(s,a) = -\alpha \ln \pi(a|s) - \beta \ln p(s'|s,a)$

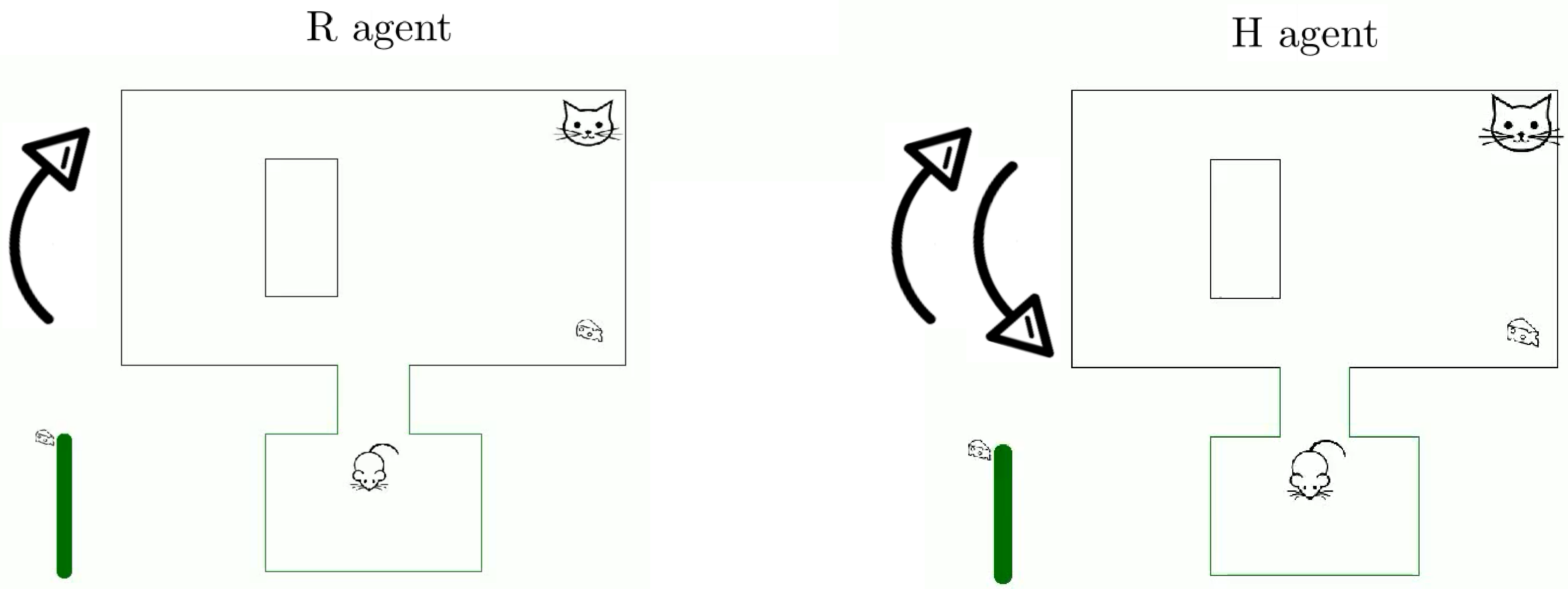
$(\alpha,\beta) = (1,1)$

$s = (x, y, E)$
$\Delta E_{food} = 10$
$\Delta E_{living} = -1$
terminal states: $E = 0$

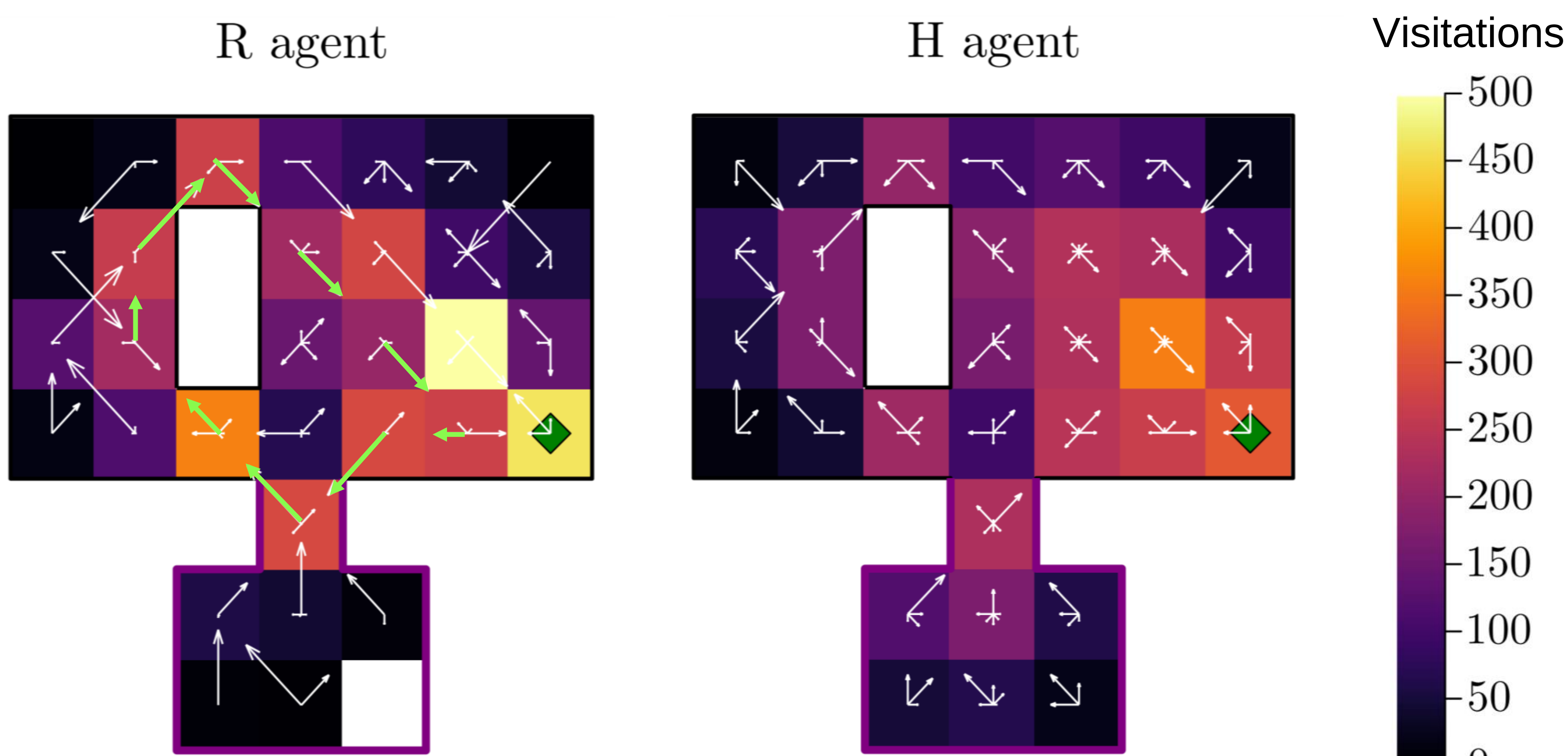
Occupancy vs reward maximization

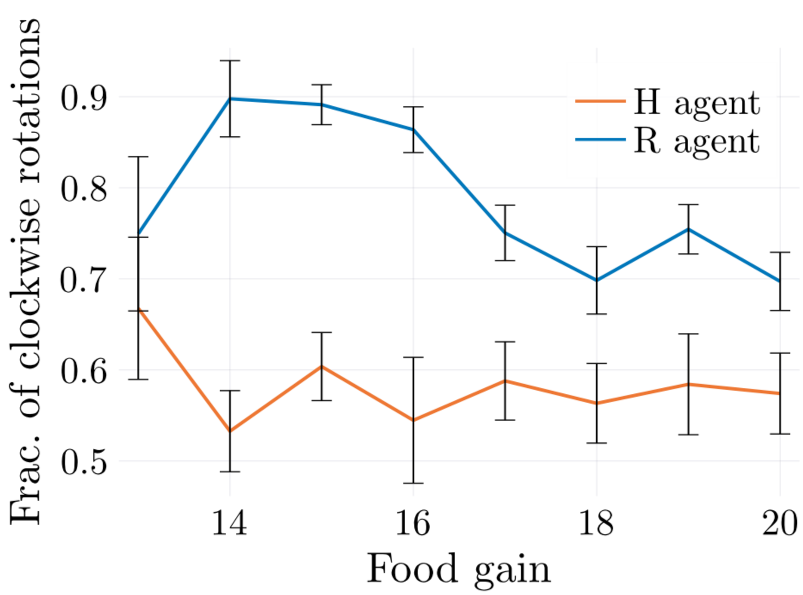


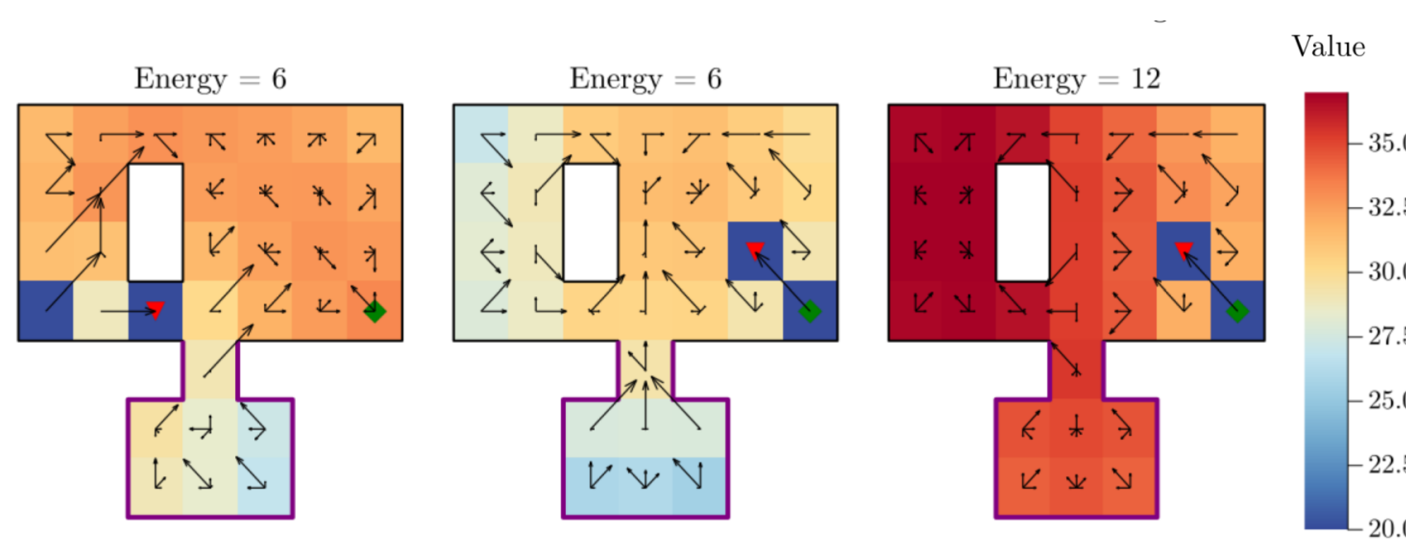
Complex behaviors in a prey-predator example



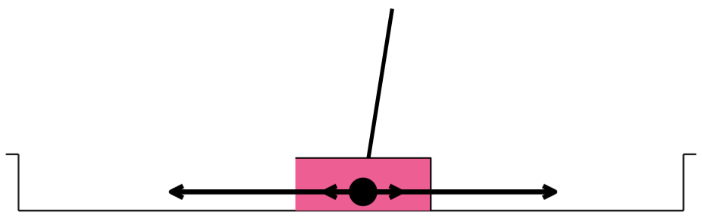
Complex behaviors in a prey-predator example





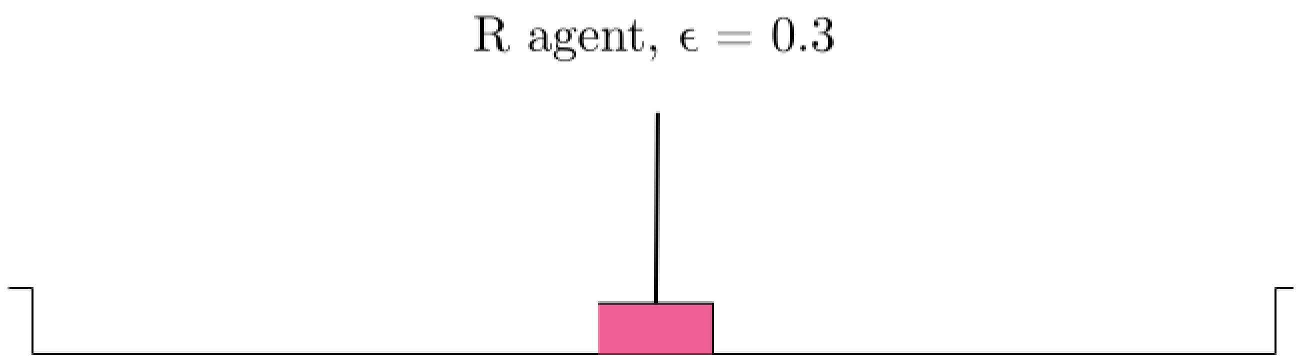


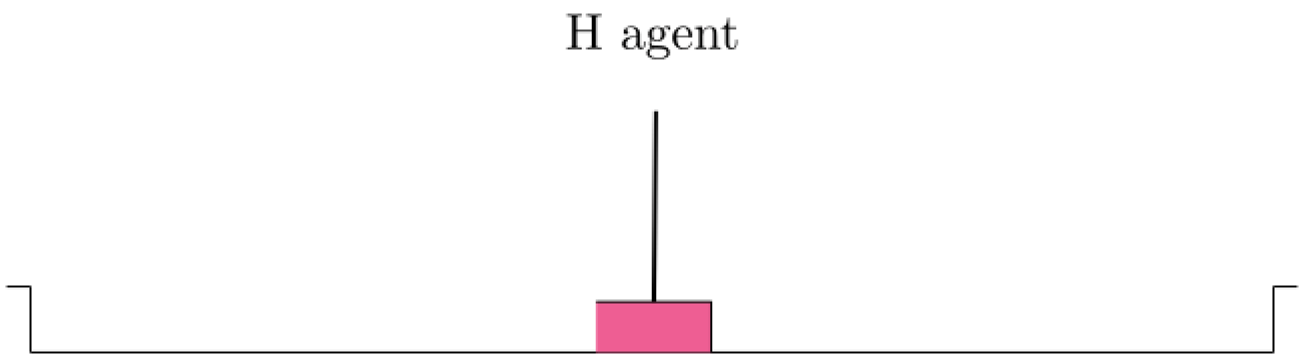
Dancing while balancing a pole

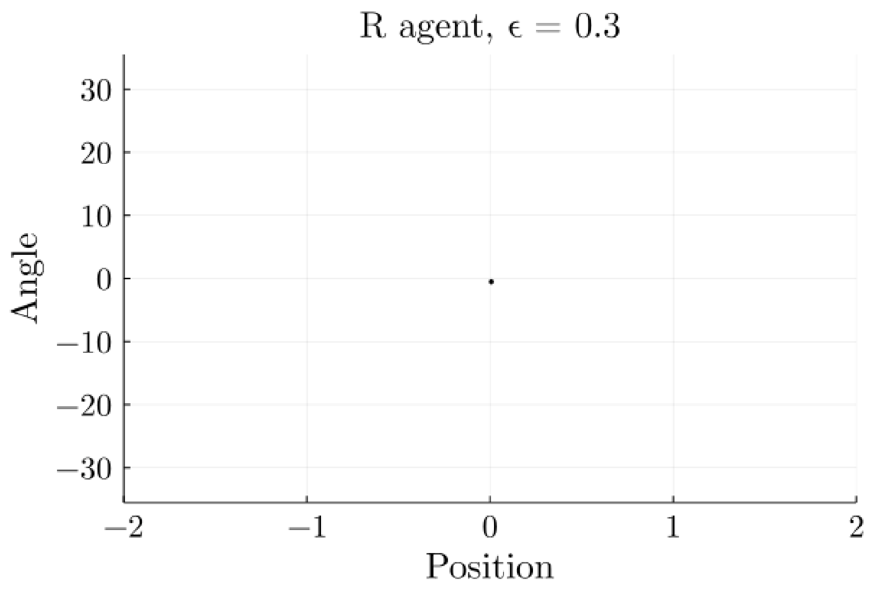


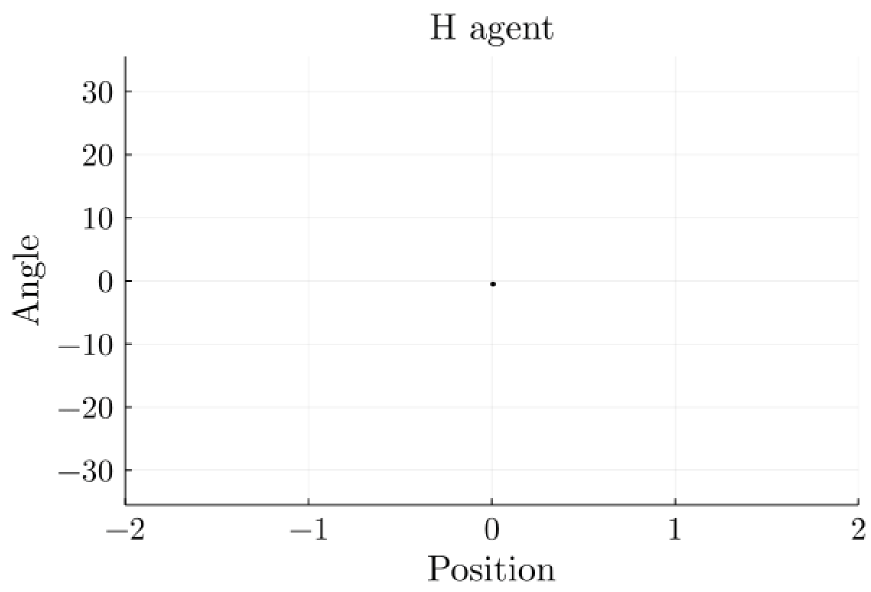
$R_\pi(s, a) = r(s, a)$: R agent

$R_\pi(s, a, s') = -\ln \pi(a|s) p(s'|s, a)$: H agent

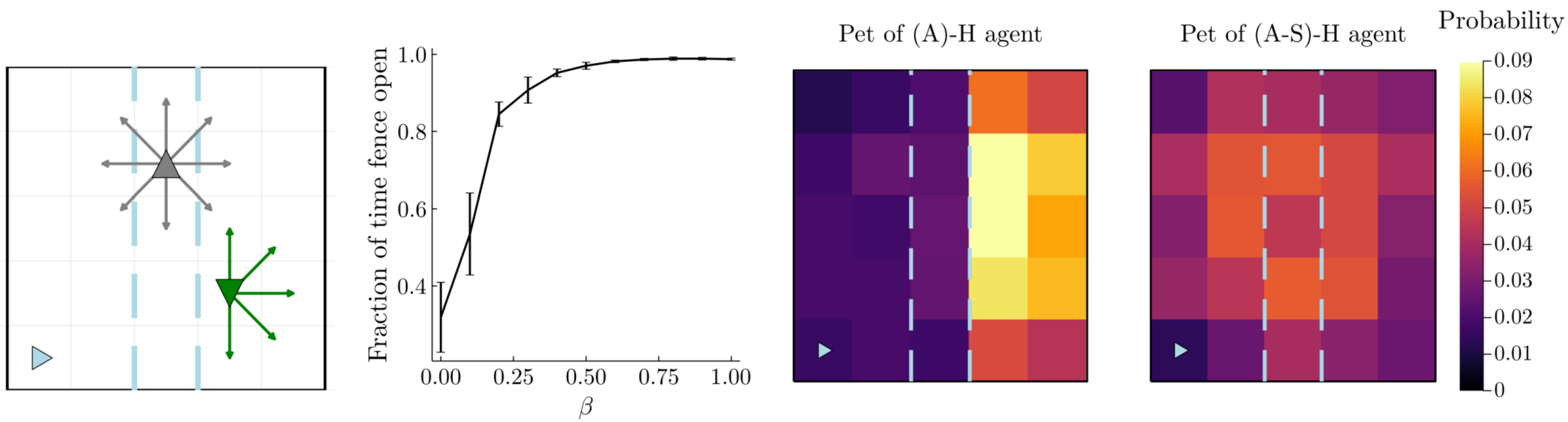








Altruistic behavior



$$R_\pi(s,a) = -\alpha \ln \pi(a|s) - \beta \ln p(s'|s,a)$$

$$s = (owner\ location, pet\ location)$$

Neural variability

- What is the mechanistic origin of neuronal variability?

Stochastic elements in the nervous systems (e.g., stochastic vesicle release) plus recurrent connections (i.e., feedback loops) (Moreno-Bote, PlosCB, 2014)

Chaotic dynamics due to strong recurrency (Van Vreeswijk & Sompolinsky, 1996)

- Hypothesis:

Variability is the result of the brain "occupying activity space"

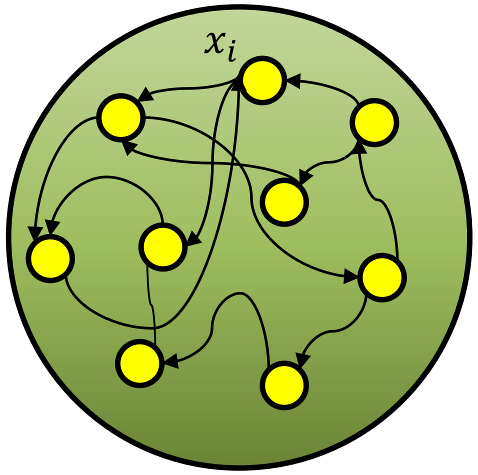
Thus, neuronal variability is promoted as long as it does not result into non-adaptive behavior or pathological activity

…but activity will be pushed close to pathological regimes

Generating and controlling variability

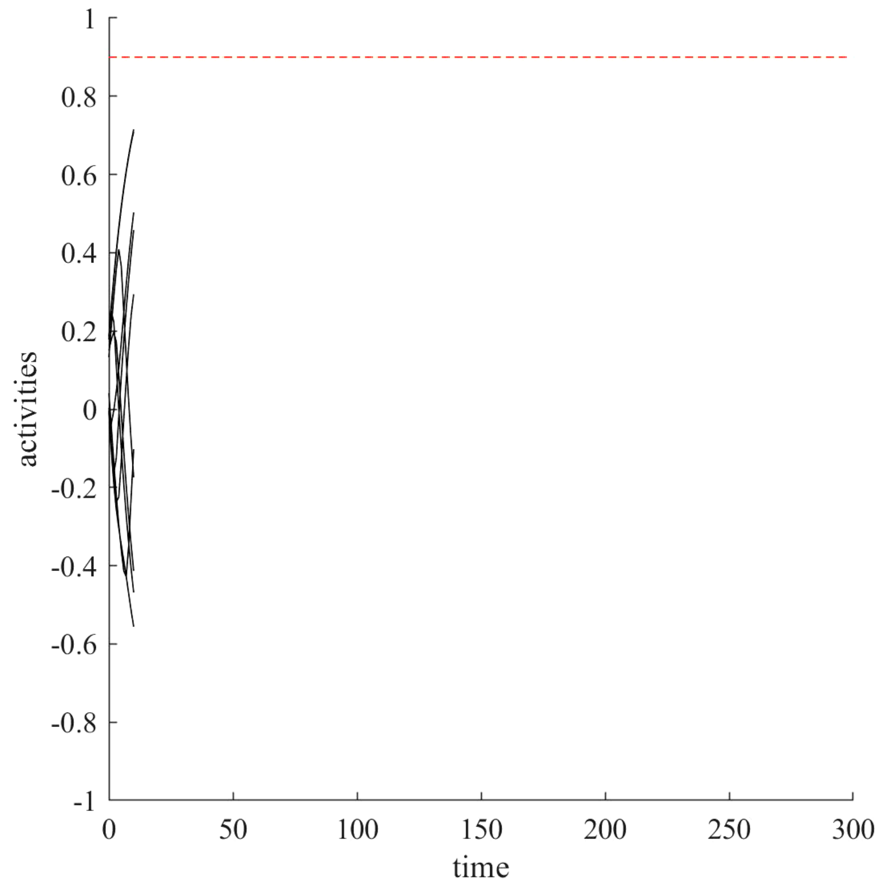
$$\frac{d}{dt}x_i = -x_i + f(\sum_j w_{ij}x_j)$$

$$N = 100$$
$$w_{ij} \sim N(0, \sigma)$$



RNN

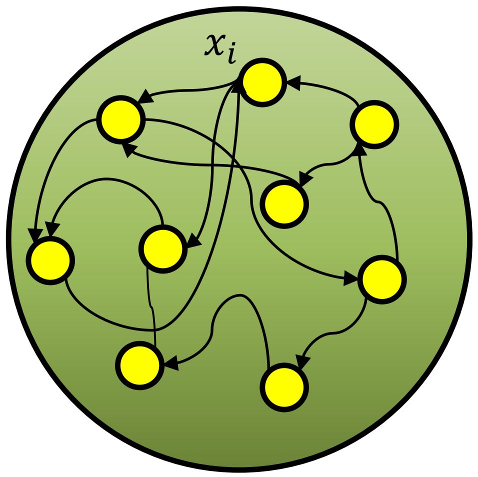
Generating and controlling variability



Uncontrolled Chaotic RNN

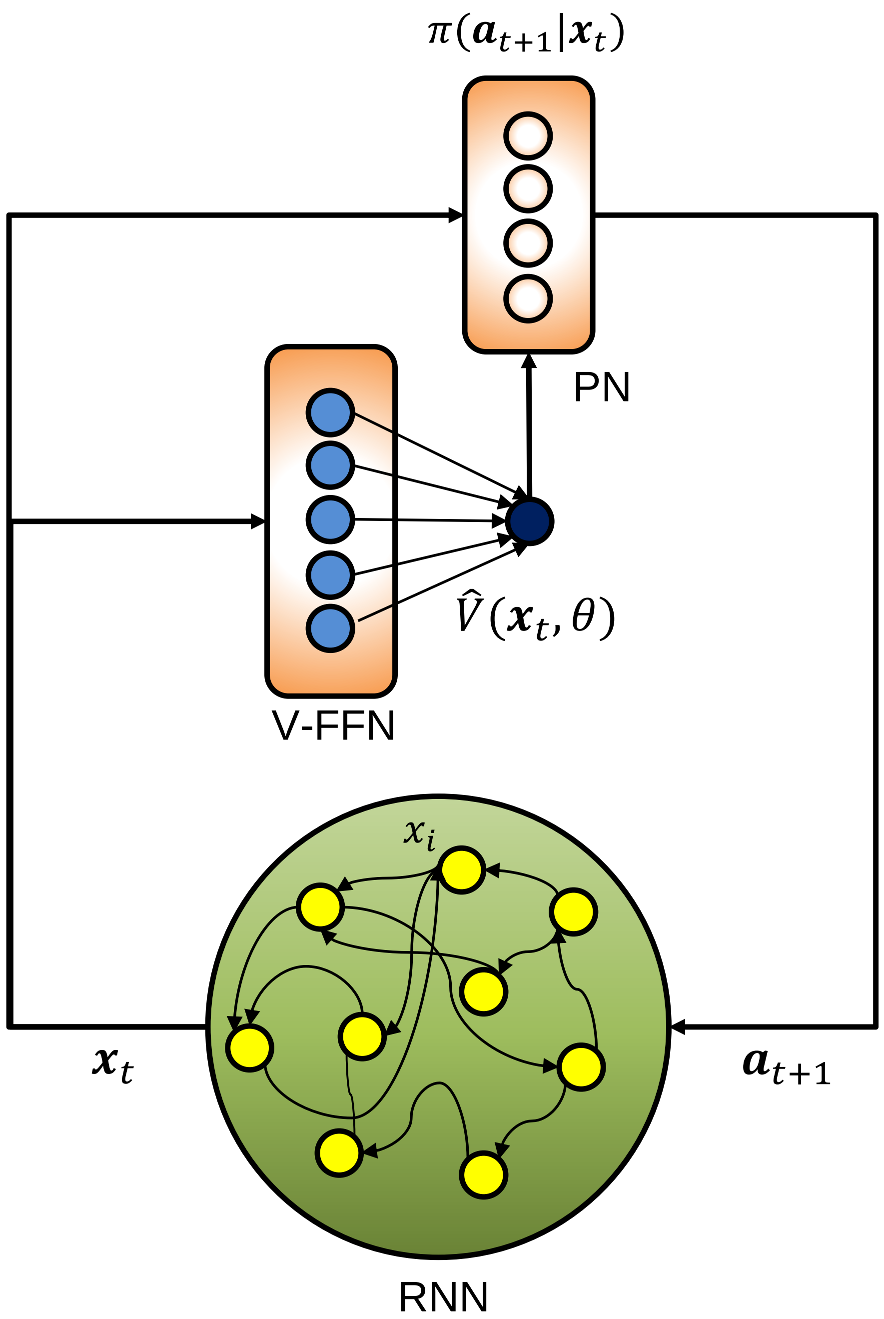
$$\frac{d}{dt}x_i = -x_i + f(\sum_j w_{ij} x_j)$$

$$N = 100$$
$$w_{ij} \sim N(0, \sigma)$$



RNN

Generating and controlling variability

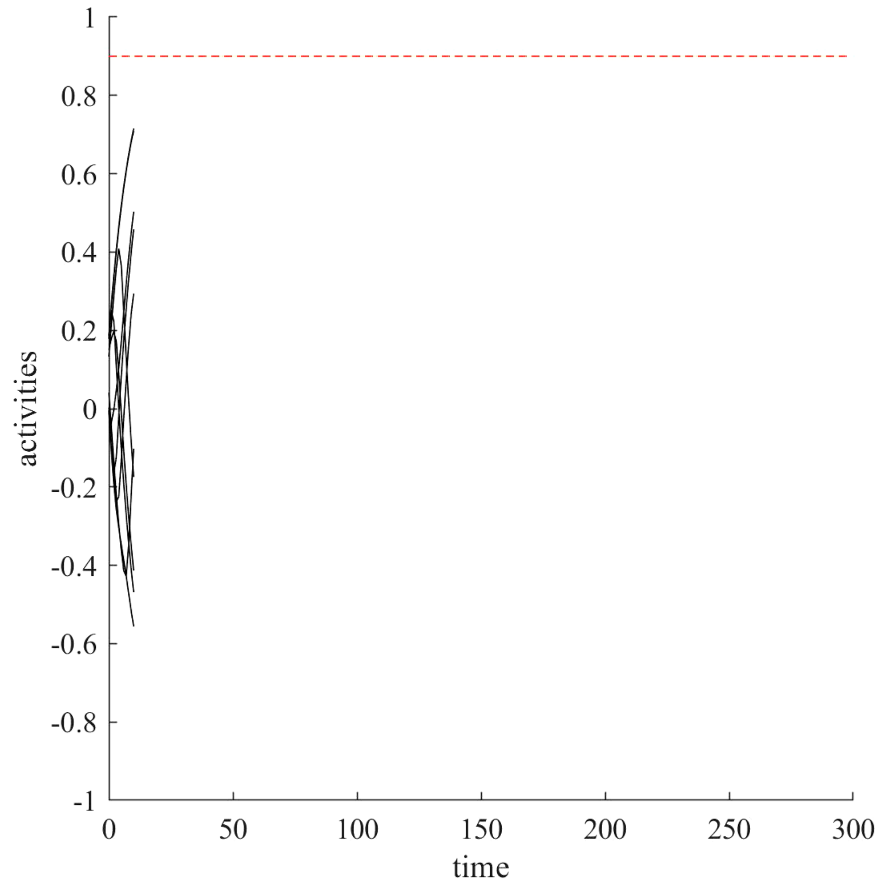


$$\frac{d}{dt}x_i = -x_i + f(\sum_j w_{ij}x_j + \sum_k v_{ik}a_k)$$

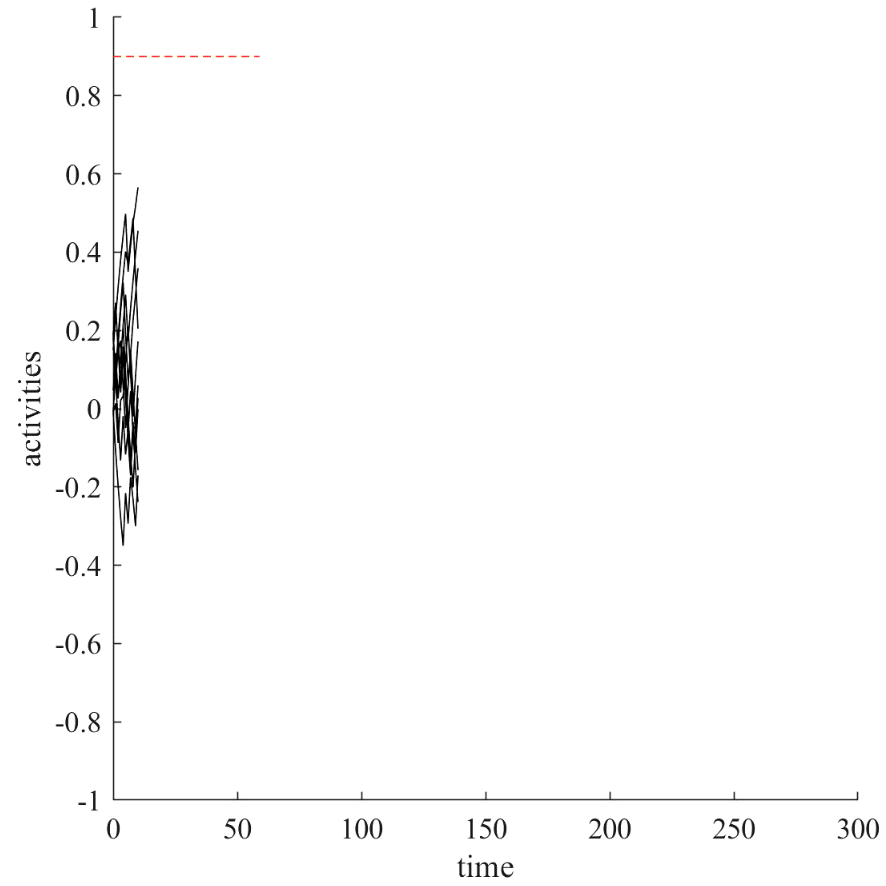
$$\theta_{k+1} = \arg\min_\theta \sum_{\boldsymbol{x}_t \in paths}\left(\hat{V}(\boldsymbol{x}_t, \theta) - \ln\left(\sum_a e^{\hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t, \boldsymbol{a}), \theta_k)}\right)\right)^2$$

$$\pi(\boldsymbol{a}|\boldsymbol{x}_t) \propto \exp(\gamma \hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t, \boldsymbol{a}), \theta))$$

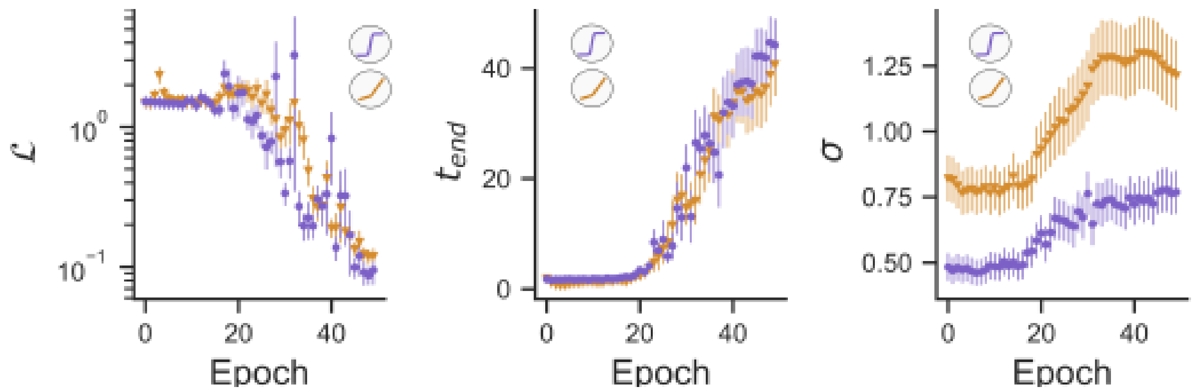
Generating and controlling variability



Uncontrolled Chaotic RNN



Controlled Chaotic RNN



Classification of (recursive) frameworks

$$V_\pi(s) = \mathbb{E}_{a,s'|s,\pi}[R_\pi(s,a,s')] + \gamma\, \mathbb{E}_{s'|s,\pi}[V_\pi(s')]$$

$$R_\pi(s,a,s') = \underbrace{\underbrace{\underbrace{\underbrace{r(s,a,s')}_{\text{Standard RL (policy-independent reward)}} - \alpha \ln \pi(a|s)}_{\text{Entropy-regularized RL (action-entropy RL)}} + \alpha_0 \ln \pi_0(a|s)}_{\text{KL RL (action-KL RL)}} - \beta \ln p(s'|s,a) + \beta_0 \ln p_0(s'|s,a)}_{\text{action-state KL RL (*)}}$$

Soft RL

(*) New

Reward-free frameworks

$$R_\pi(s,a,s') = \underbrace{\underbrace{-\alpha \ln \pi(a|s)}_{\text{max occupancy principle (MOP) (*)}} - \beta \ln p(s'|s,a)}_{\substack{\text{action-state MOP } (\alpha > 0, \beta \geq 0) \text{ (*)} \\ \text{empowerment } (\alpha > 0, \beta < 0) \text{ (*)}}}$$

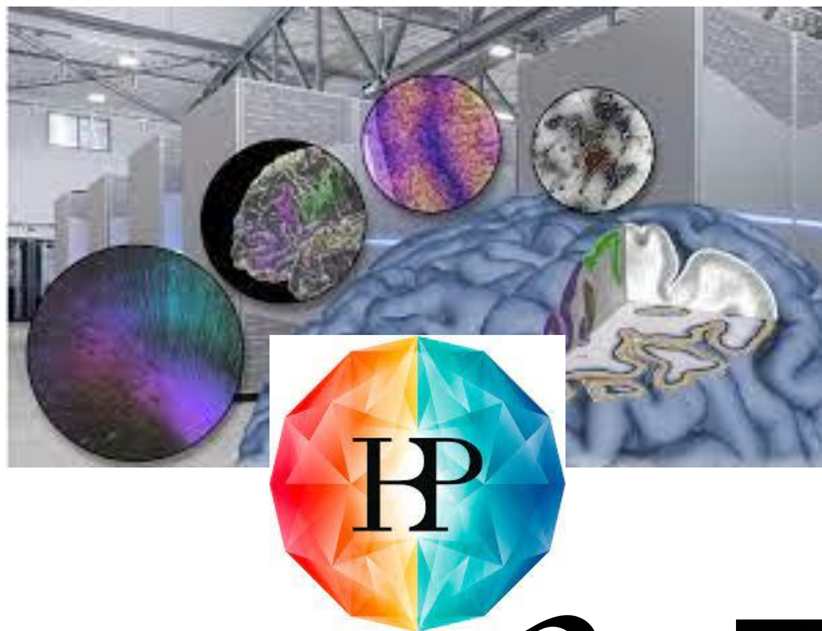
Conclusions

- Are we really utility maximizers?
- Defining reward functions is problematic, even dangerous
- MOP principle: *the goal is to occupy action-state path space*
- External rewards are the means to accomplish that objective
- Entropy seeking behavior is fun, lively and energetic
- Goal-directed behavior emerges

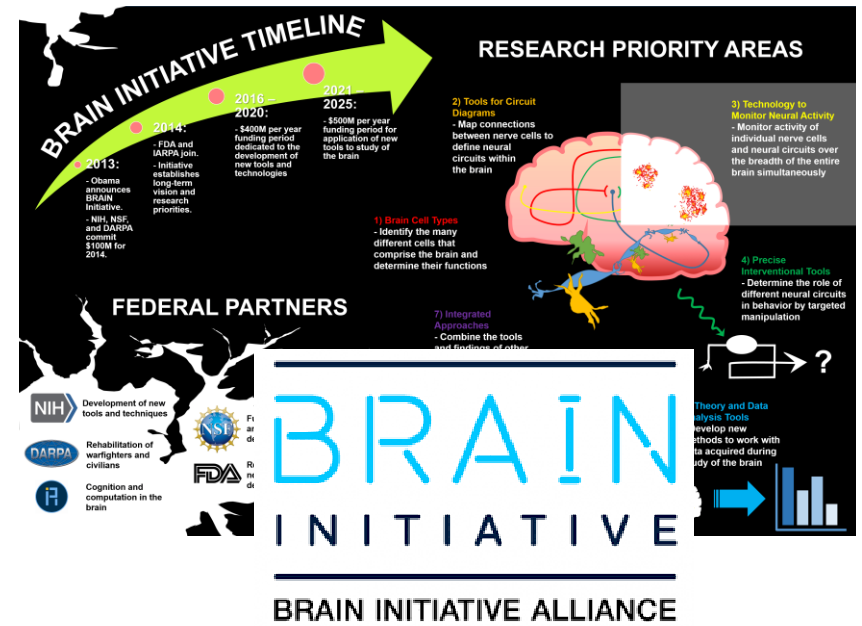
(terminal states and internal states are critical)

- A possible account of neural variability

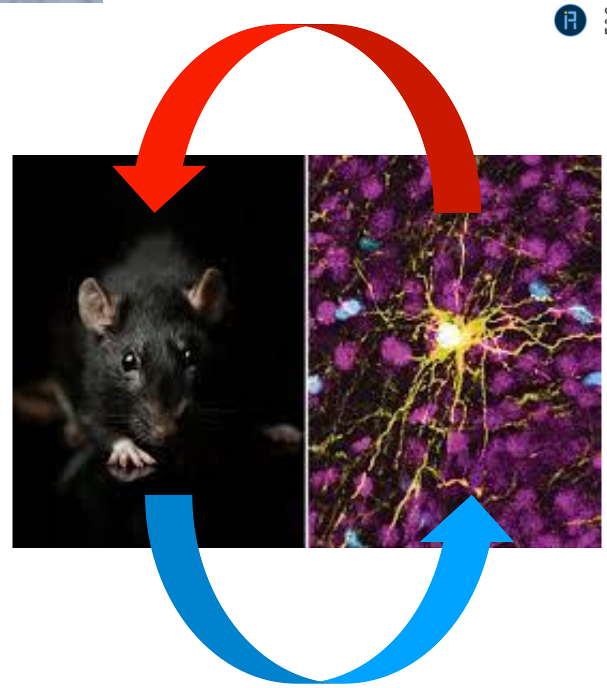
Approaches for Brain and Behavior



HBBP



BBI



Getty Images

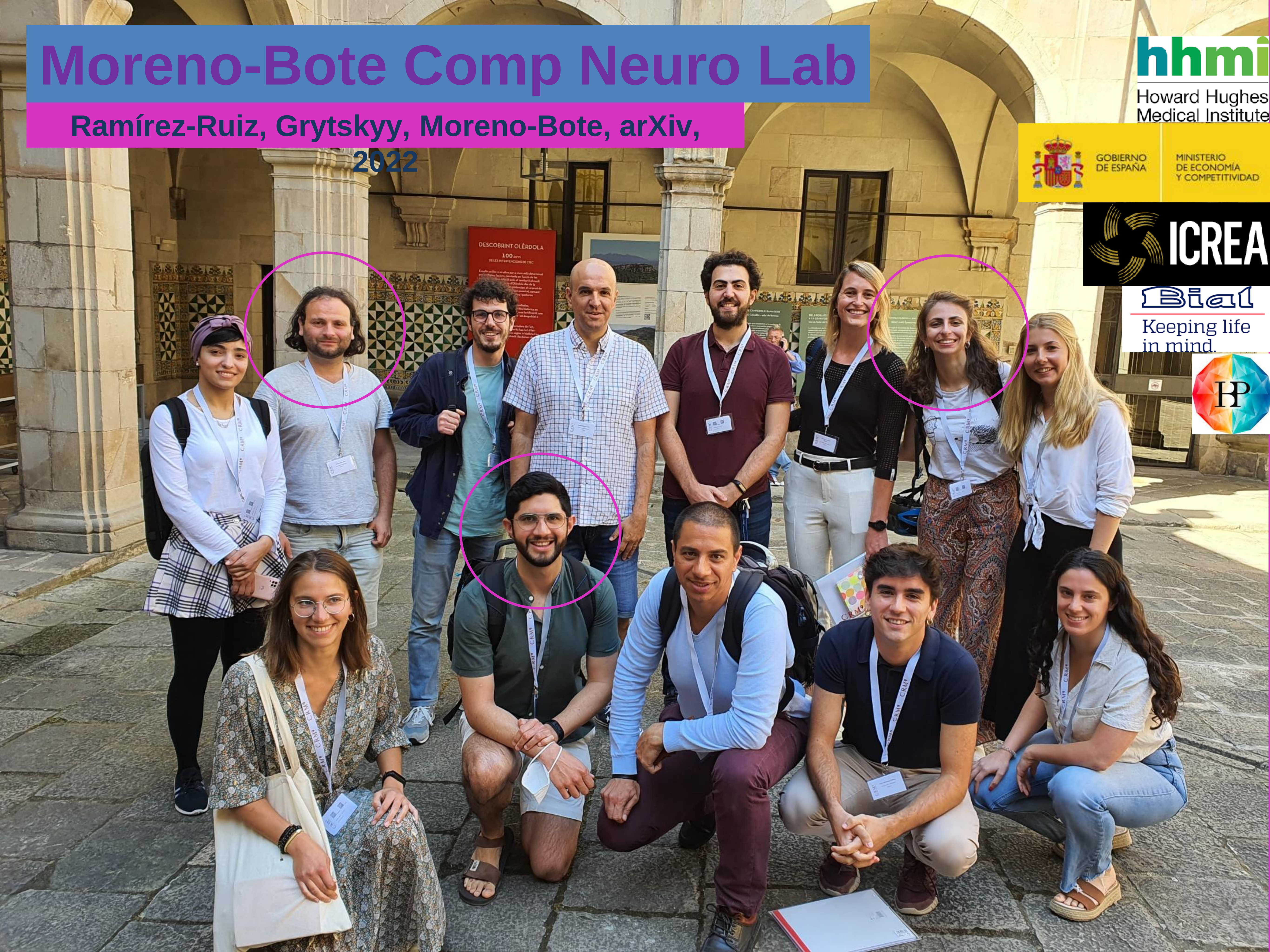
Bottom-Up Approach: from synapses neurons and circuits to emerging behaviors

- emphasis on data collection and simulation, but not on theory
- no emphasis on behavior

Proposal. **Top-Down approach**: from behavior to synapses, neurons and circuits

Moreno-Bote Comp Neuro Lab

Ramírez-Ruiz, Grytskyy, Moreno-Bote, arXiv, 2022



hhmi
Howard Hughes Medical Institute

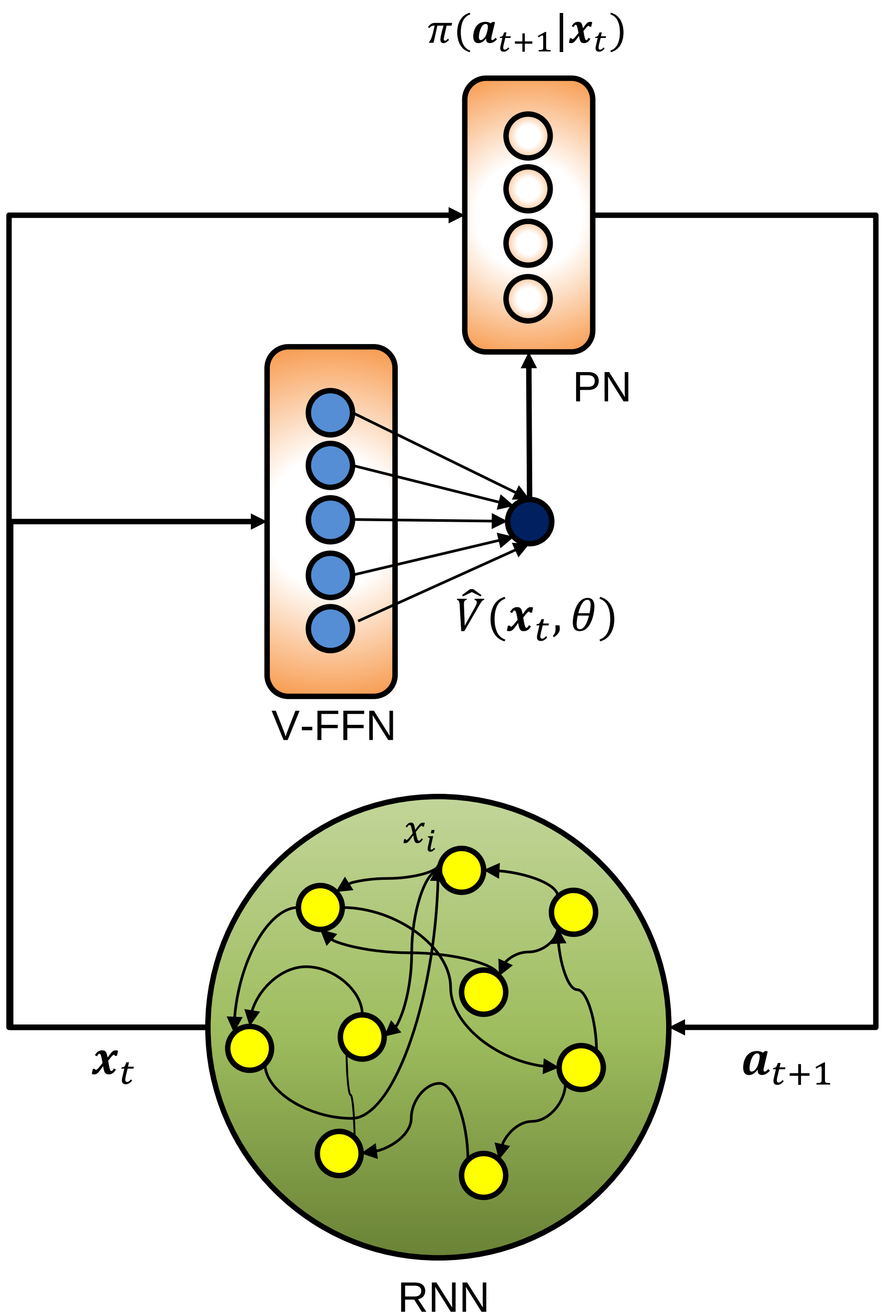
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

ICREA

Bial
Keeping life in mind.

HP

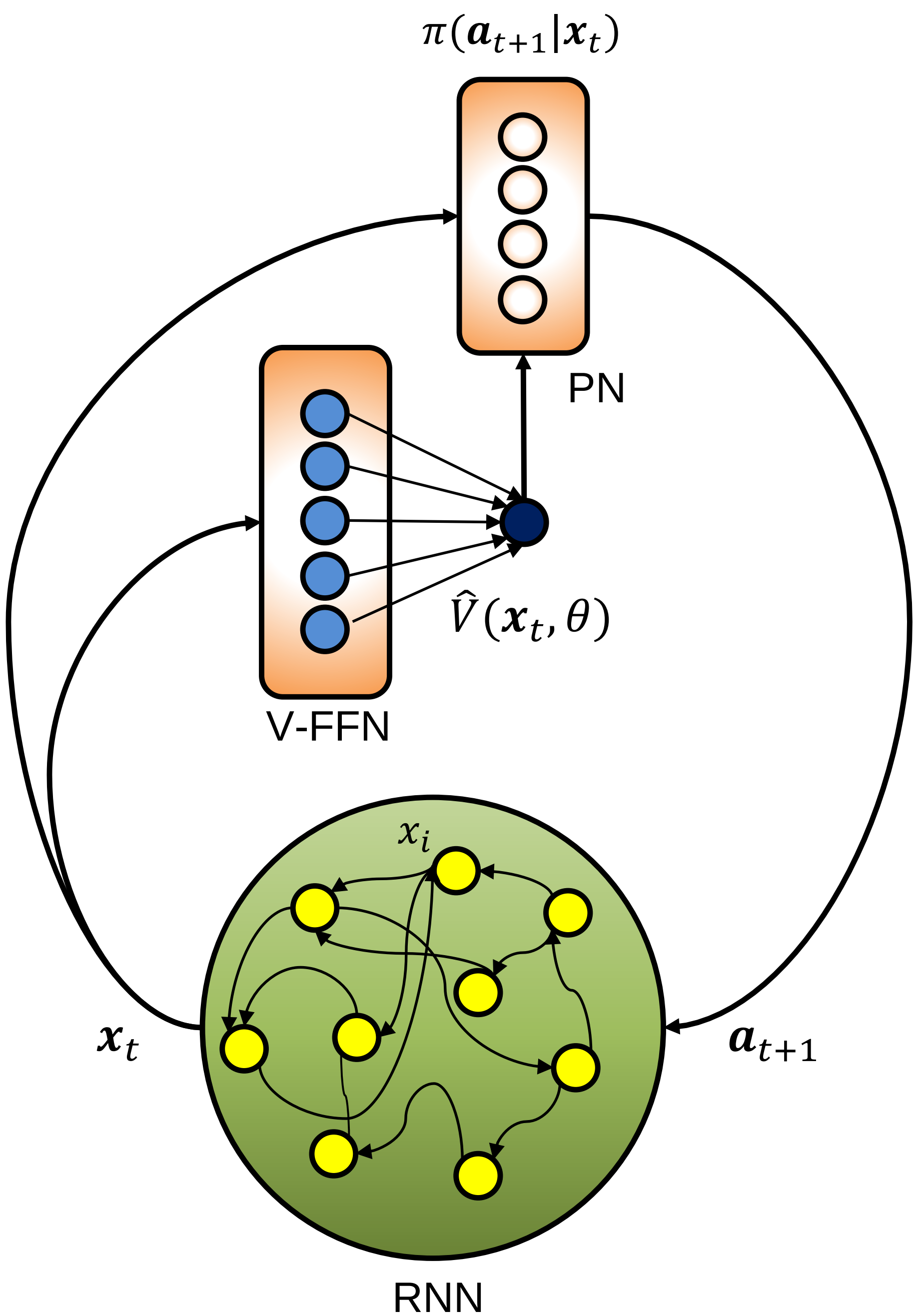
Generating and controlling variability



$$\frac{d}{dt}x_i = -x_i + f(\sum_j w_{ij}x_j + \sum_k v_{ik}a_k)$$

$$\theta_{k+1} = \arg\min_{\theta} \sum_{\boldsymbol{x}_t \in paths} \left(\hat{V}(\boldsymbol{x}_t,\theta) - \ln\left(\sum_a e^{\hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t,\boldsymbol{a}),\theta_k)}\right)\right)^2$$

$$\pi(\boldsymbol{a}|\boldsymbol{x}_t) \propto \exp(\gamma \hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t,\boldsymbol{a}),\theta))$$



$$\frac{d}{dt}x_i = -x_i + f(\sum_j w_{ij}x_j + a_i)$$

$$\pi(\boldsymbol{a}|\boldsymbol{x}_t) \propto \exp(\gamma\hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t,\boldsymbol{a}),\theta))$$

$$\theta^* = \arg\min_\theta \sum_{\boldsymbol{x}_t \in paths}\left(\hat{V}(\boldsymbol{x}_t,\theta) - \ln\left(\sum_a e^{\hat{V}(\boldsymbol{x}'(\boldsymbol{x}_t,\boldsymbol{a}),\theta)}\right)\right)^2$$