Bayesian Control Rule

Pedro A. Ortega

Max Planck Institute for Intelligent Systems
Max Planck Institute for Biological Cybernetics

Overview

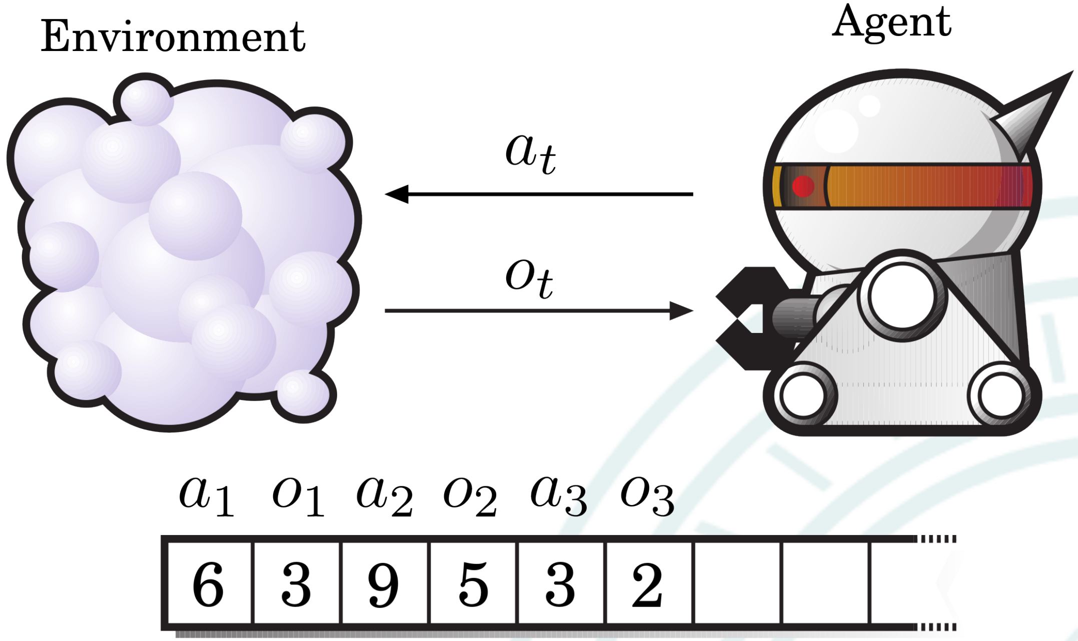
- Introduction
- Adaptation
- Causality
- Bayesian control rule
- Conclusions

Introduction

Agent-Environment Setup

$$o_t \sim Q(o_t|a_{\le t}, o_{<t})$$

$$o_t \in \mathcal{O}$$



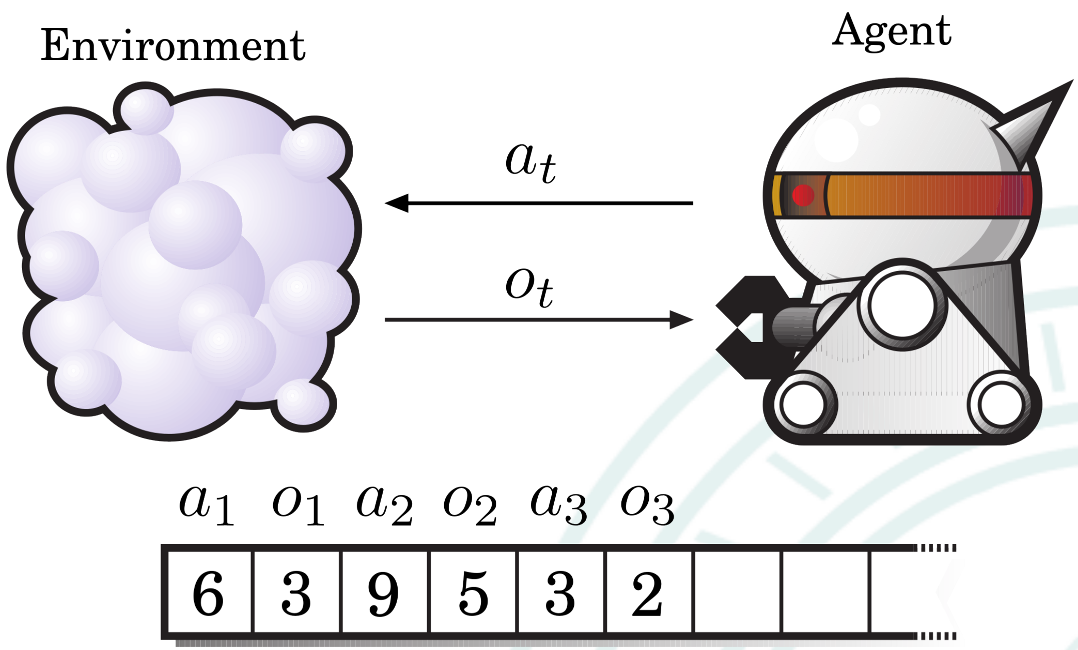
$$a_t \sim P(a_t|a_{<t}, o_{<t})$$

$$a_t \in \mathcal{A}$$

Agent-Environment Setup

$o_t \sim Q(o_t | a_{\leq t}, o_{<t})$

$o_t \in \mathcal{O}$

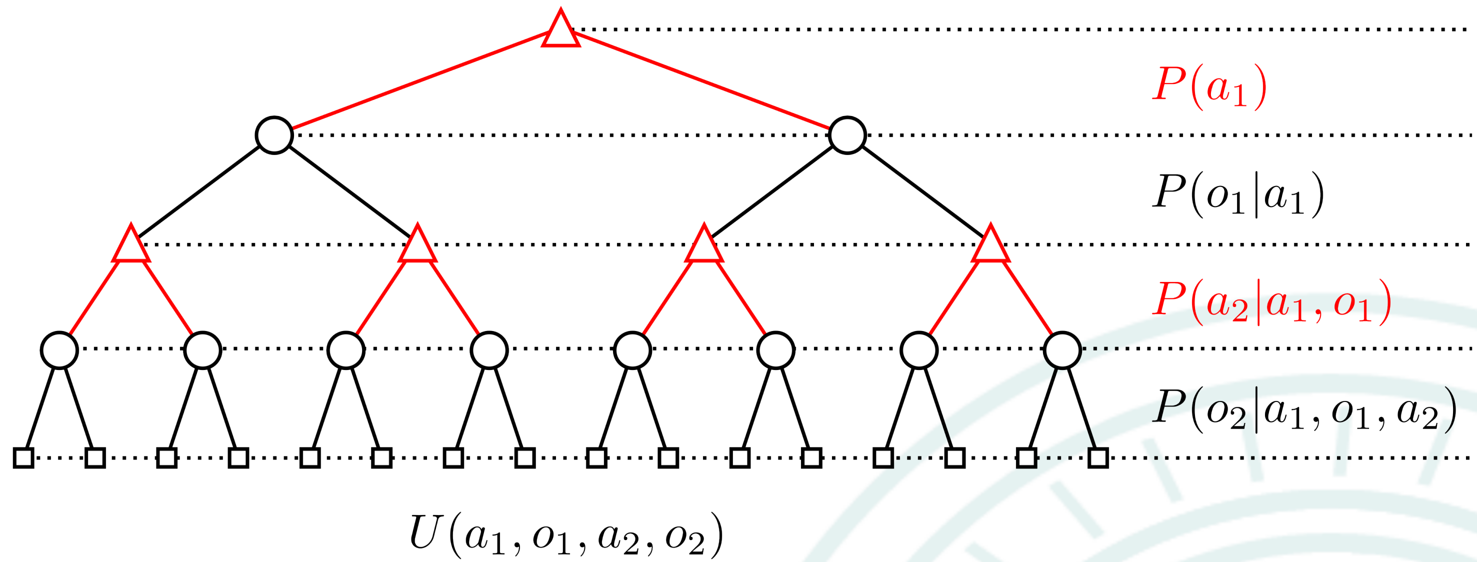


$a_t \sim P(a_t | a_{<t}, o_{<t})$

$a_t \in \mathcal{A}$

Environment can be a bandit, MDP, POMDP or any other **controllable stochastic process**.

Adaptive Control



In theory:

- Choose policy maximizing **subjective expected utility**.

In practice: intractable!

- Policy space **grows exponentially** with planning horizon.
- Policy choice **causally precedes** interactions.

Choose policy **before** interacting?

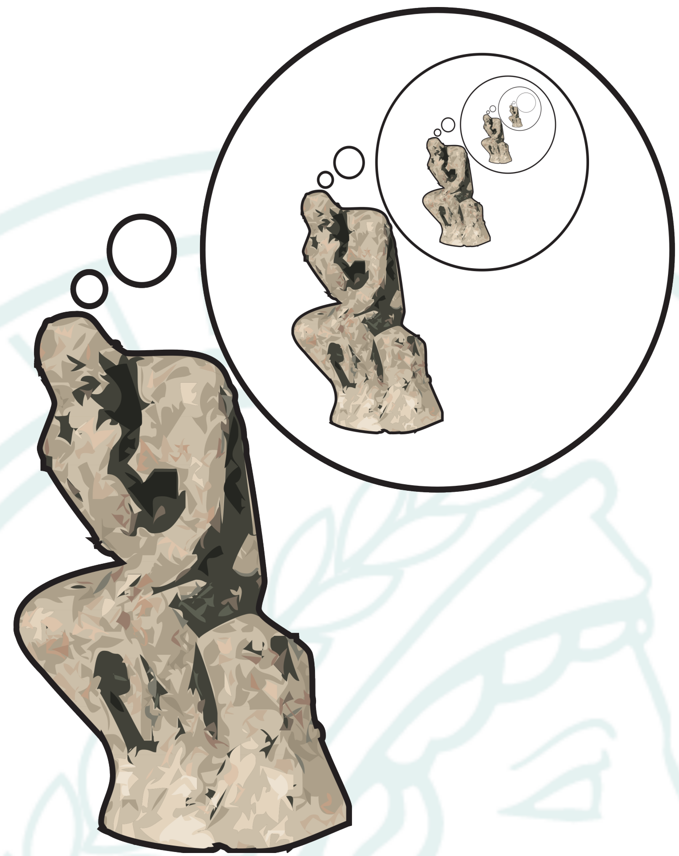
What if choosing the optimal policy was tractable?

This implies:

- precomputing **all the possible lives**,
- and then picking the **optimal policy**.

However:

- Prediction has no accuracy, because it is **not supported** by any data.
- The optimal policy is **statistically meaningless in the beginning**!



Can we choose policies **dynamically**?

- **Delay** choice of optimal policy – when **justified** by data.
- Agent is **uncertain** about the optimal policy.
- **Practical** adaptive control and RL **do this** explicitly/implicitly.
- Implementation of "**intuition**"

Questions

How do we choose the optimal policy **dynamically**?

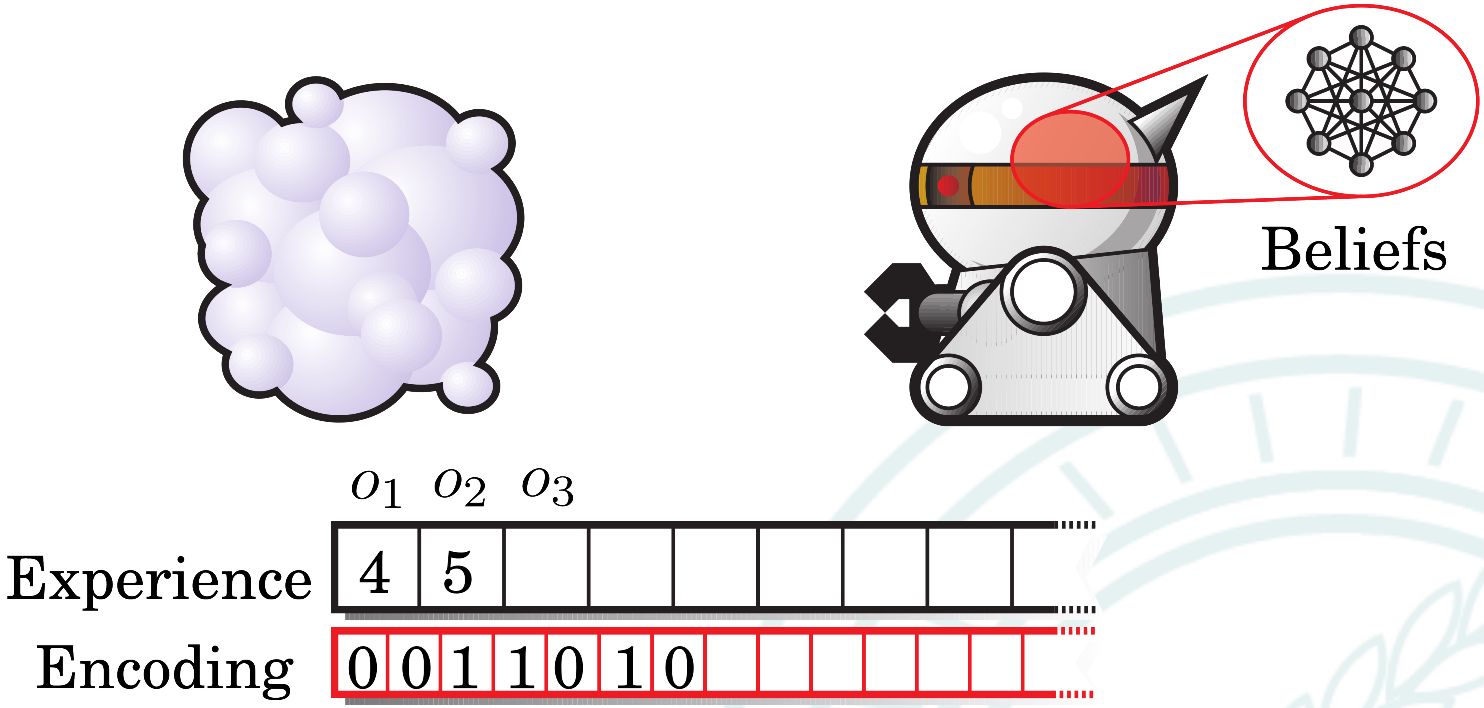
How is uncertainty over the policy **represented**?

How are **actions issued** when the policy is uncertain?

How is this uncertainty **reduced**?

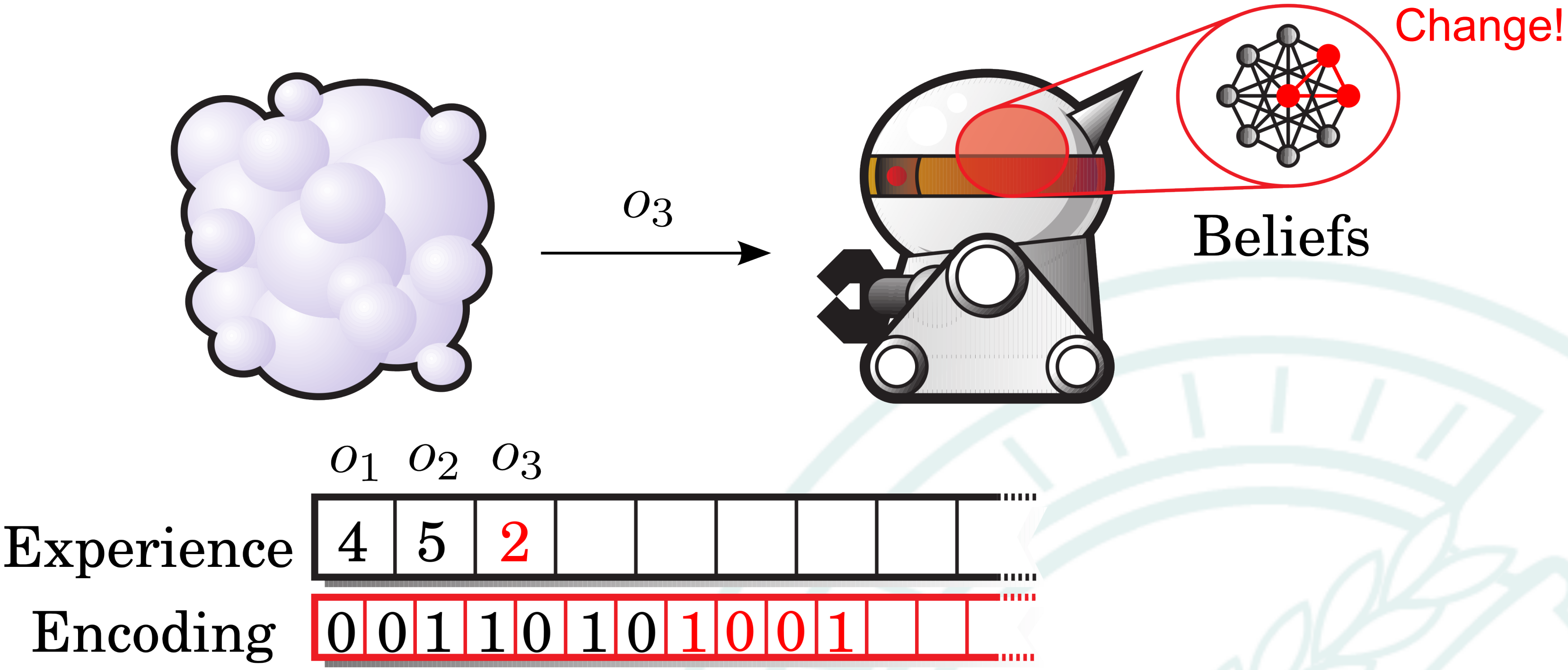
Adaptation

The Cost of Experience



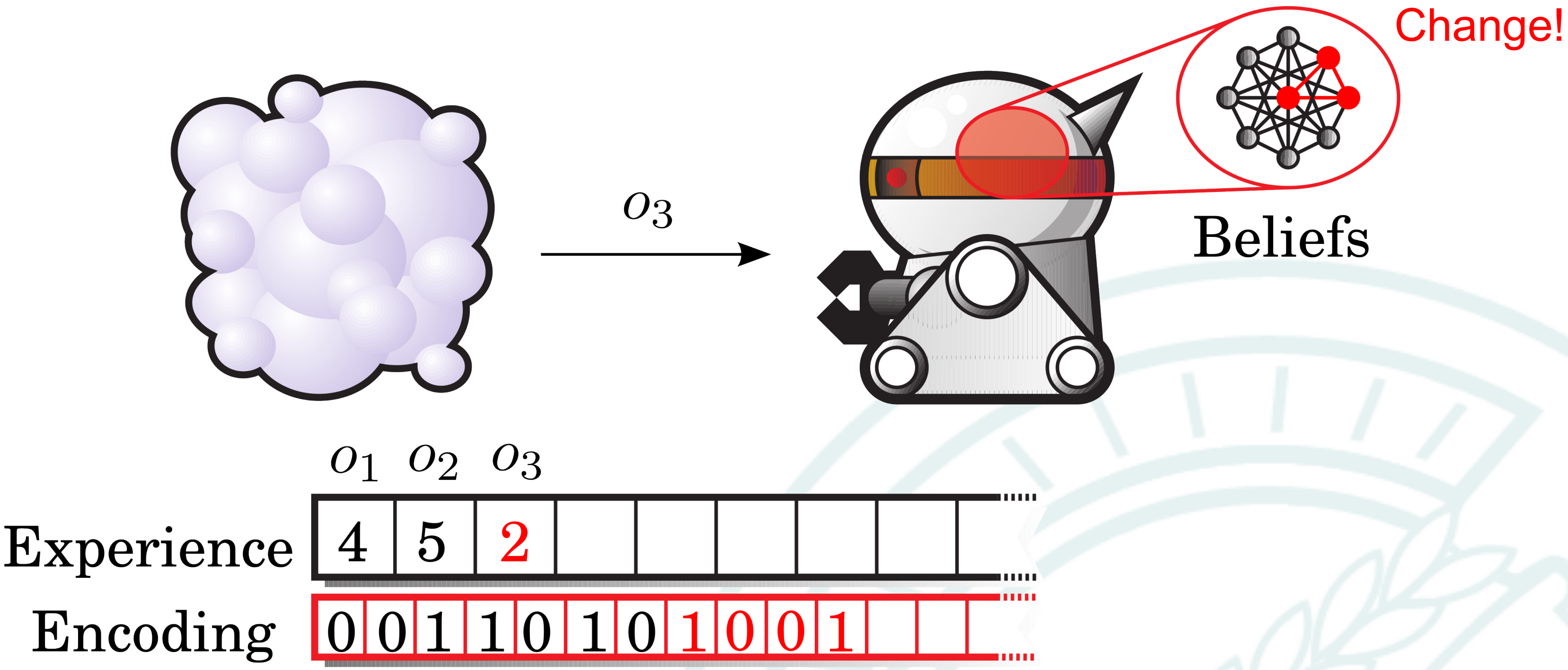
- Agent records observations.

The Cost of Experience



- Agent records observations.
- Acquiring experience implies **changes** in the belief structure.

The Cost of Experience



- Agent records observations.
- Acquiring experience implies **changes** in the belief structure.
- **Can we minimize these changes?**

Adaptive Compression

- When the environment is **known**, maximal compression is achieved when codeword lengths are chosen as

$$l(o_{\leq t}) := -\log Q(o_{\leq t})$$

- Conversely, every code **implies predictions**

$$P(o_{\leq t}) = 2^{-l(o_{\leq t})}$$

- The belief structure of the agent **embodies the assumptions** about the environment.

Adaptive Compression (cont.)

- How to compress when the environment is **unknown**?
- Consider set of possible environments Θ, with probabilities $P(\theta)$ and models $P(o_{\le t}|\theta)$.
- Choose a predictor $\tilde{P}$ minimizing expected codeword length:

Choice of θ — Environment θ — Predictor

$$L_t[\tilde{P}] = \sum_{\theta} P(\theta) \left\{ \sum_{o_{\le t}} P(o_{\le t}|\theta) \log \frac{P(o_{\le t}|\theta)}{\tilde{P}(o_{\le t})} \right\}$$

Adaptive Compression (cont.)

- Solution: **Bayesian mixture**

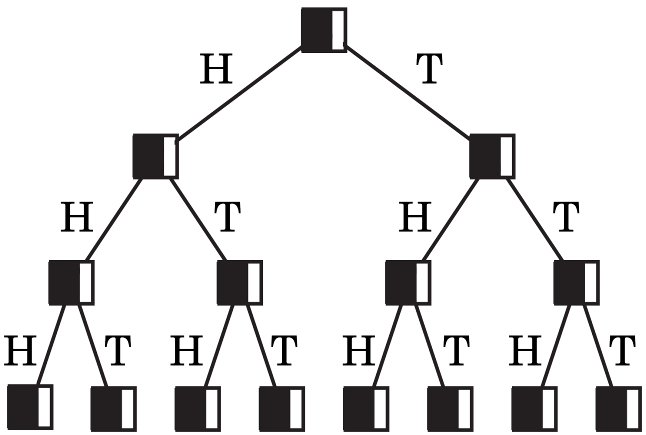
$$\tilde{P}(o_{\leq t}) := \sum_{\theta} P(o_{\leq t}|\theta)P(\theta) = P(o_{\leq t})$$

- Predictive distribution

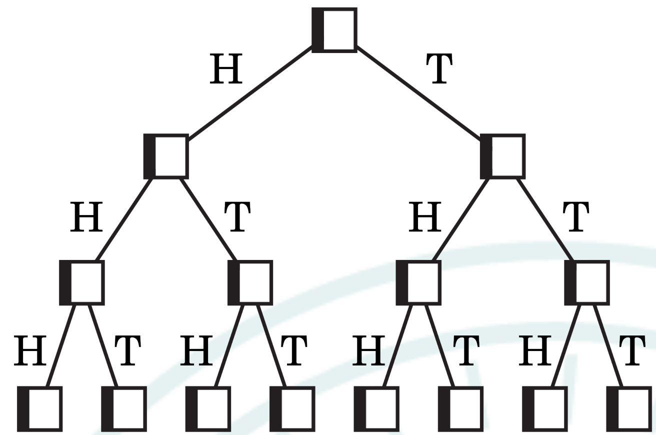
$$P(o_t|o_{<t}) = \sum_{\theta} P(o_t|o_{<t})P(\theta|o_{<t})$$

- Bottom line: adaptive compression is solved by **pretending** that the Bayesian mixture is the true environment

Example: Prediction of Biased Coin

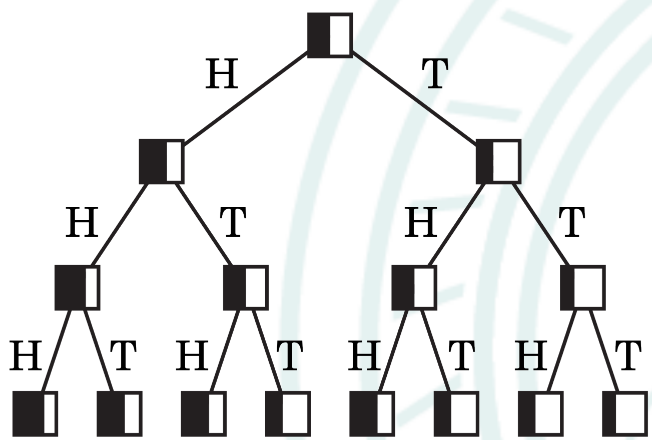


bias $\theta = \frac{3}{4}$



bias $\theta = \frac{1}{4}$

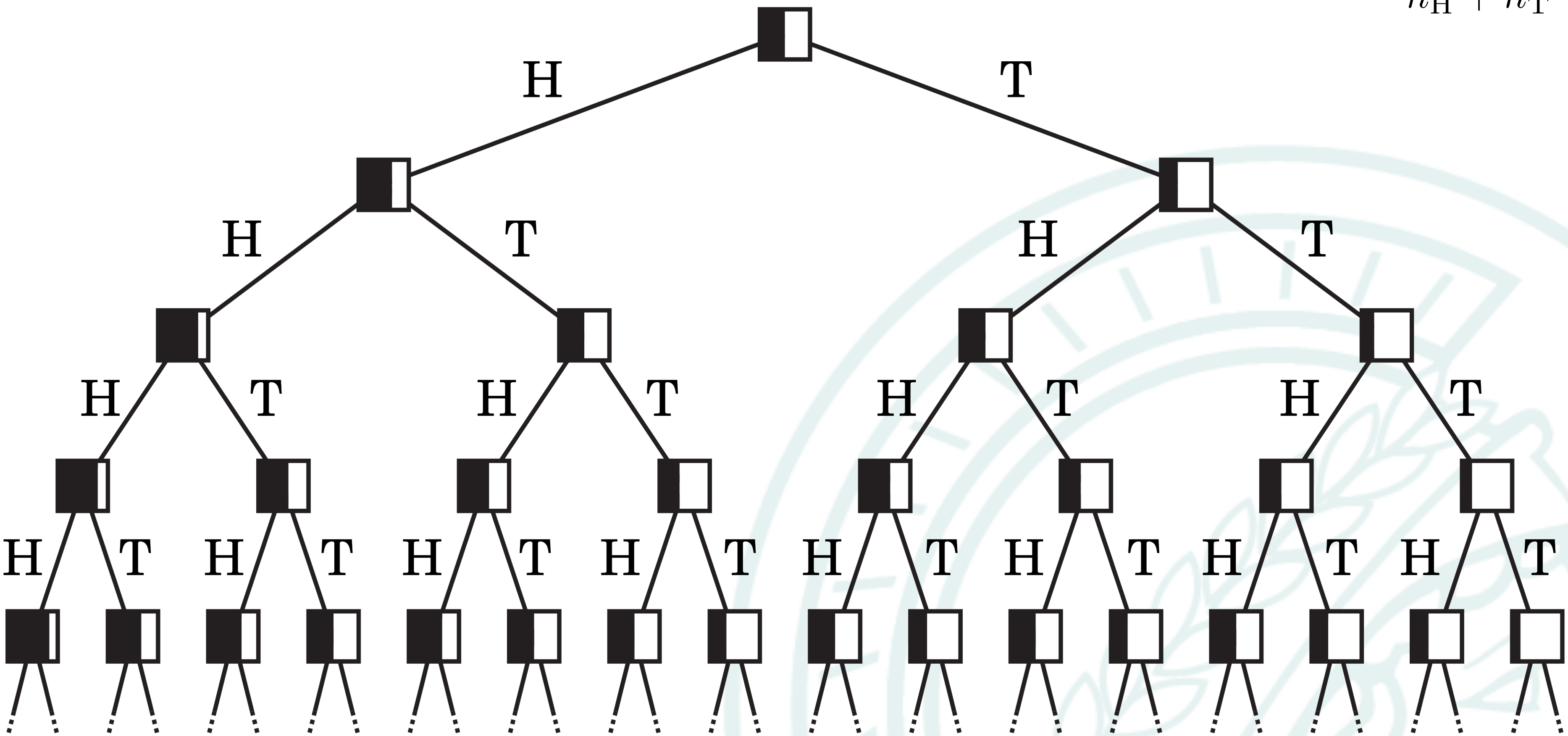
$P(\theta) = \frac{1}{2}$



mixture

Example: Prediction of Biased Coin II

$$P(\text{H}|\text{experience}) = \frac{n_\text{H} + 1}{n_\text{H} + n_\text{T} + 2}$$



mixture over all biases in [0,1]

Summary

The Bayesian mixture is the optimal compressor of experience for an unknown environment.

Causality

Extension to Actions

- Can we use this for **adaptive behavior**?

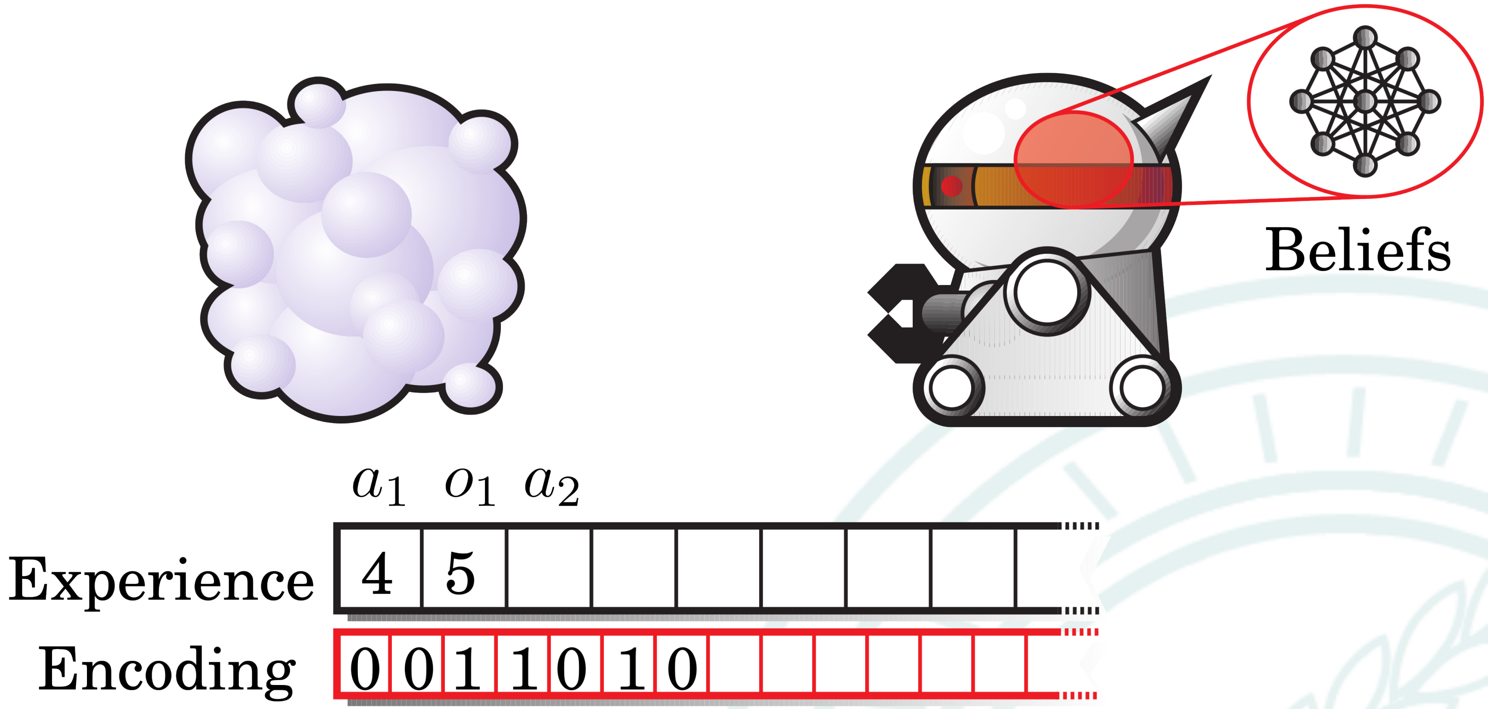
- Instead of **competing hypotheses**, we would have **competing behaviors** $(\theta, \pi) \in \Theta \times \Pi$:

$$P(a_{\leq t}, o_{\leq t} | \theta, \pi) \qquad P(\theta, \pi)$$

- Would lead to

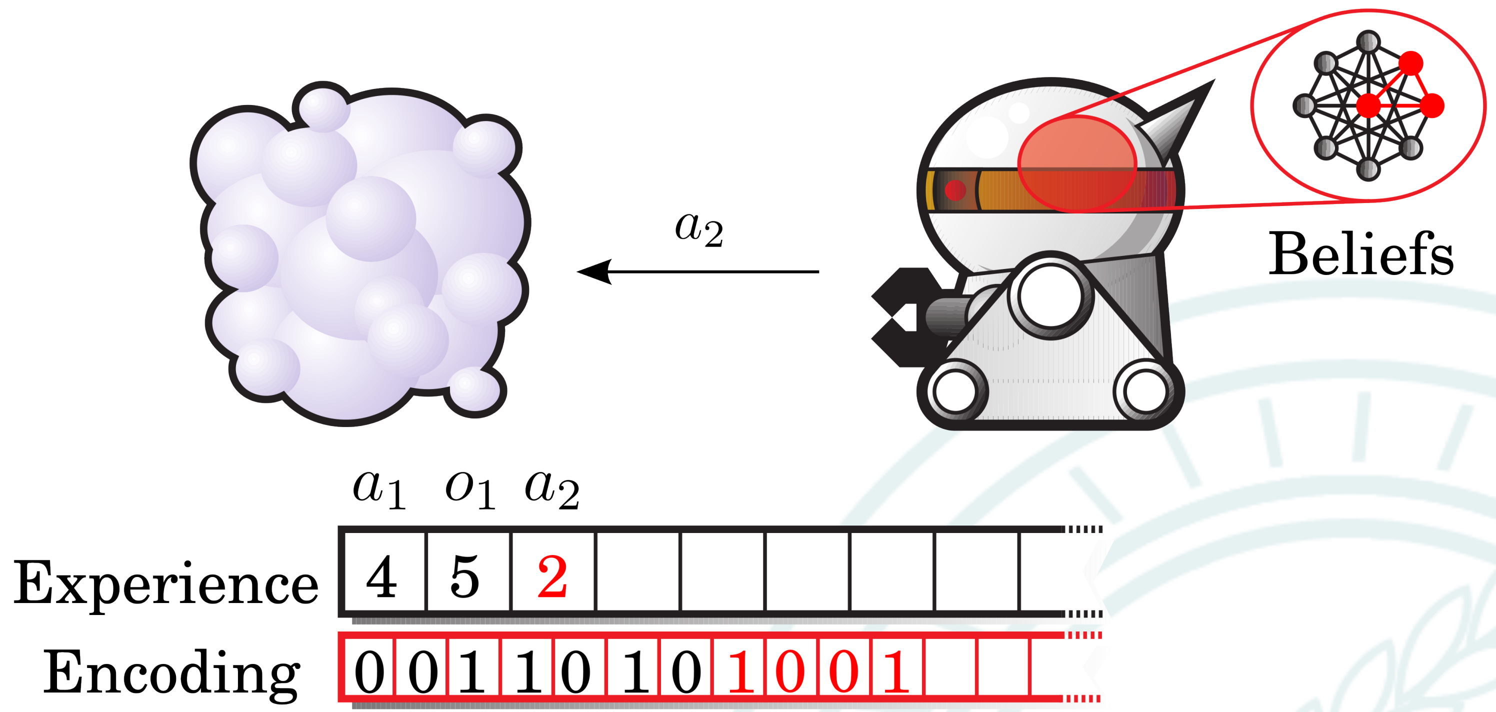
$$P(\text{next action}|\text{experience}) = P(a_t | a_{<t}, o_{<t})$$

The Cost of Experience II



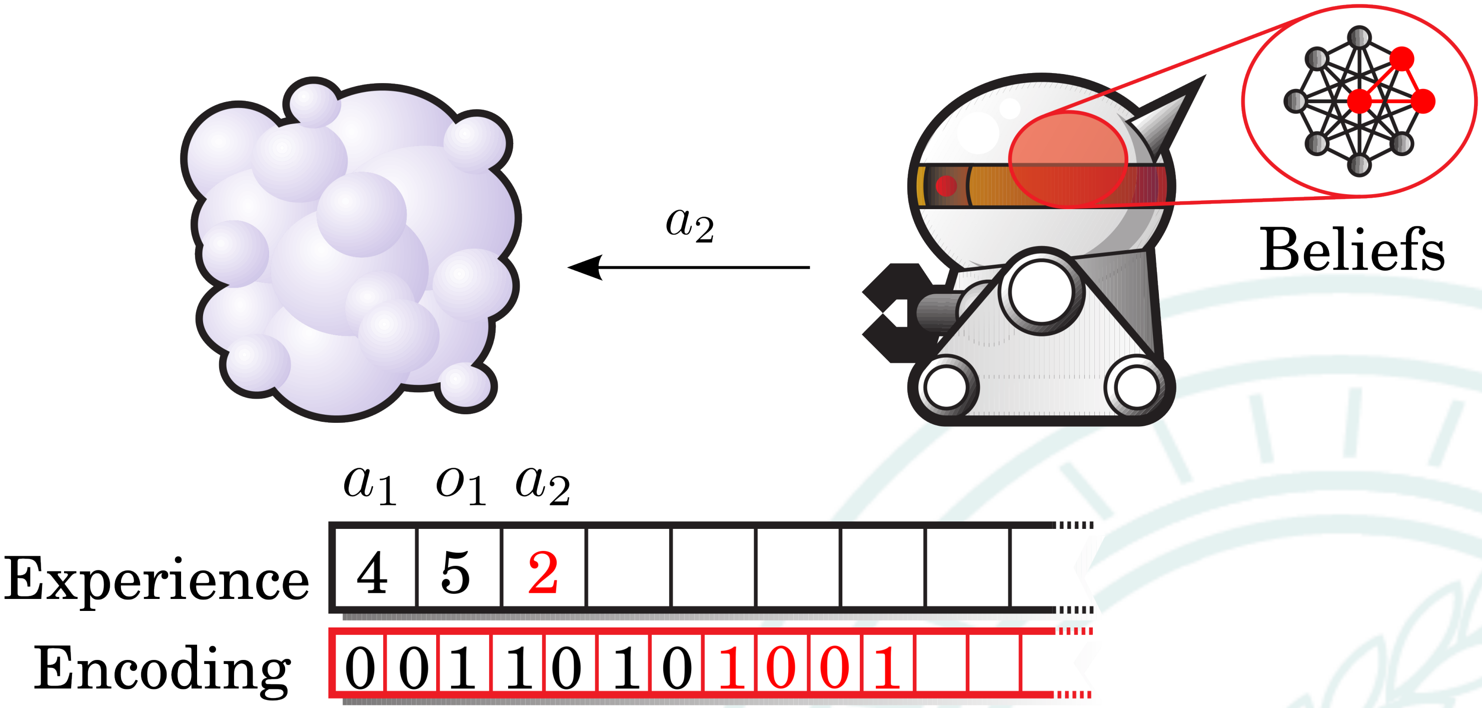
- Agent records actions & observations.

The Cost of Experience II



- Agent records actions & observations.
- Again, actions **change** the belief structure.

The Cost of Experience II



- Agent records actions & observations.
- Again, actions **change** the belief structure.
- However, they **do not change the beliefs**.

Problems of the Bayesian Update

- Posterior beliefs

$$
\begin{aligned}
P(\theta, \pi | a_t, o_t, \ldots) \\
&\propto \text{likelihood} \times \text{prior} \\
&= P(o_t | \theta, a_t, \ldots) P(a_t | \pi, \ldots) \times P(\theta, \pi | \ldots)
\end{aligned}
$$

Problems of the Bayesian Update

- Posterior beliefs

$$
\begin{aligned}
P(\theta, \pi | a_t, o_t, \ldots) \\
&\propto \text{likelihood} \times \text{prior} \\
&= P(o_t|\theta, a_t, \ldots) P(a_t|\pi, \ldots) \times P(\theta, \pi | \ldots)
\end{aligned}
$$

...but **our actions produce evidence**, we conclude the optimal policy from our own actions.

Problems of the Bayesian Update

- Posterior beliefs

$$P(\theta, \pi | a_t, o_t, \ldots)$$
$$\propto \text{likelihood} \times \text{prior}$$
$$= P(o_t | \theta, a_t, \ldots) P(a_t | \pi, \ldots) \times P(\theta, \pi | \ldots)$$

...but **our actions produce evidence**, we conclude the optimal policy from our own actions.

Problems of the Bayesian Update

- Posterior beliefs

$$P(\theta, \pi | a_t, o_t, \ldots) \\ \propto \text{likelihood} \times \text{prior} \\ = P(o_t | \theta, a_t, \ldots) P(a_t | \pi, \ldots) \times P(\theta, \pi | \ldots)$$

...but **our actions produce evidence**, we conclude the optimal policy from our own actions.

- **We cannot change events that causally precede the present.**

Causality

- Solution: treat actions as **causal interventions**

$$
\begin{aligned}
P(\theta, \pi | \hat{a}_t, o_t, \ldots) & \\
&\propto \text{likelihood} \times \text{prior} \\
&= P(o_t | \theta, \hat{a}_t, \ldots) P(\hat{a}_t | \pi, \ldots) \times P(\theta, \pi | \ldots) \\
&= P(o_t | \theta, a_t, \ldots) \times P(\theta, \pi | \ldots)
\end{aligned}
$$

Causality

- Solution: treat actions as **causal interventions**

$$
\begin{aligned}
P(\theta, \pi | \hat{a}_t, o_t, \ldots) \\
&\propto \text{likelihood} \times \text{prior} \\
&= P(o_t | \theta, \hat{a}_t, \ldots) P(\hat{a}_t | \pi, \ldots) \times P(\theta, \pi | \ldots) \\
&= P(o_t | \theta, a_t, \ldots) \times P(\theta, \pi | \ldots)
\end{aligned}
$$

- Causal intervention informs us that we have to **ignore the evidence** produced by the action.

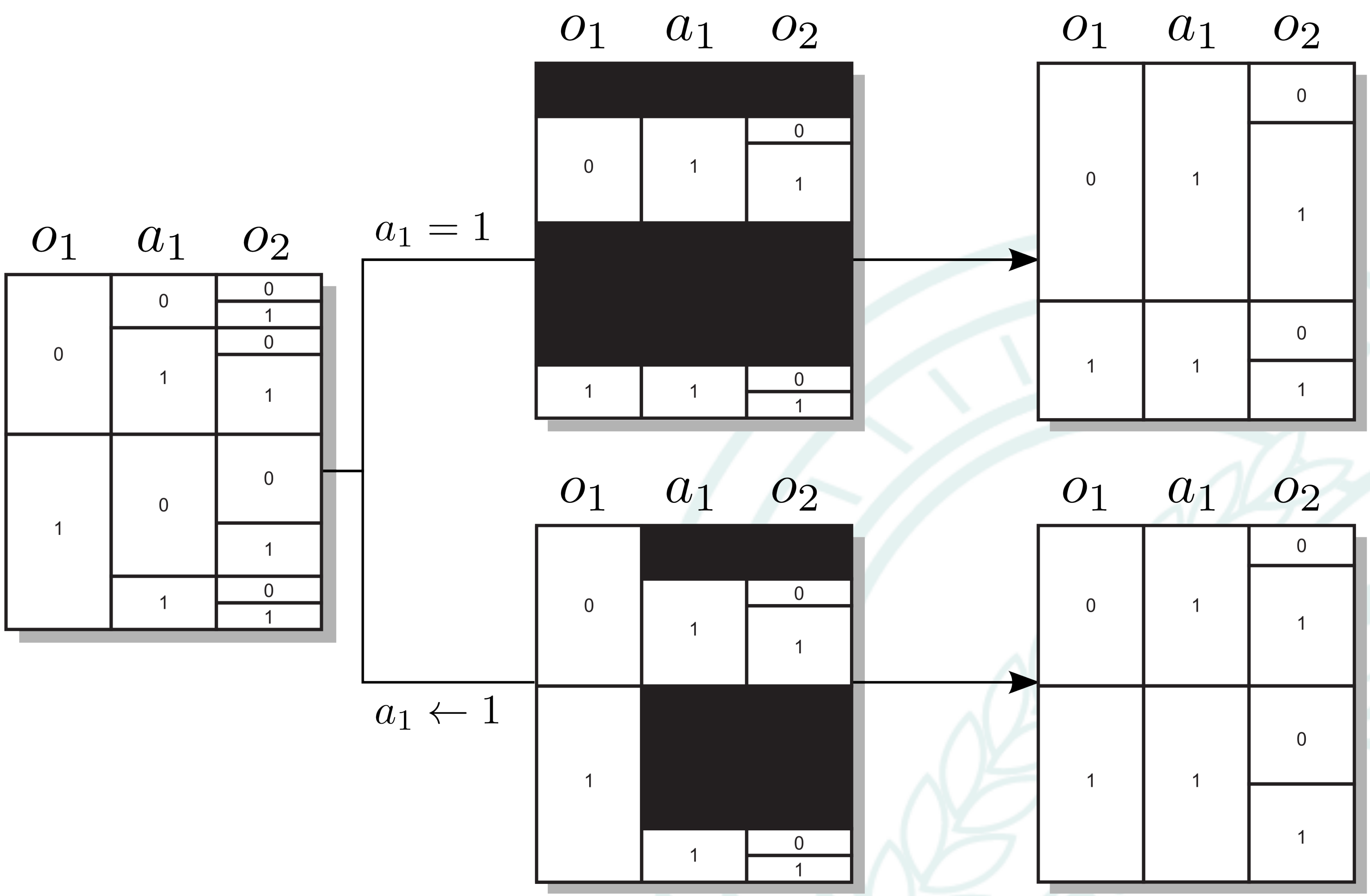
Causality

- Solution: treat actions as **causal interventions**

$$
\begin{aligned}
P(\theta, \pi | \hat{a}_t, o_t, \ldots) \\
&\propto \text{likelihood} \times \text{prior} \\
&= P(o_t | \theta, \hat{a}_t, \ldots) P(\hat{a}_t | \pi, \ldots) \times P(\theta, \pi | \ldots) \\
&= P(o_t | \theta, a_t, \ldots) \times P(\theta, \pi | \ldots)
\end{aligned}
$$

- Causal intervention informs us that we have to **ignore the evidence** produced by the action.
- Caveat: $\pi = \pi(\theta)$

Bayesian versus Causal Update



Summary

Actions are produced by the agent itself and thus need to be treated as causal interventions.

Bayesian Control Rule

Bayesian Control Rule

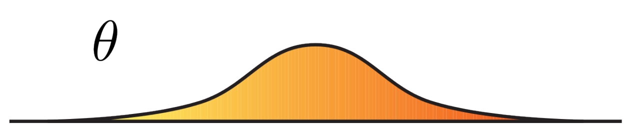
Given a set Θ of

- behaviors $P(a_{\leq t}, o_{\leq t} | \theta)$
- prior probabilities $P(\theta)$

sample actions from $P(a_t | \hat{a}_{<t}, o_{<t})$

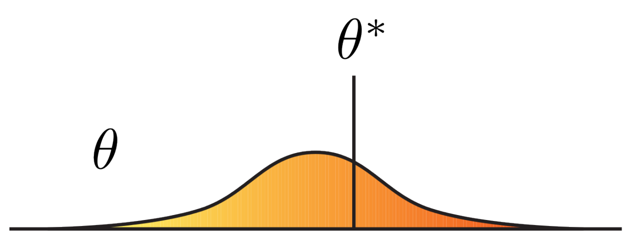
Bayesian Control Rule (cont.)

Time t



Prior:

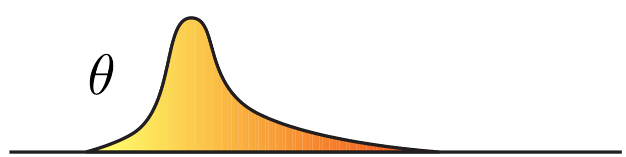
$P(\theta|\hat{a}_{<t}, o_{<t})$



Acting:

$\theta^* \sim P(\theta|\hat{a}_{<t}, o_{<t})$

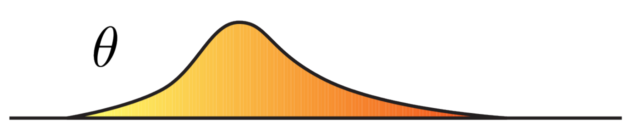
$a_t^* \sim P(a_t|\theta^*, \hat{a}_{<t}, o_{<t})$



Observing:

$$P(\theta|\hat{a}_{\leq t}, o_{\leq t}) \propto P(\theta|\hat{a}_{\leq t}, o_{\leq t}) P(o_t|\theta, a_{\leq t}, o_{<t})$$

Time t+1



Posterior:

$P(\theta|\hat{a}_{\leq t}, o_{\leq t})$

Example: 2-Armed Bandit

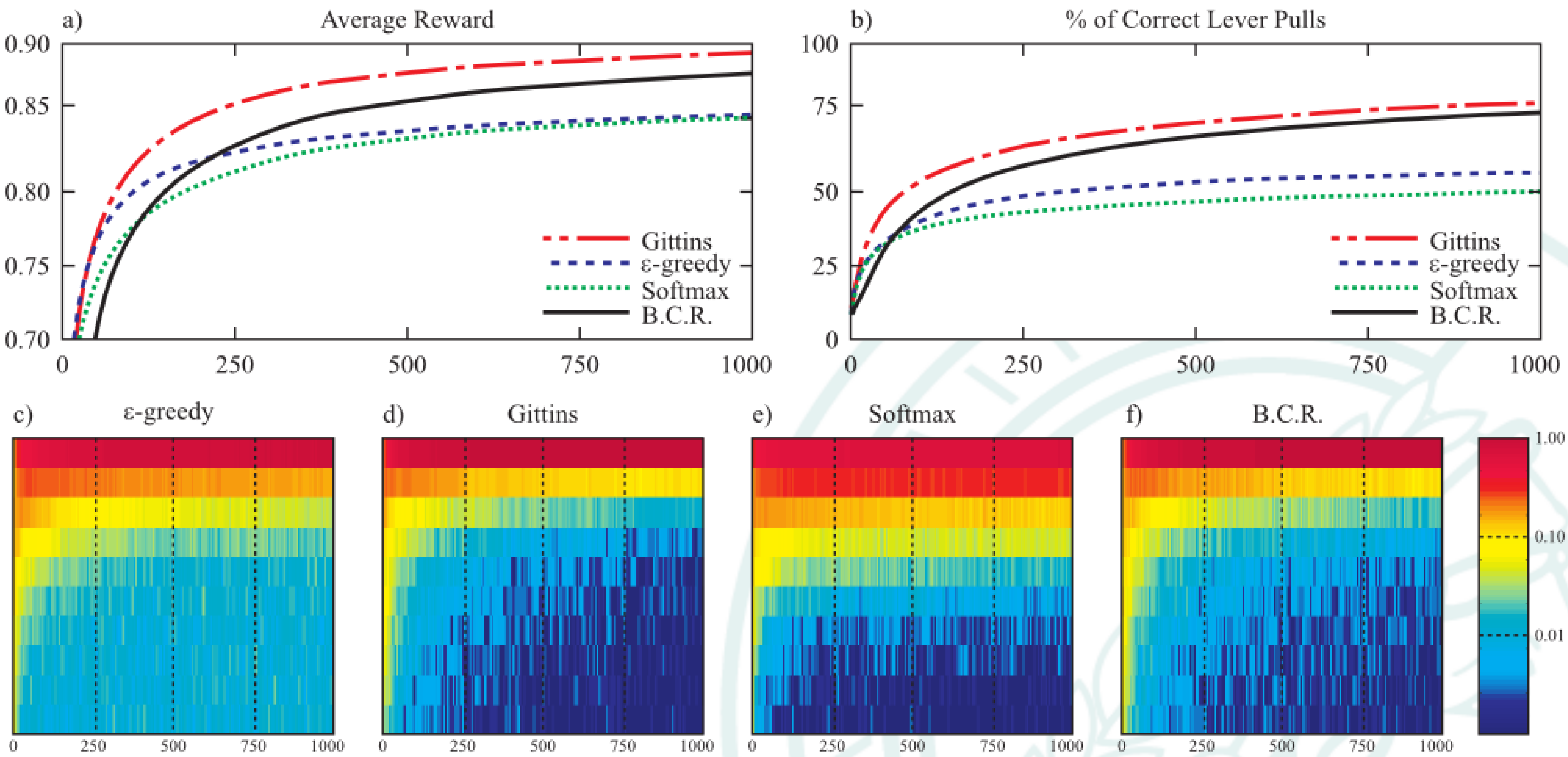
- Bernoulli-distributed rewards, unknown biases.
- Hypotheses: $\Theta = [0,1] \times [0,1]$
- Prior: $P(\theta) = U(0,1) \times U(0,1)$
- Observations: $P(o|\theta, a) = B(o; \theta_a)$
- Actions: $P(a|\theta) = \delta_a^{\arg\max_i \theta_i}$

Example: 2-Armed Bandit

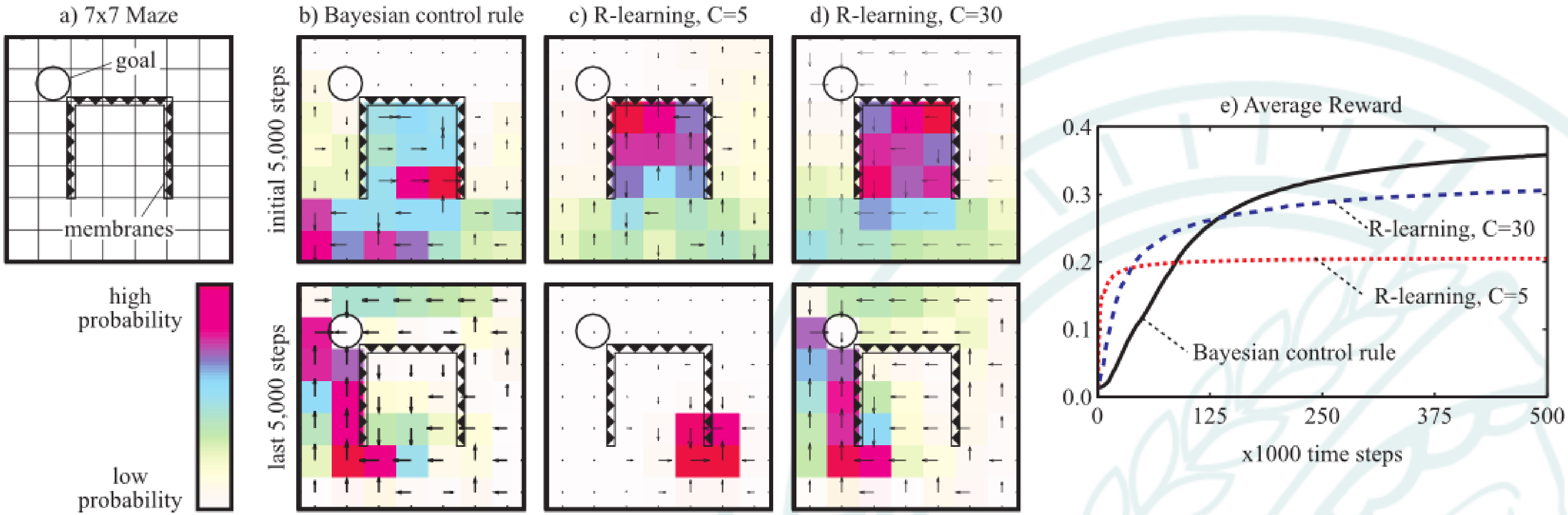
- Bernoulli-distributed rewards, unknown biases.
- Hypotheses: $\Theta = [0,1] \times [0,1]$
- Prior: $P(\theta) = U(0,1) \times U(0,1)$
- Observations: $P(o|\theta,a) = B(o;\theta_a)$
- Actions: $P(a|\theta) = \delta_a^{\arg\max_i \theta_i}$

- Recently proven to be **asymptotically optimal** [Kaufmann, Korda, Munos 2012].

Results for 10-Armed Bandit

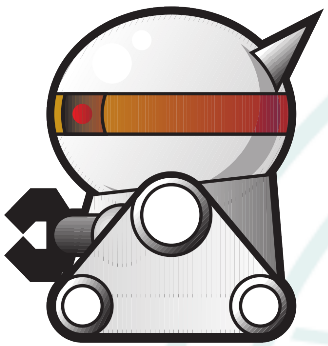


Markov Decision Processes

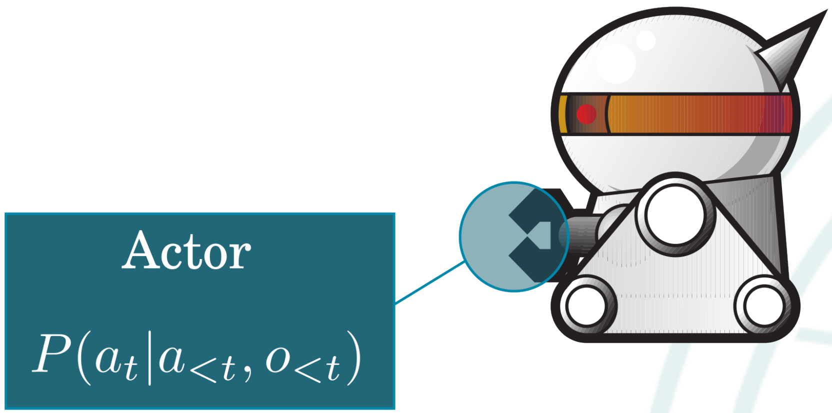


Conclusions

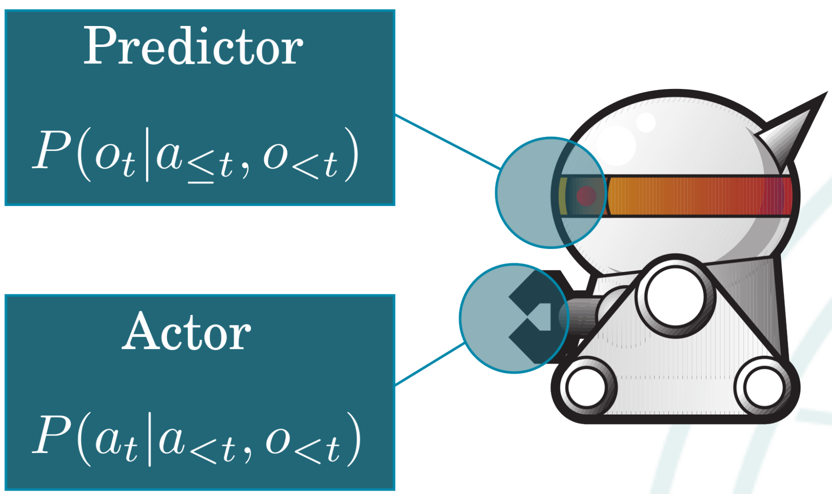
Blueprint of the Agent



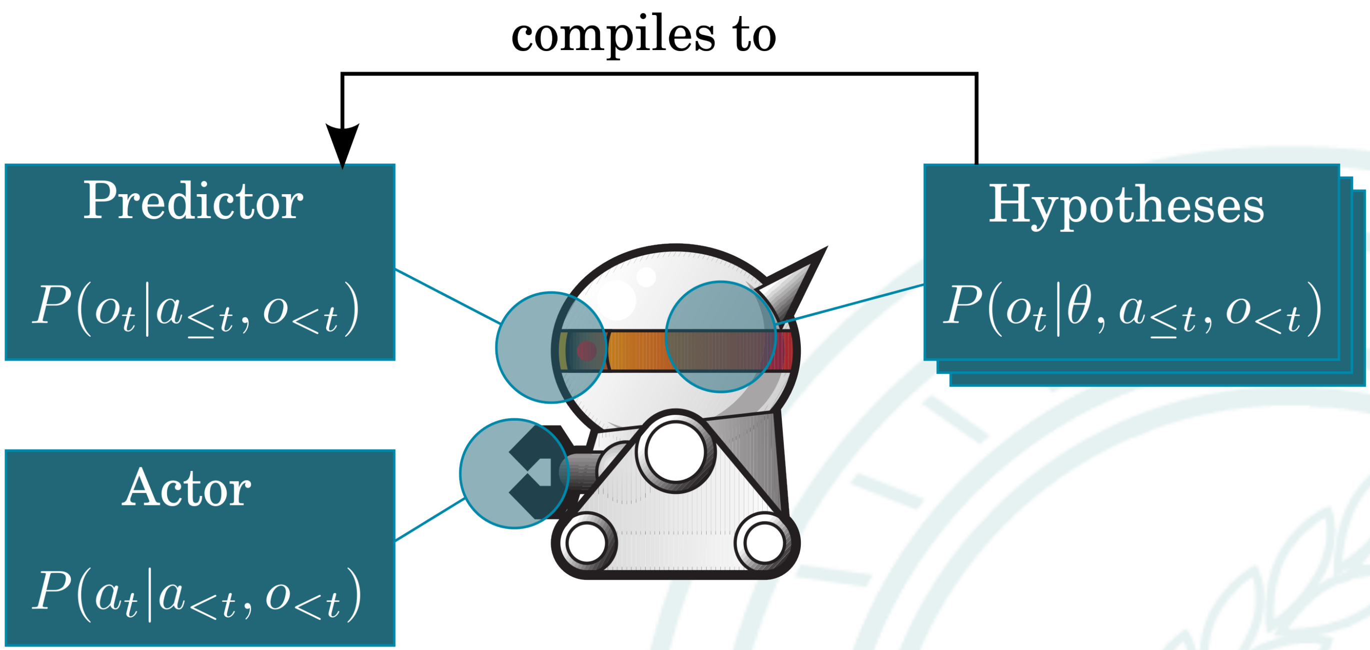
Blueprint of the Agent



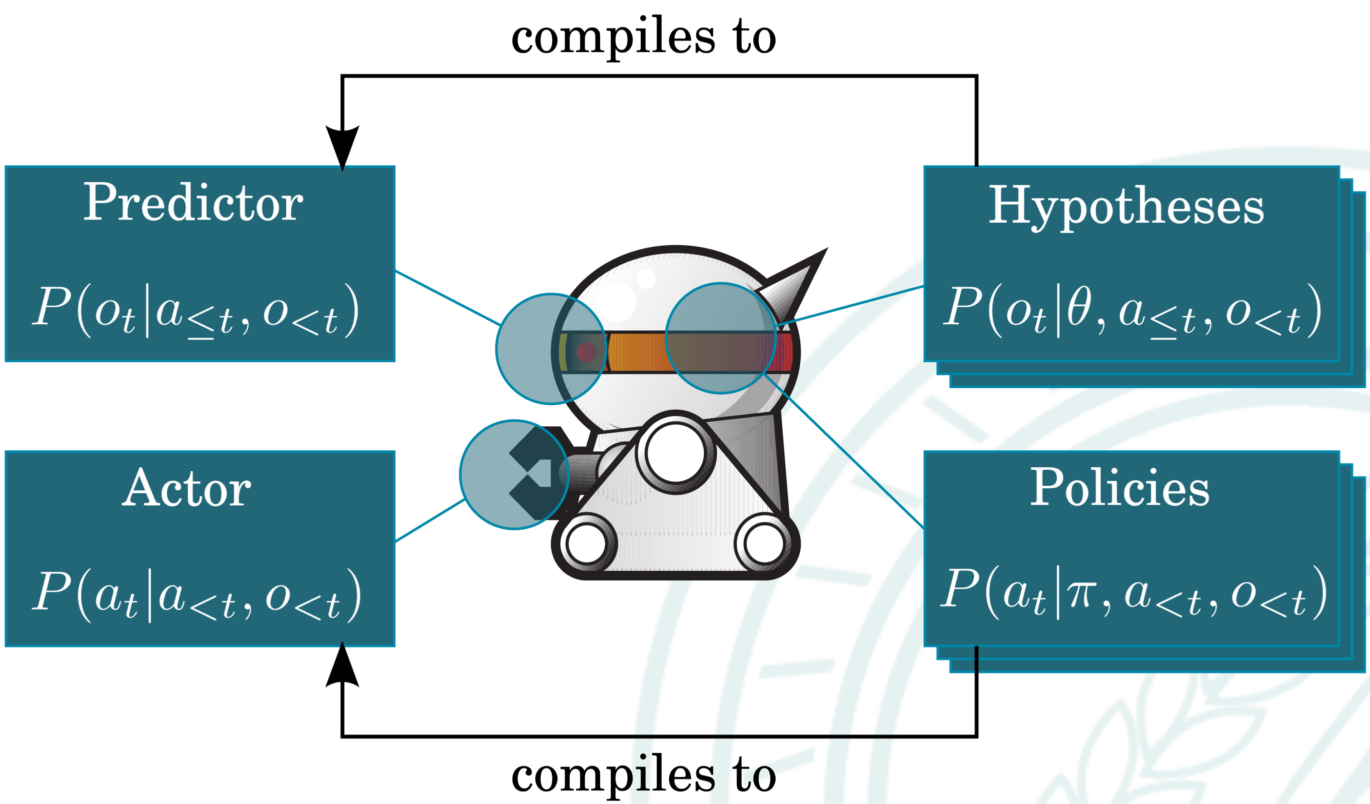
Blueprint of the Agent



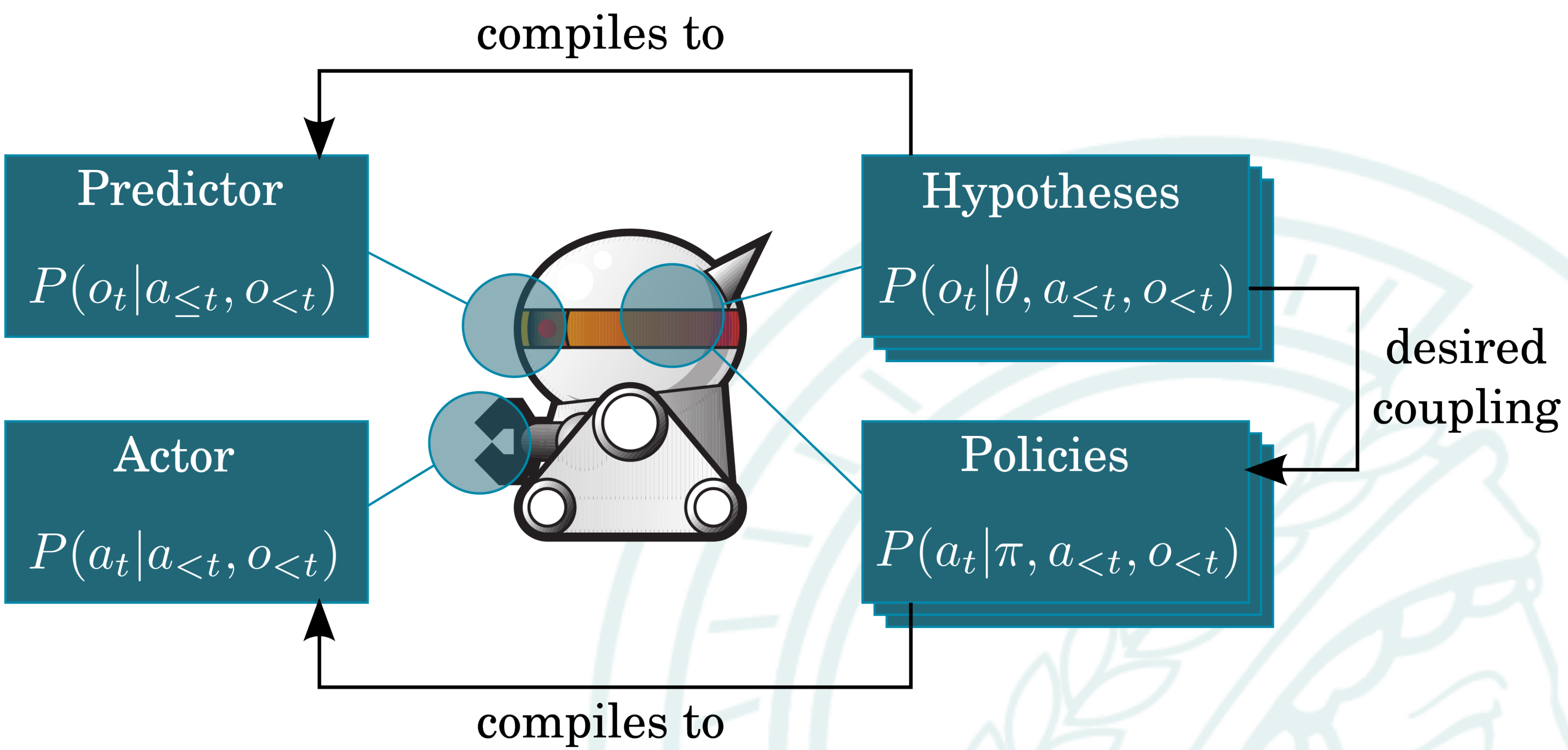
Blueprint of the Agent



Blueprint of the Agent



Blueprint of the Agent



Properties

- Stochastic controller that **refines its policy with experience.**
- Ingredients: **Bayes + Causality**.
- Transforms control into inference.
- Related to **Random Beliefs** & **Thompson sampling.**
- Allows tackling **game-theoretic** problems.
- Exploits **built-in reward mechanism** of Bayes' rule.
- Works also with **complex causal models**.

Pros and Cons

Pros

- Simple and general.
- Converges to desired behavior in “ergodic” tasks.
- Suitable for on-line.
- Trades-off exploration versus exploitation.
- Automatic temporal credit assignment.

Cons

- Sub-optimal in the transient.
- Does not converge in non-ergodic environments.
- Convergence speed highly depends on environment.
- Design of behaviors can be difficult.