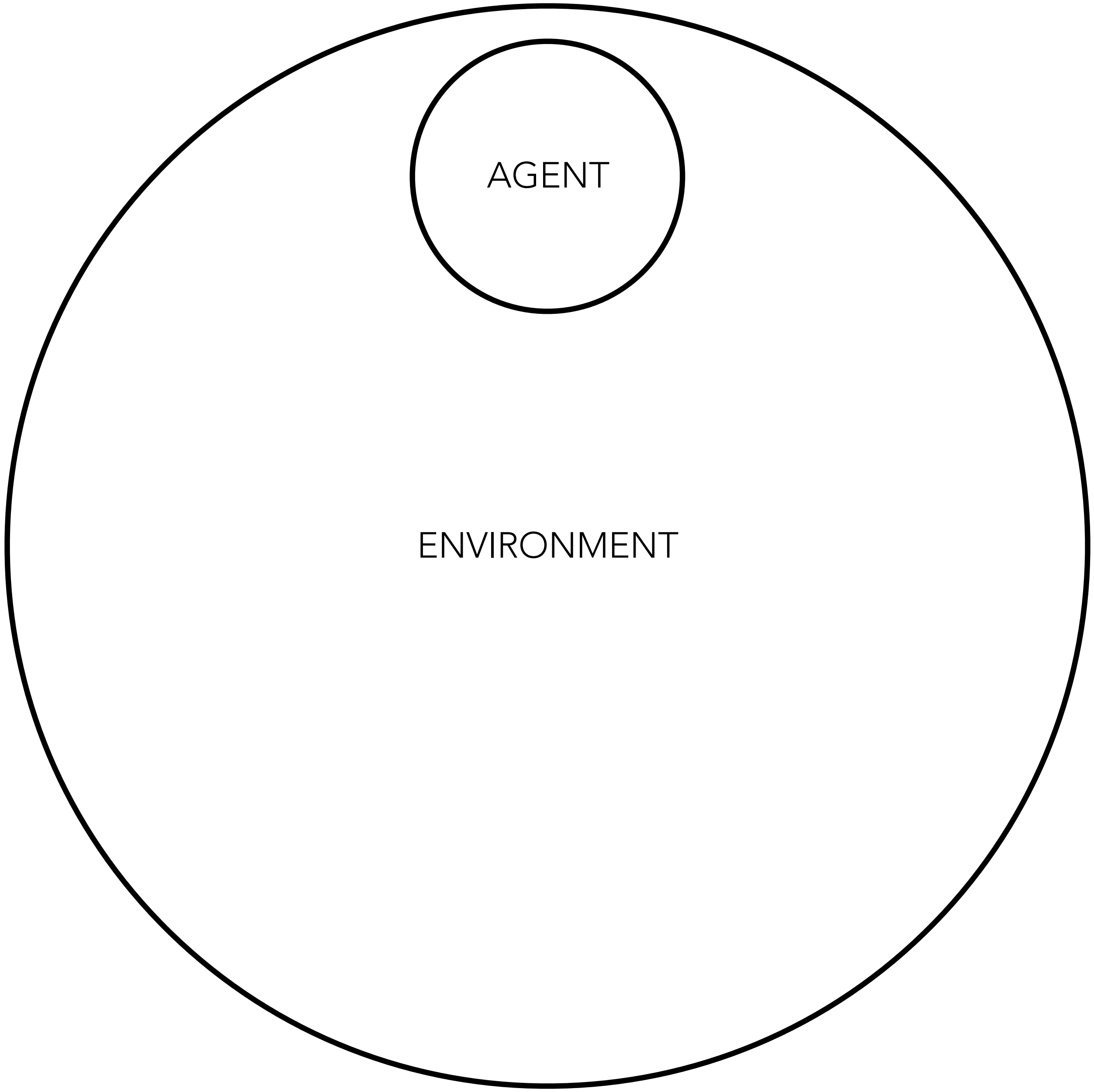
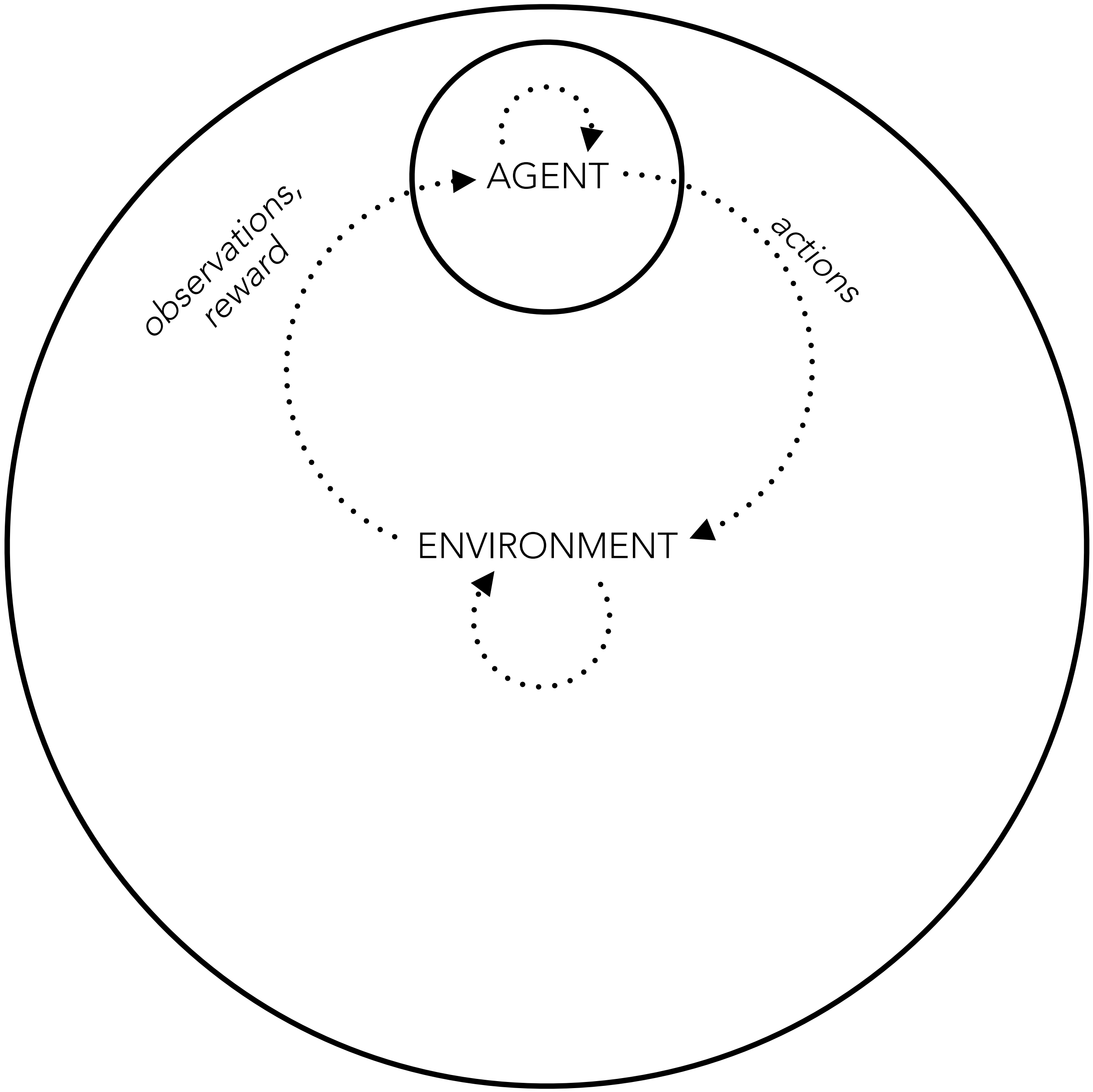
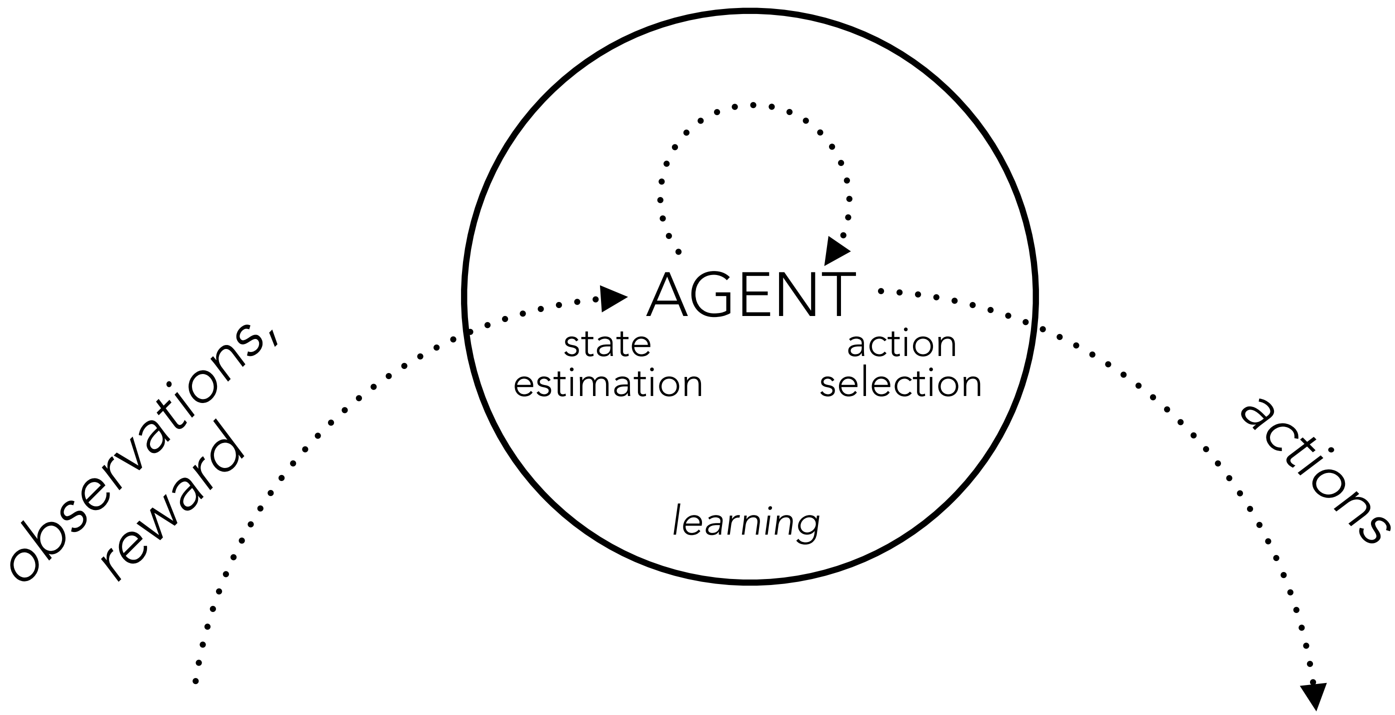
An Inference Perspective on Model-Based Reinforcement Learning

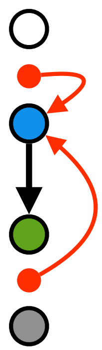
Joseph Marino, Yisong Yue

Caltech


AGENT
ENVIRONMENT

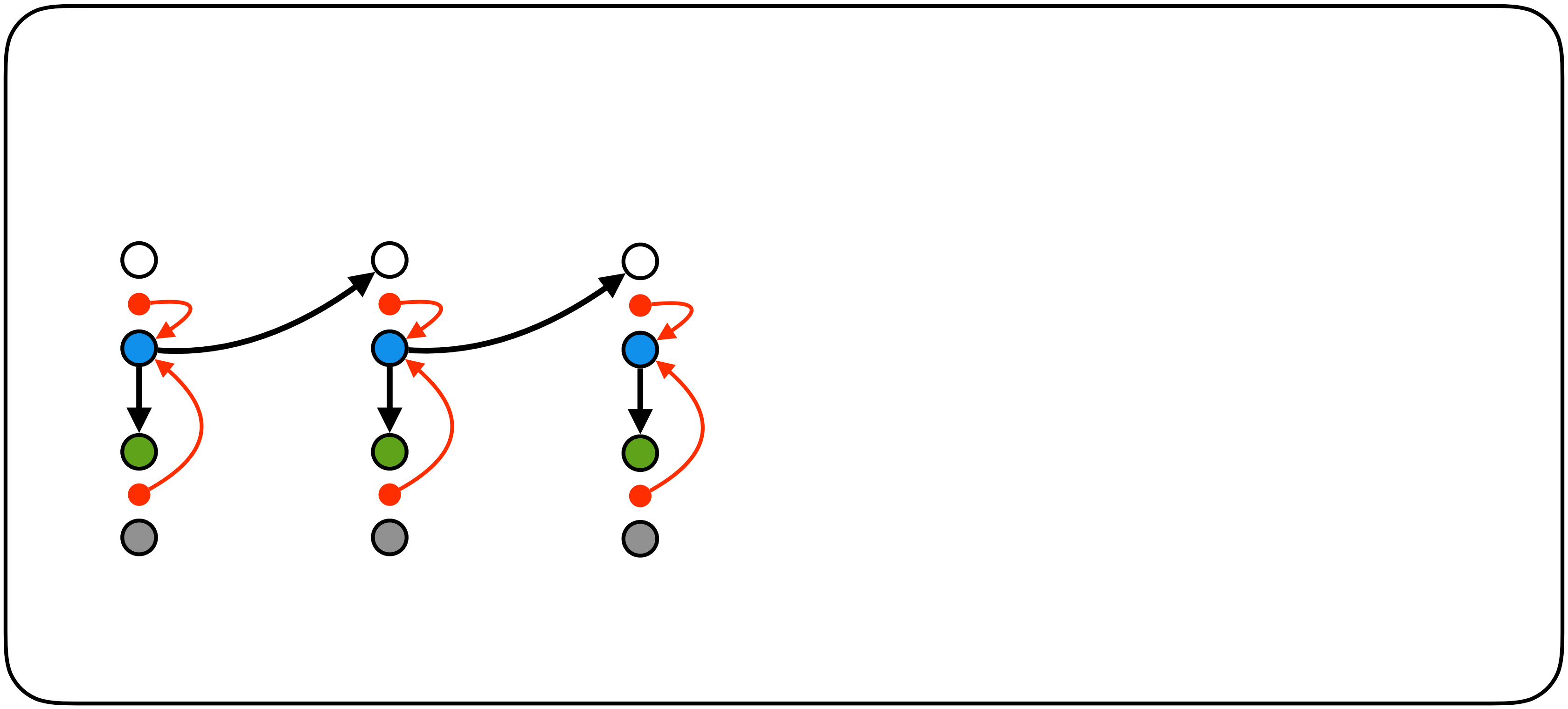

AGENT
observations,
reward
actions
ENVIRONMENT


AGENT
state
estimation
action
selection
learning
observations,
reward
actions



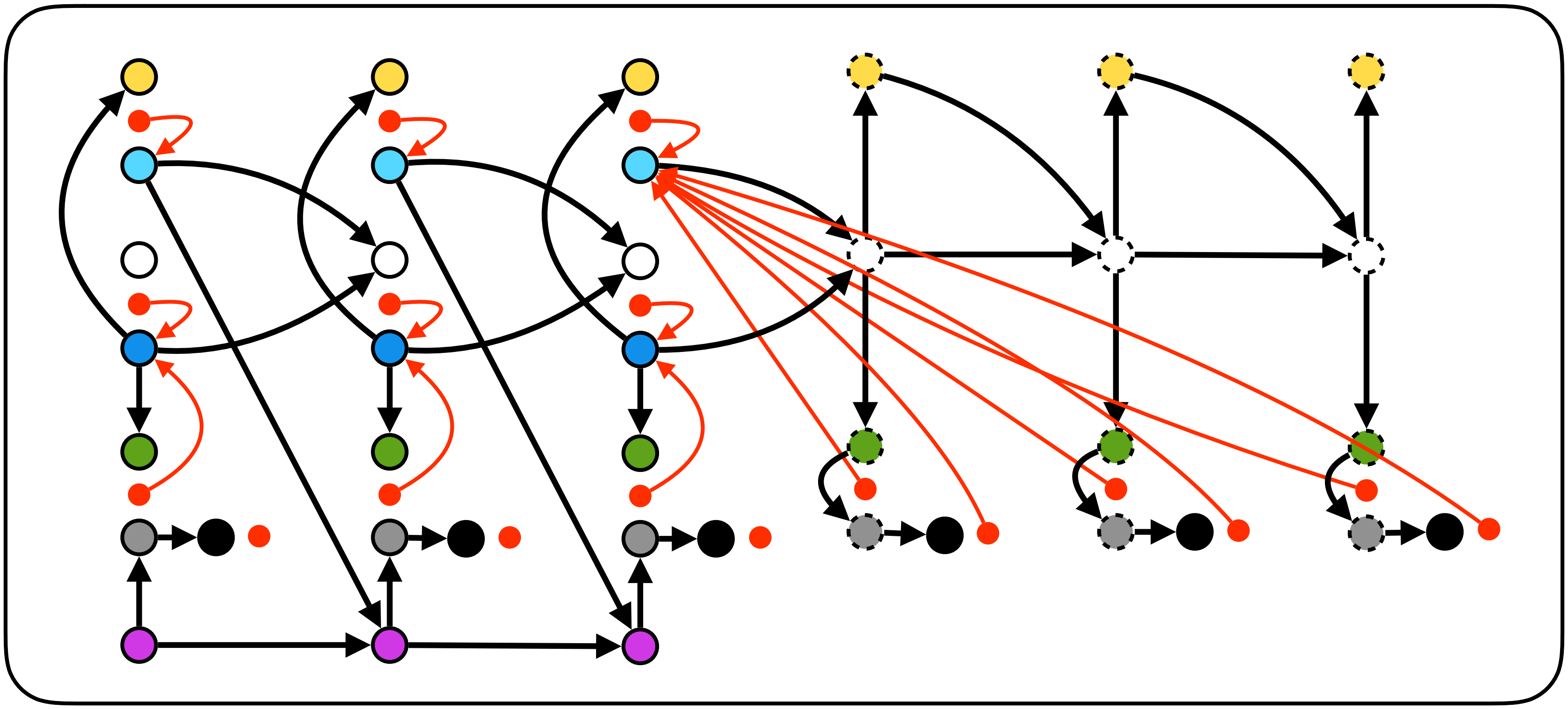
Iterative
Amortized Inference

Marino, *et al.*, 2018a



Iterative Amortized Inference
Marino, *et al.*, 2018a

Amortized Variational Filtering
Marino, *et al.*, 2018b



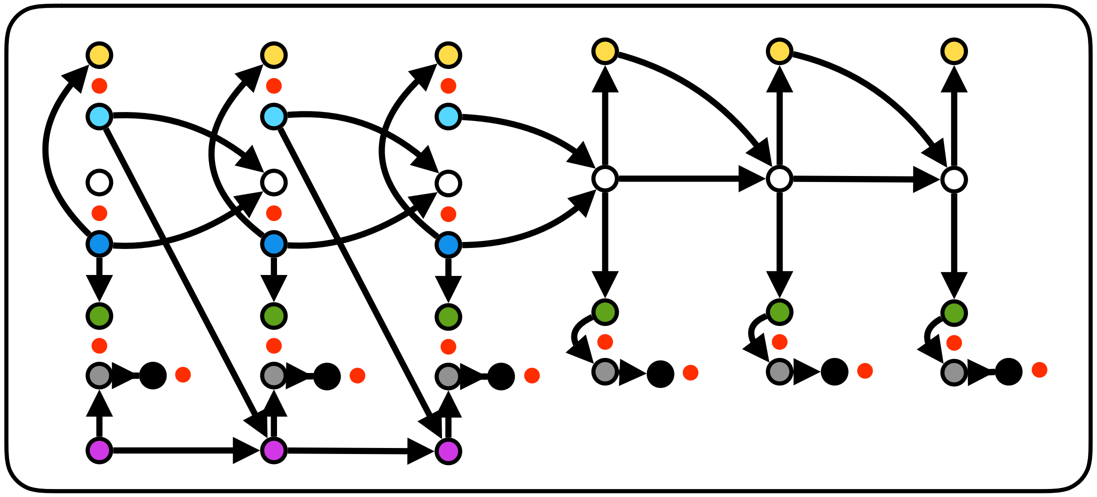
Iterative
Amortized Inference
Marino, *et al.*, 2018a

Amortized
Variational Filtering
Marino, *et al.*, 2018b

An Inference Perspective
on Model-Based RL
Marino & Yue, 2019

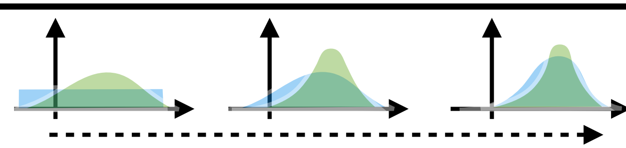
OVERVIEW

frame model-based RL as probabilistic inference & learning



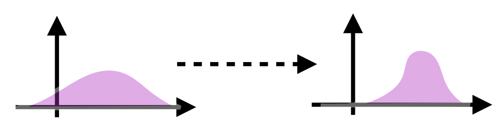
two additional terms:

prior over actions



combine <u>model-based likelihood</u> with a <u>model-free prior</u>

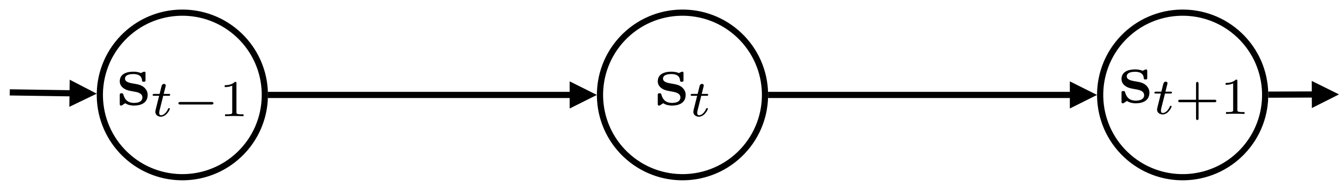
information gain from observations



model task-relevant state information, biases planning

REINFORCEMENT LEARNING AS INFERENCE

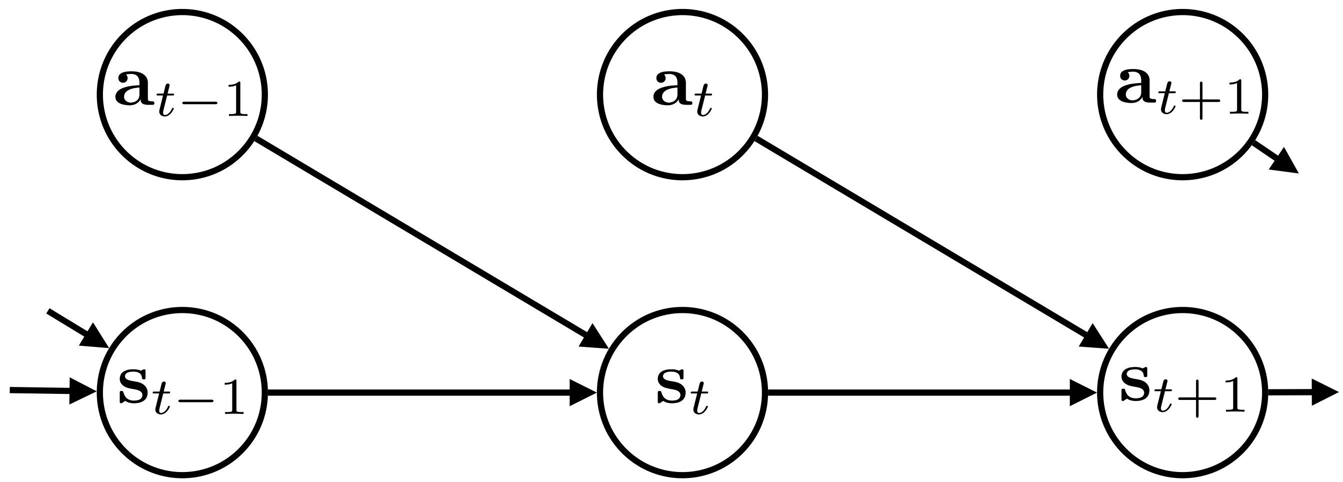
reformulate RL as a probabilistic inference problem



Levine, 2018

REINFORCEMENT LEARNING AS INFERENCE

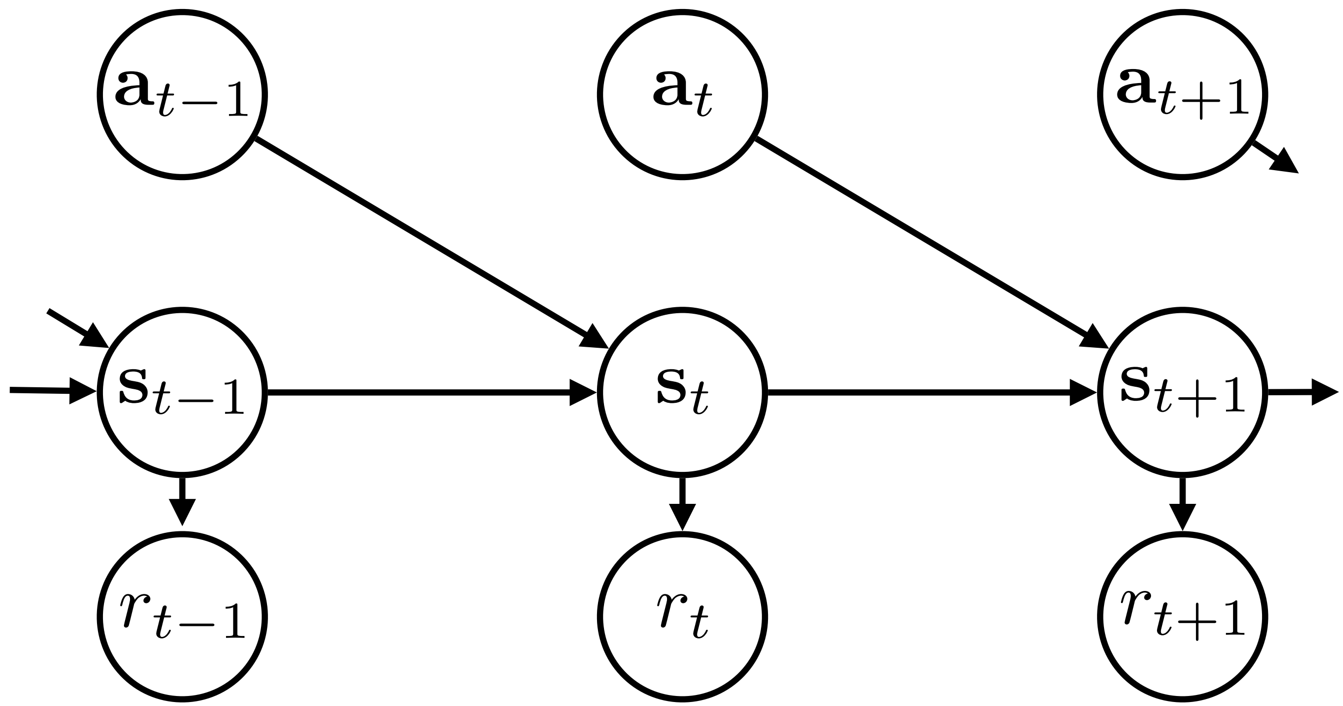
reformulate RL as a probabilistic inference problem



Levine, 2018

REINFORCEMENT LEARNING AS INFERENCE

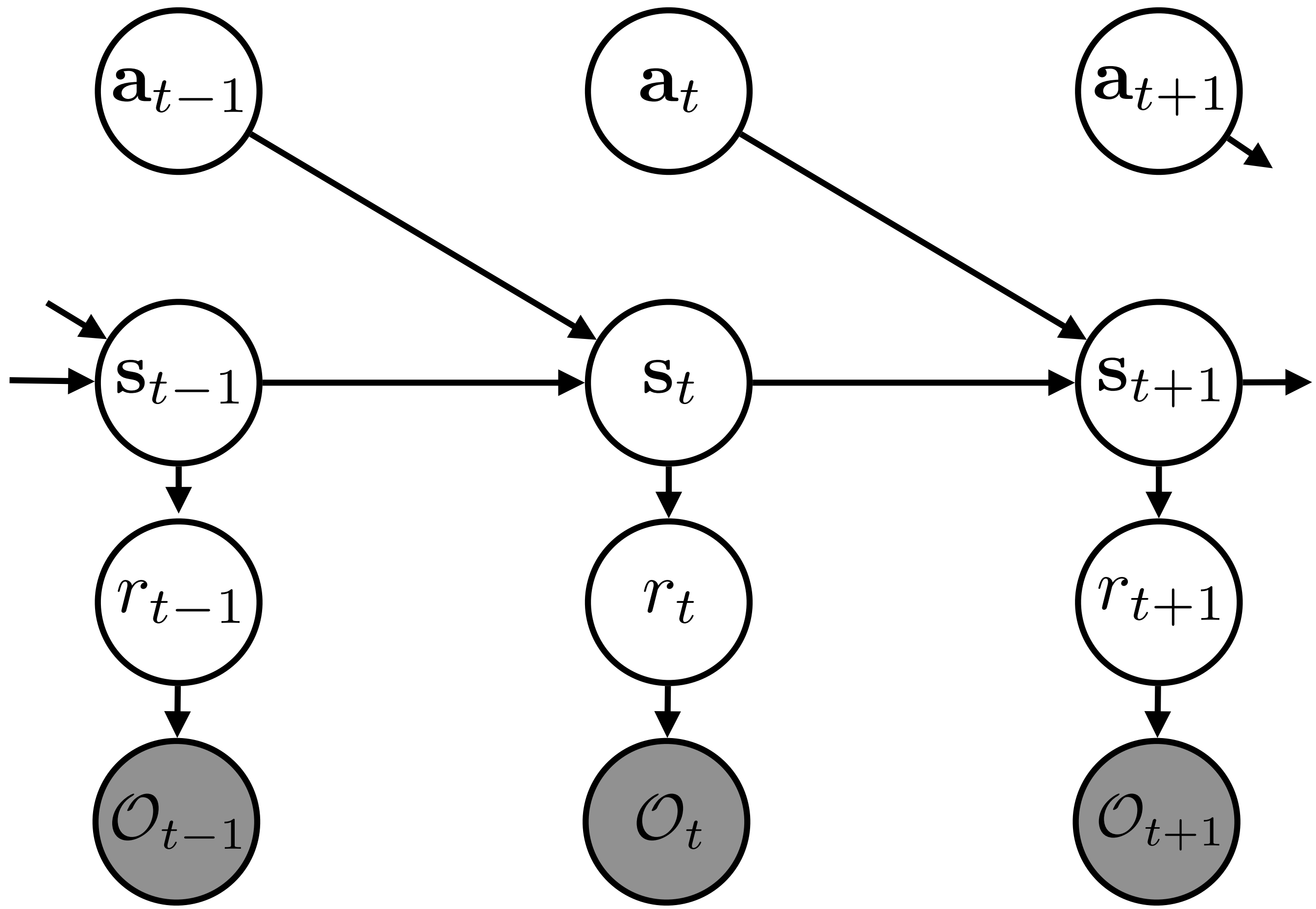
reformulate RL as a probabilistic inference problem



Levine, 2018

REINFORCEMENT LEARNING AS INFERENCE

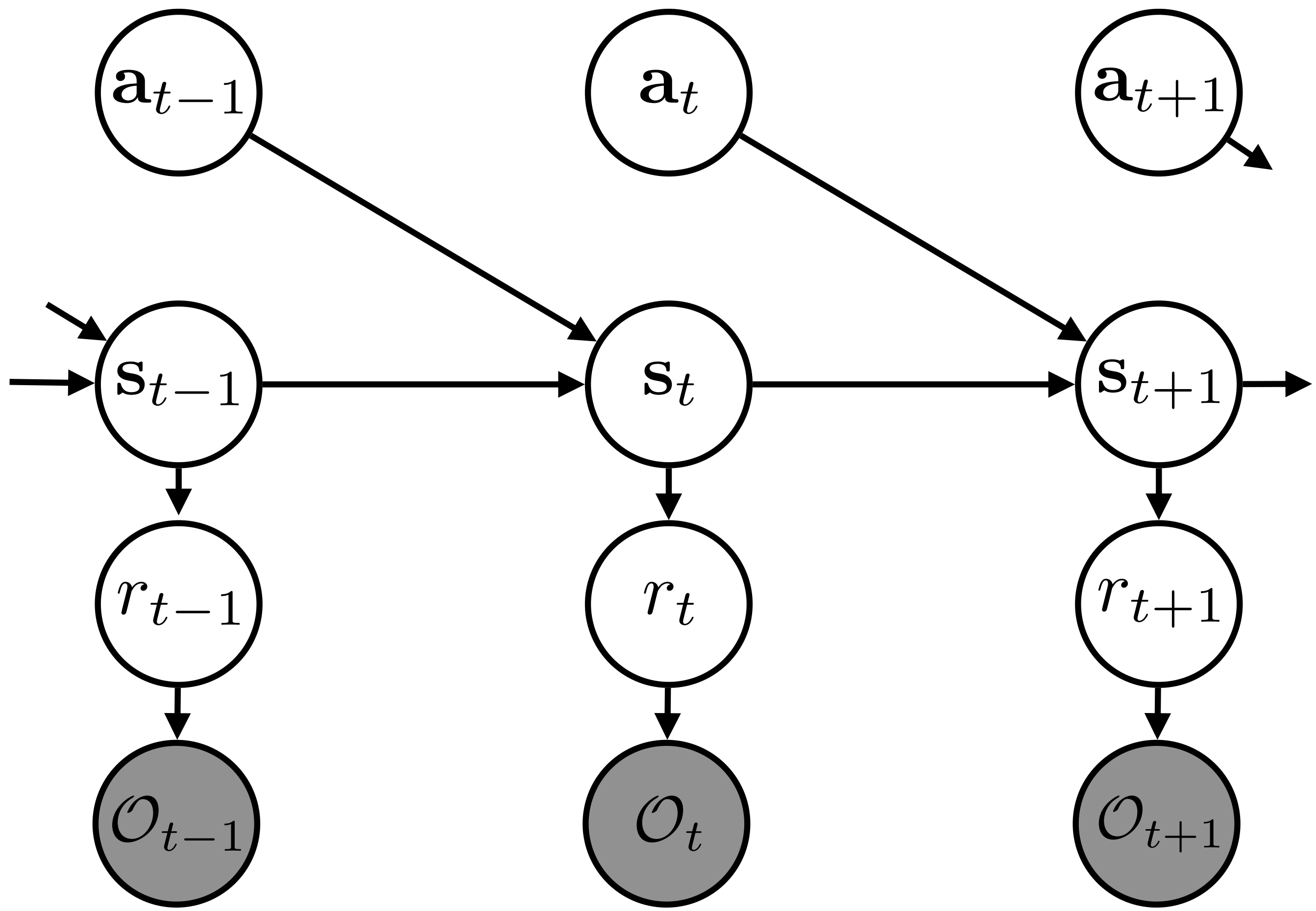
reformulate RL as a probabilistic inference problem



Levine, 2018

PLANNING AS INFERENCE

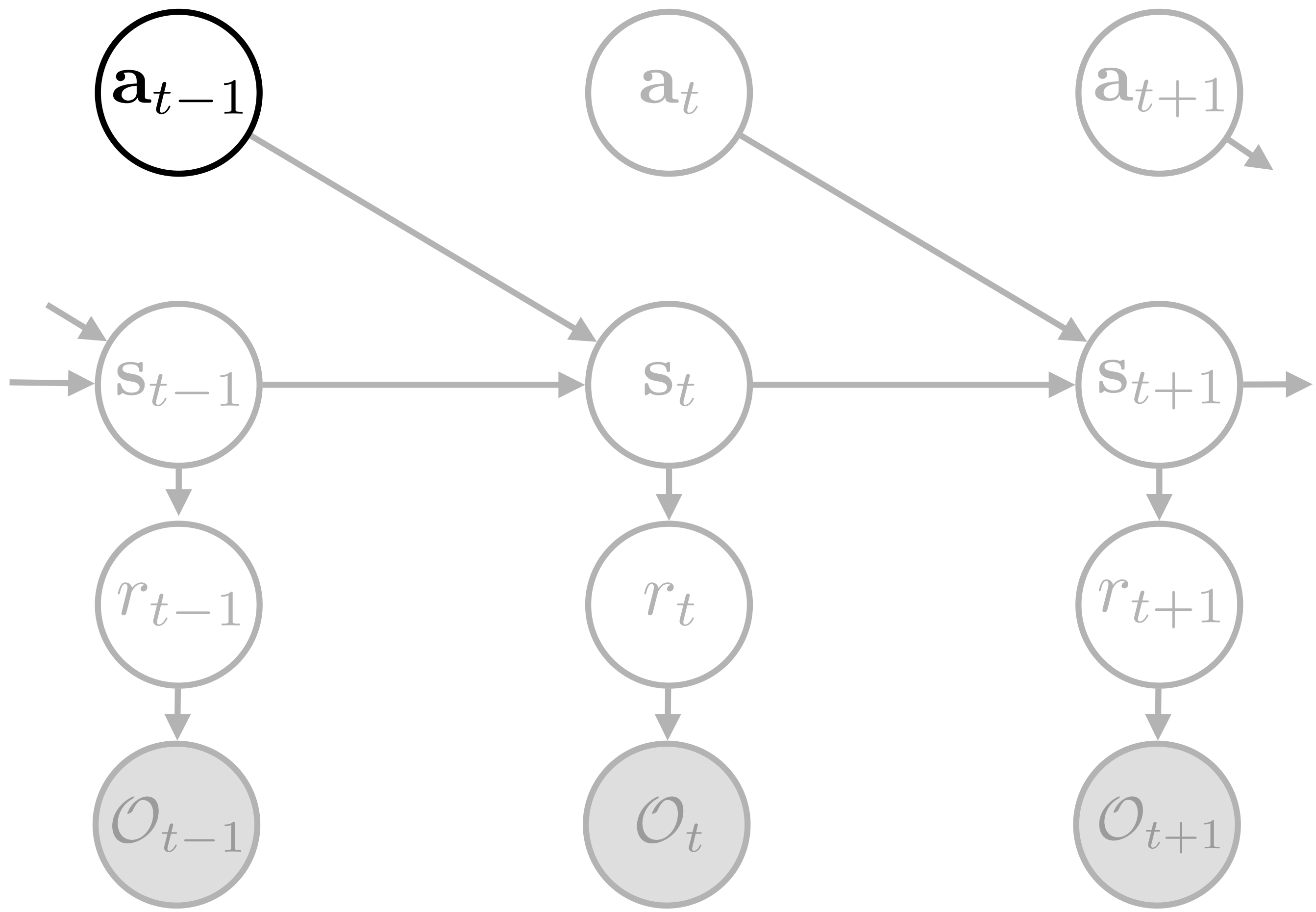
can also reformulate planning as inference



Attias, 2003
Botvinick & Toussaint, 2012

PLANNING AS INFERENCE

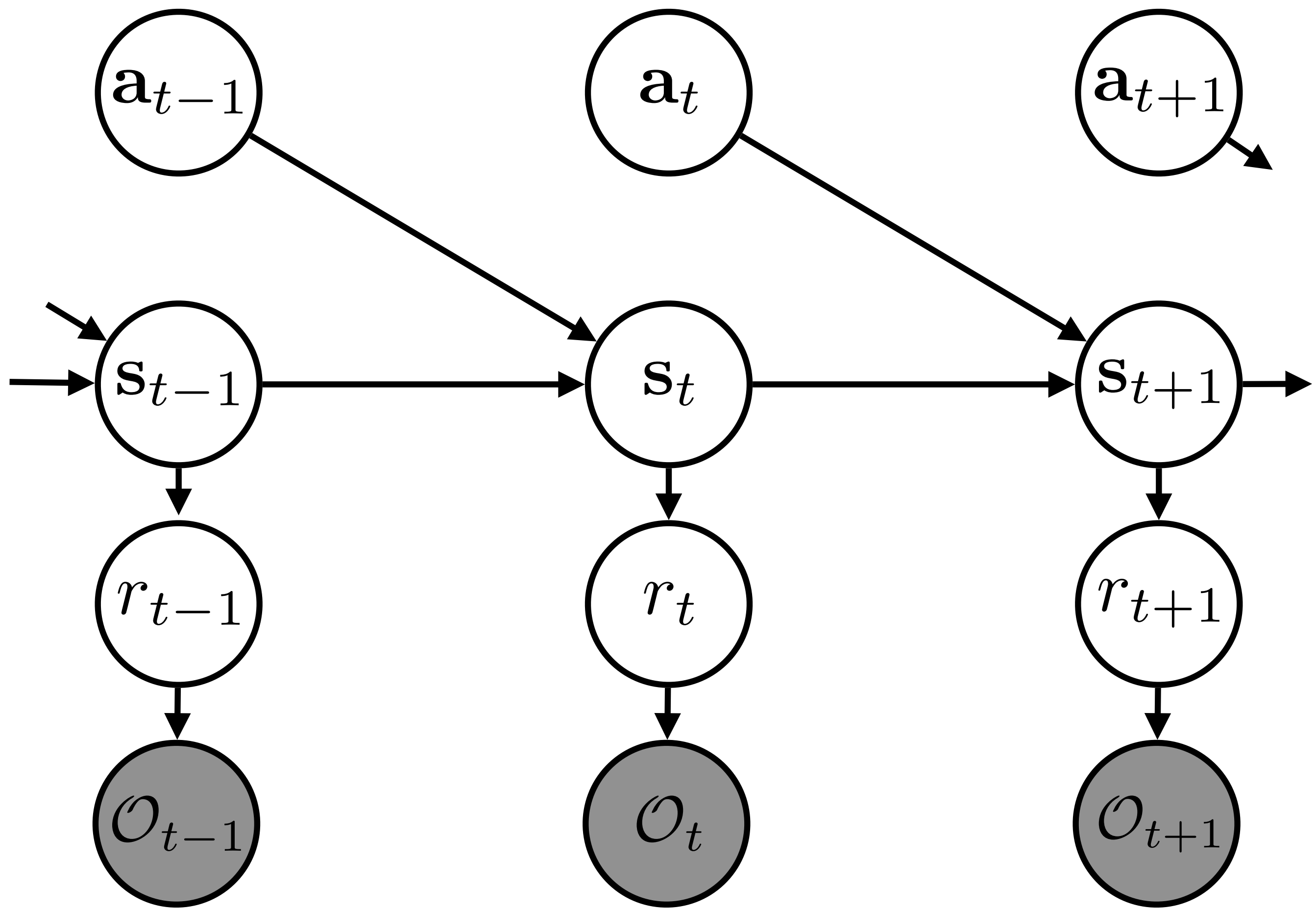
can also reformulate planning as inference



Attias, 2003
Botvinick & Toussaint, 2012

PLANNING AS INFERENCE

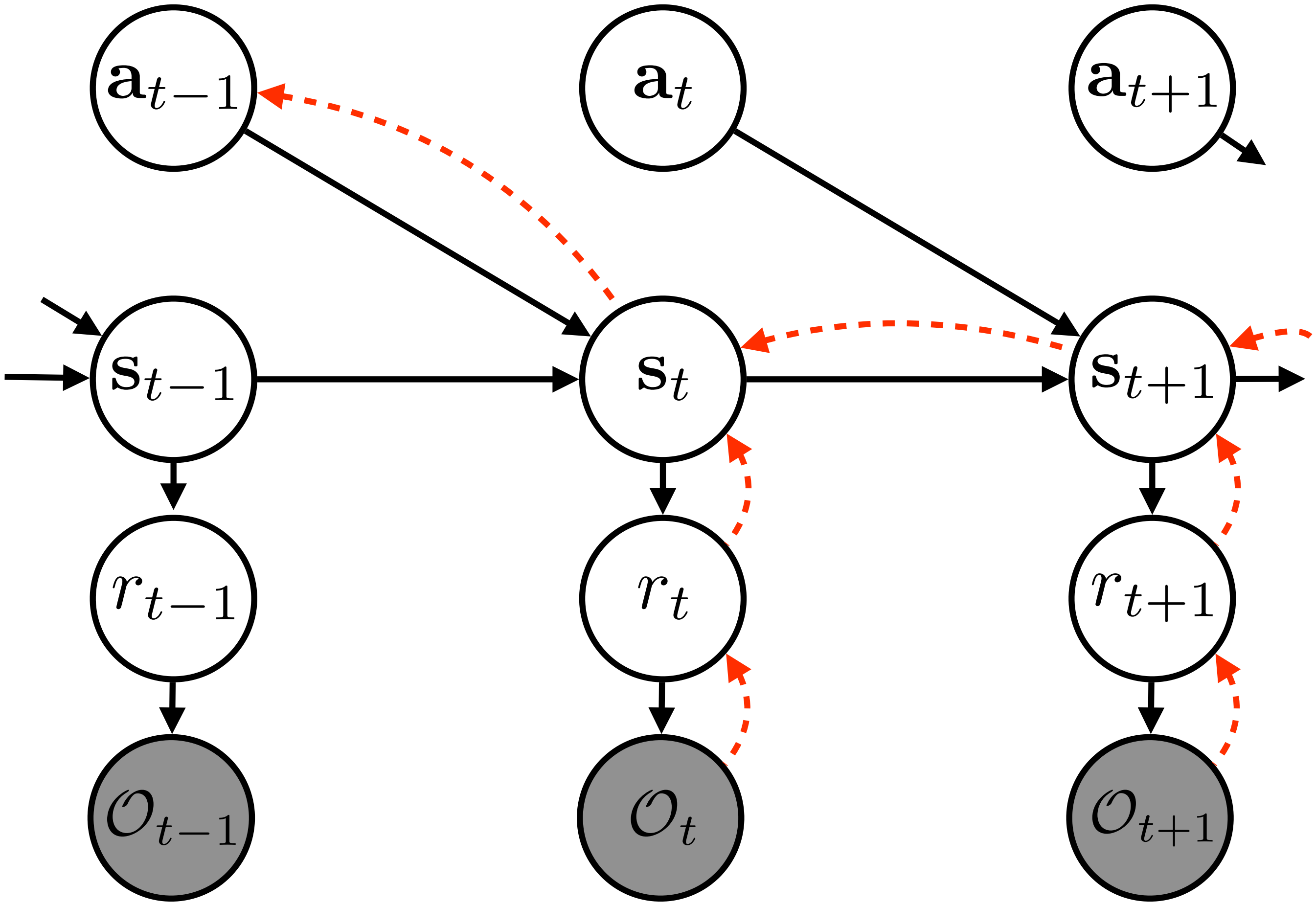
can also reformulate planning as inference



Attias, 2003
Botvinick & Toussaint, 2012

PLANNING AS INFERENCE

can also reformulate planning as inference

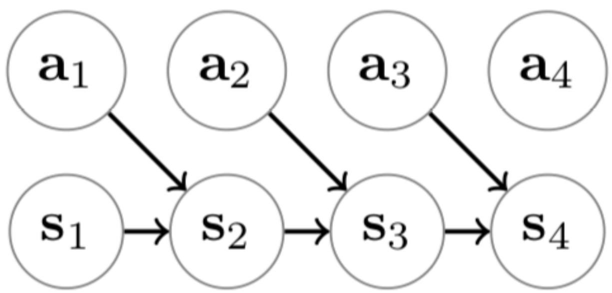


Attias, 2003
Botvinick & Toussaint, 2012

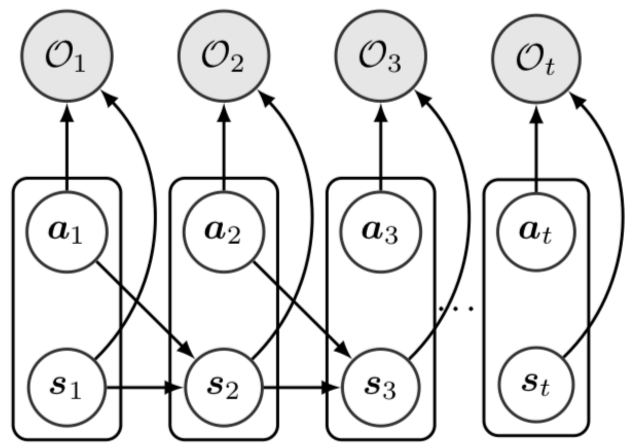
PREVIOUS WORKS

model priors are not learned

focused on fully-observable environments

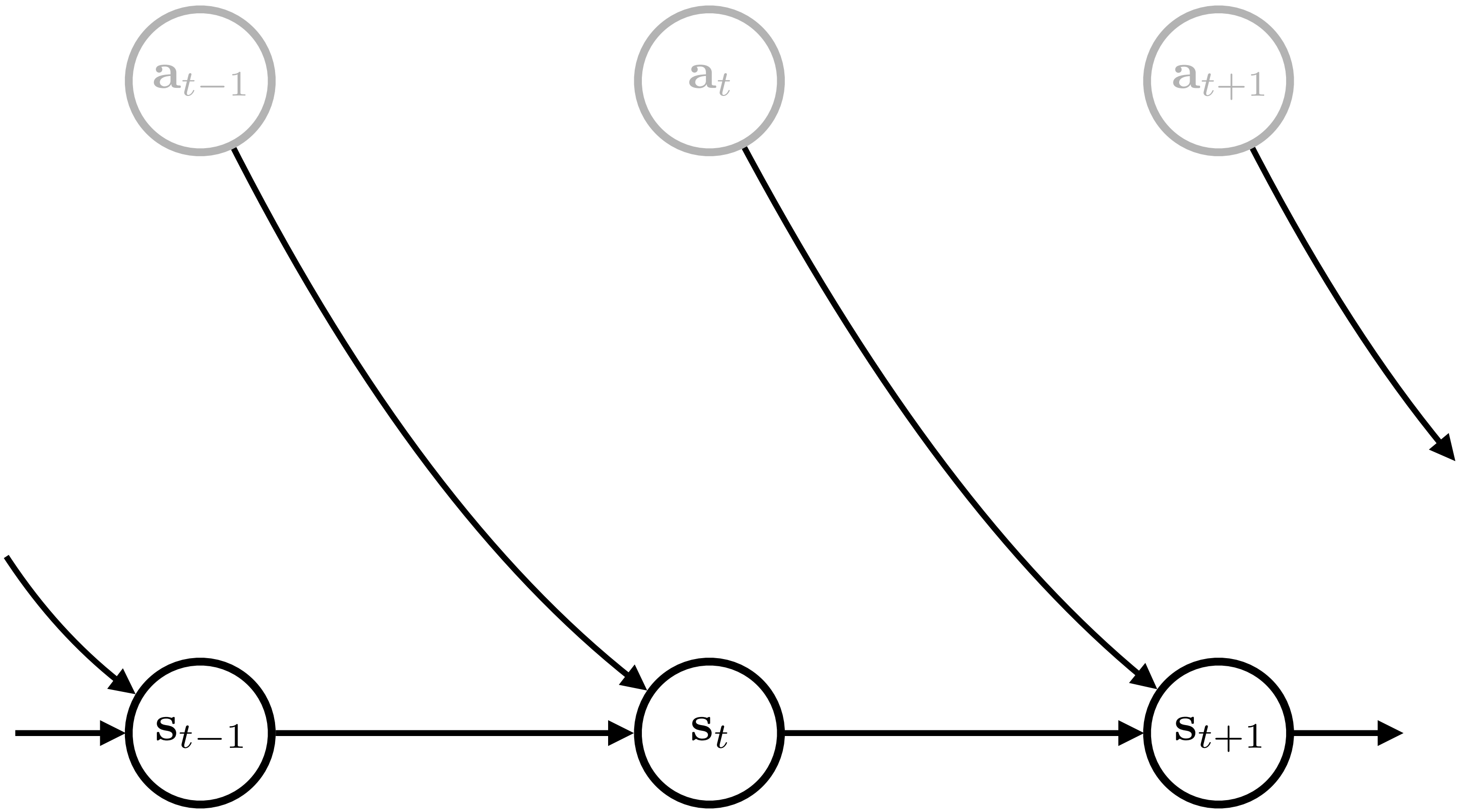


Levine, 2018



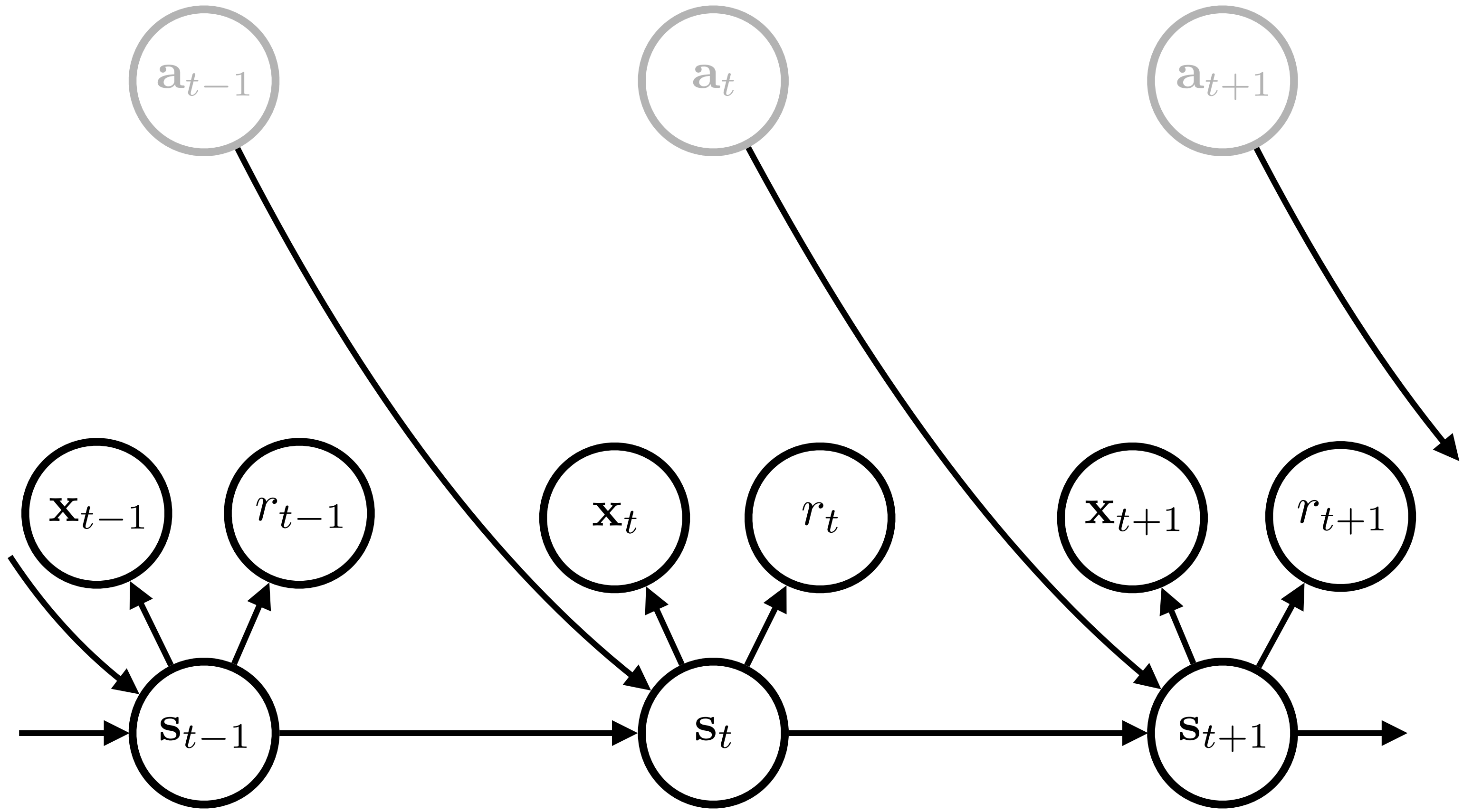
Piché, Thomas, *et al.*, 2019

ENVIRONMENT



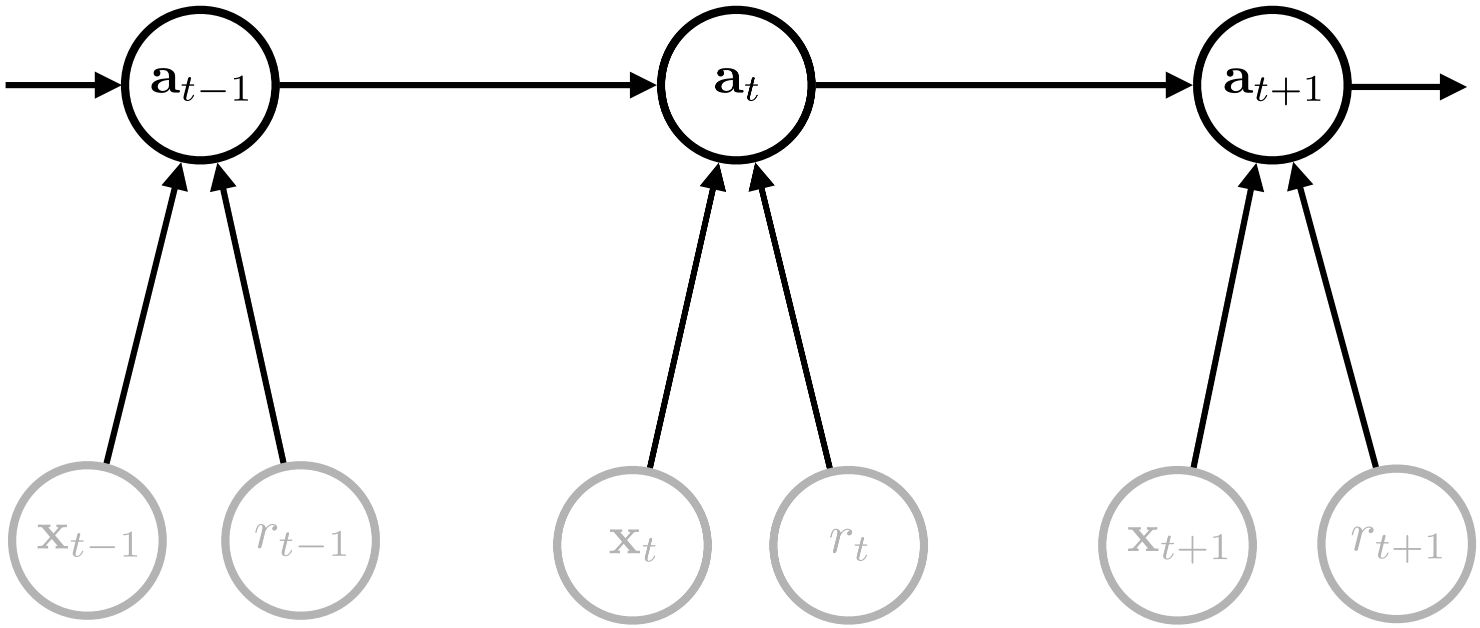
$$p_{\mathrm{e}}(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T} | \mathbf{a}_{1:T-1}) = \prod_{t=1}^{T} \underbrace{p_{\mathrm{e}}(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}_{\text{state dynamics}}$$

ENVIRONMENT



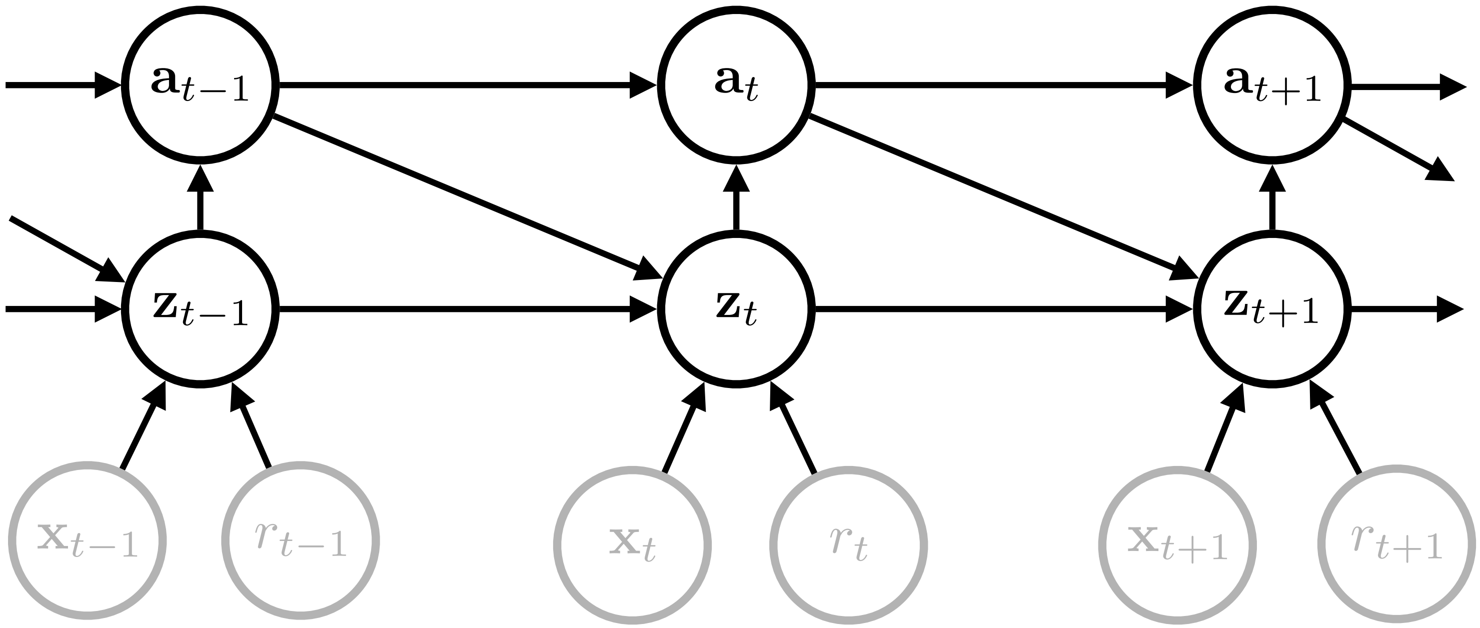
$$p_{\mathrm{e}}(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T} | \mathbf{a}_{1:T-1}) = \prod_{t=1}^{T} \underbrace{p_{\mathrm{e}}(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}_{\text{state dynamics}} \underbrace{p_{\mathrm{e}}(\mathbf{x}_t | \mathbf{s}_t) p_{\mathrm{e}}(r_t | \mathbf{s}_t)}_{\text{observations/rewards}}$$

AGENT



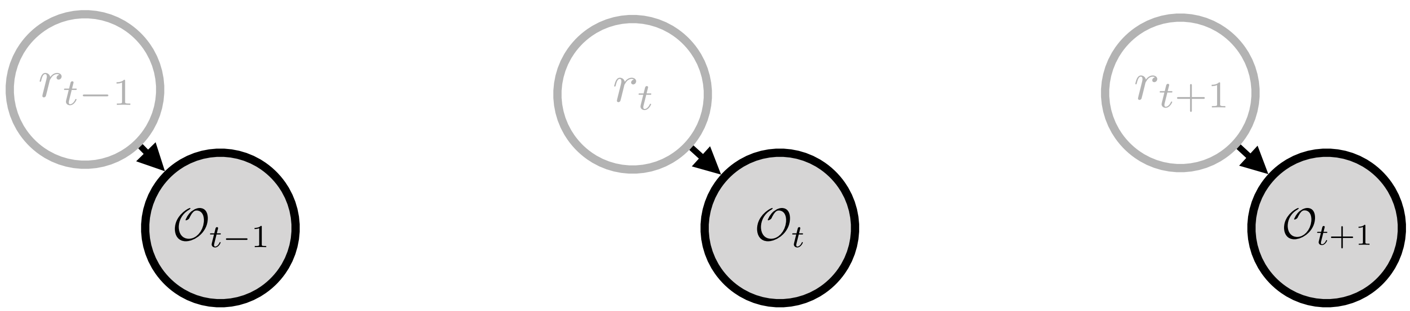
$$p_{\mathrm{a}}(\mathbf{a}_{1:T}|\mathbf{x}_{1:T}, r_{1:t})$$

AGENT



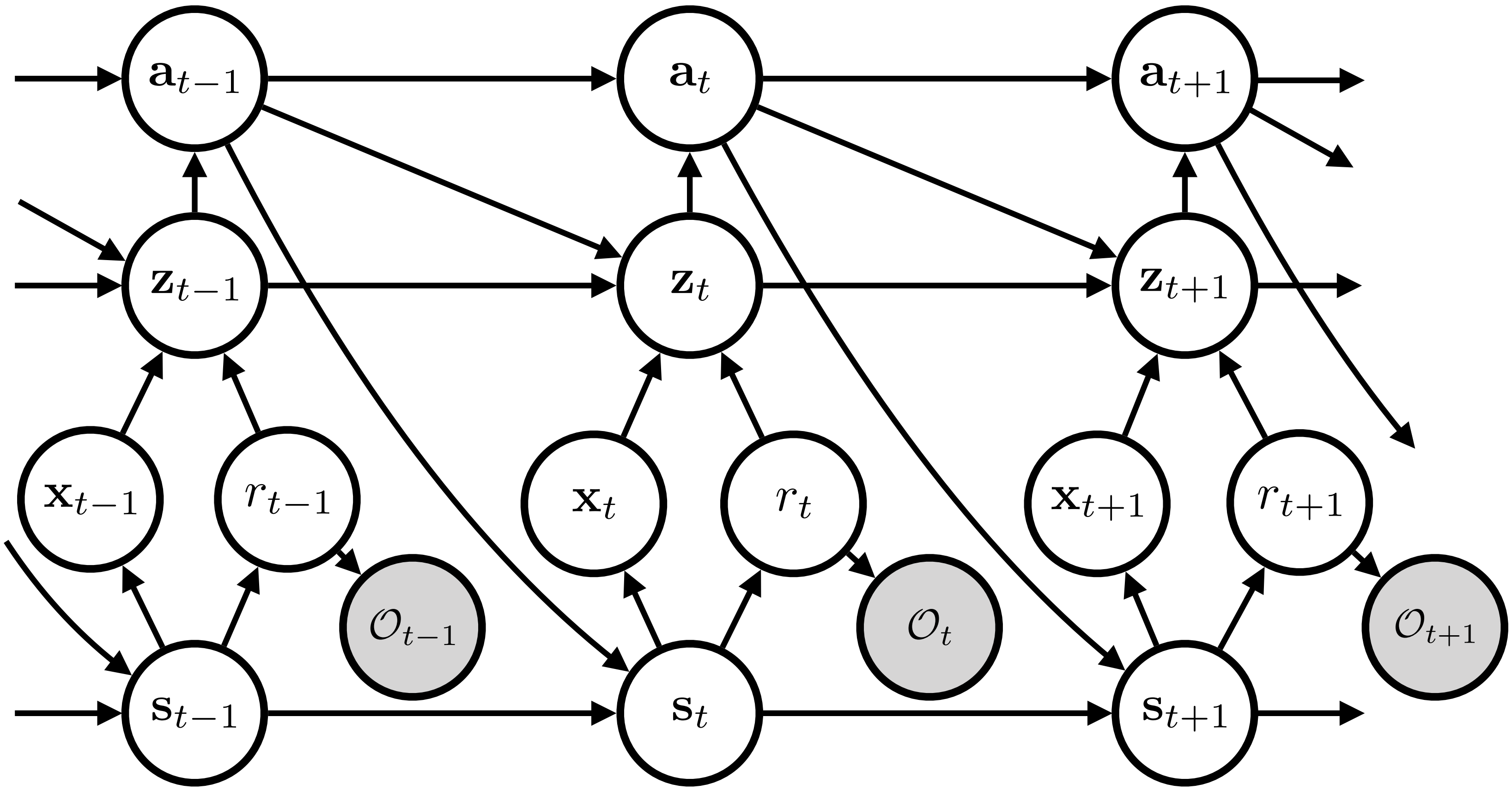
$$p_{\mathrm{a}}(\mathbf{a}_{1:T}, \mathbf{z}_{1:T} | \mathbf{x}_{1:T}, r_{1:T}) = \prod_{t=1}^{T} \underbrace{p_{\mathrm{a}}(\mathbf{a}_t | \mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})}_{\text{action prior}} \underbrace{p_{\mathrm{a}}(\mathbf{z}_t | \mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t})}_{\text{internal state prior}}$$

OPTIMALITY



$$p(\mathcal{O}_{1:T}|r_{1:T}) = \prod_{t=1}^{T} \underbrace{p(\mathcal{O}_t|r_t)}_{\text{cond. likelihood of optimality}}$$

ENVIRONMENT-AGENT-OPTIMALITY



$$p(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathbf{z}_{1:T}, \mathcal{O}_{1:T})$$

joint distribution

LEARNING

The agent's distributions are parameterized by θ.

Maximum Log-Likelihood Objective

$$\begin{aligned}\theta^* &= \arg\max_\theta \mathbb{E}_{\mathcal{O}_{1:T} \sim \delta(\mathbf{1})} \left[\log p(\mathcal{O}_{1:T})\right] \\ &= \arg\max_\theta \log p(\mathcal{O}_{1:T} = \mathbf{1}).\end{aligned}$$

where

$$p(\mathcal{O}_{1:T}) = \int p(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathbf{z}_{1:T}, \mathcal{O}_{1:T}) d\mathbf{x}_{1:T} dr_{1:T} d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} d\mathbf{z}_{1:T}$$

VARIATIONAL INFERENCE

We cannot evaluate $\log p(\mathcal{O}_{1:T})$ due to the intractable marginalization.

Introduce a *structured* approximate posterior, q:

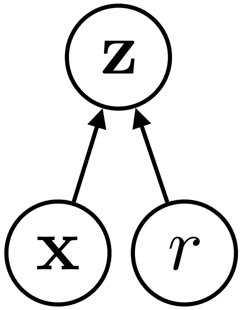
$$q(\mathbf{z}_{1:T}, \mathbf{a}_{1:T} | \mathbf{x}_{1:T}, r_{1:T}, \mathcal{O}_{1:T}) = \prod_{t=1}^{T} \overbrace{q(\mathbf{z}_t | \mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}^{\text{internal state}} \cdot \underbrace{q(\mathbf{a}_t | \mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}_{\text{action}}$$

This results in a lower bound on the objective:

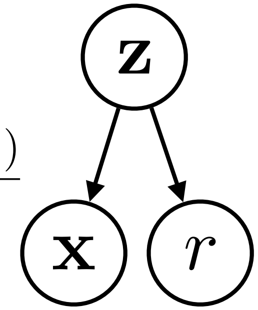
$$\mathcal{L}(q) \leq \log p(\mathcal{O}_{1:T})$$

GENERATIVE AGENT LOWER BOUND

Convert state estimation into a generative mapping using Bayes' Rule:



$$p_{\mathrm{a}}(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}) = \frac{p_{\mathrm{a}}(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{<t}, r_{<t}) p_{\mathrm{a}}(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}{p_{\mathrm{a}}(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}$$



Plugging this into the bound yields:

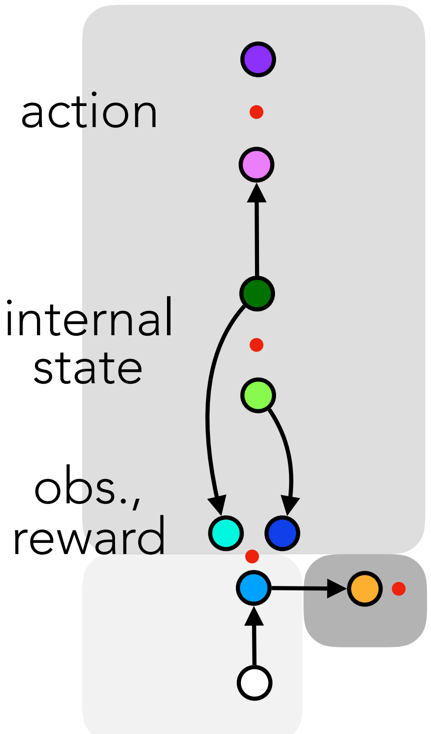
$$\mathcal{L} = \mathbb{E}_{\substack{\mathbf{s},\mathbf{x},r \sim p_{\mathrm{e}} \\ \mathbf{z},\mathbf{a} \sim q}} \Bigg[\sum_{t=1}^{T} \underbrace{r_t}_{\text{reward}} + \overbrace{\log \frac{p_{\mathrm{a}}(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{<t}, r_{<t})}{p_{\mathrm{a}}(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}}^{\text{information gain}} - \overbrace{\log \frac{q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_{\mathrm{a}}(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}}^{\text{internal state consistency}}$$

$$\underbrace{- \log \frac{q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_{\mathrm{a}}(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})}}_{\text{action consistency}} \Bigg]$$

generative agent bound

GENERATIVE AGENT LOWER BOUND

Computation Graph:



• log likelihood, log ratio

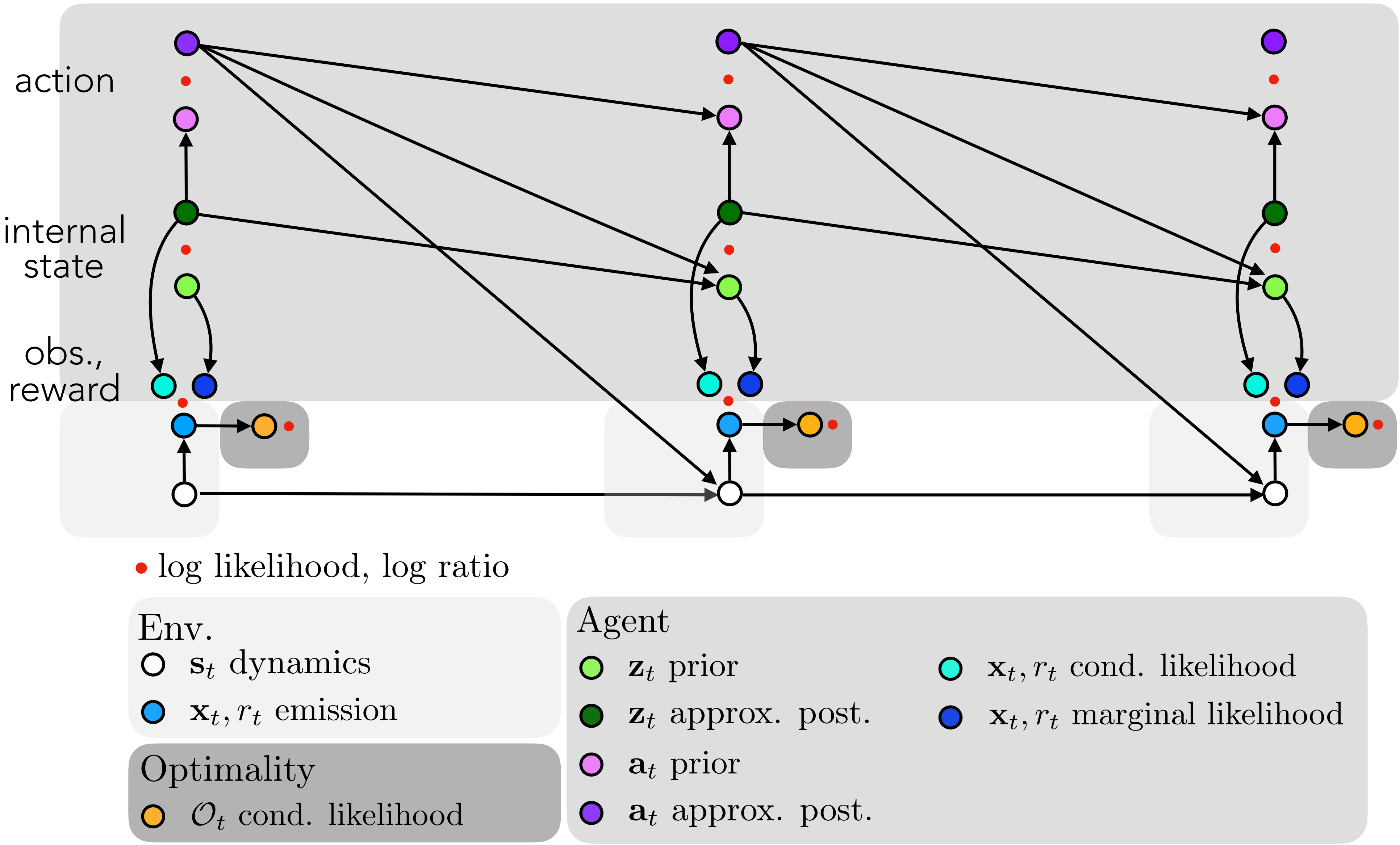
Env.
- $\mathbf{s}_t$ dynamics
- $\mathbf{x}_t, r_t$ emission

Optimality
- $\mathcal{O}_t$ cond. likelihood

Agent
- $\mathbf{z}_t$ prior
- $\mathbf{z}_t$ approx. post.
- $\mathbf{a}_t$ prior
- $\mathbf{a}_t$ approx. post.
- $\mathbf{x}_t, r_t$ cond. likelihood
- $\mathbf{x}_t, r_t$ marginal likelihood

GENERATIVE AGENT LOWER BOUND

Computation Graph:



VARIATIONAL EM

while θ ***not converged:***

$\mathbf{x}_1, r_1, \mathbf{s}_1 \sim p_\mathrm{e}(\mathbf{x}_1|\mathbf{s}_1)p_\mathrm{e}(r_1|\mathbf{s}_1)p_\mathrm{e}(\mathbf{s}_1)$

for $t = 1 \ldots T$:

 # inference (simulated rollouts)

$q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T}) \leftarrow \arg\max_q \mathcal{L}_{t:T}$

$q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T}) \leftarrow \arg\max_q \mathcal{L}_{t:T}$

 # interaction

$\mathbf{a}_t \sim q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})$

$\mathbf{x}_{t+1}, r_{t+1}, \mathbf{s}_{t+1} \sim p_\mathrm{e}(\mathbf{x}_{t+1}|\mathbf{s}_{t+1})p_\mathrm{e}(r_{t+1}|\mathbf{s}_{t+1})p_\mathrm{e}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

 # learning

$\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}$

We typically cannot simulate the environment to evaluate $\mathcal{L}_{t:T}$

ACTION INFERENCE VIA **PLANNING**

Estimate $\mathcal{L}_{t:T}$ using the generative agent's internal model.

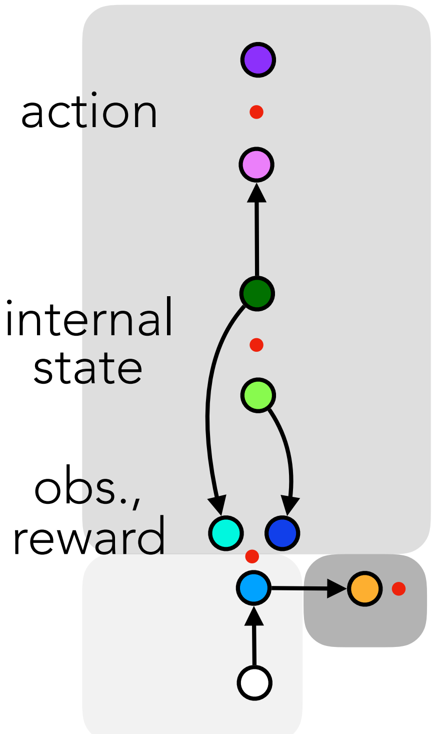
Replace p_{e} and q in future terms with p_{a}:

$$\widehat{\mathcal{L}}_{t+1:T} = \mathbb{E}_{\mathbf{x},r,\mathbf{z},\mathbf{a}\sim p_{\mathrm{a}}}\left[\sum_{\tau=t+1}^{T} \underbrace{r_\tau}_{\text{reward}} + \overbrace{\log \frac{p_{\mathrm{a}}(\mathbf{x}_\tau, r_\tau | \mathbf{a}_{<\tau}, \mathbf{z}_{\leq\tau}, \mathbf{x}_{<\tau}, r_{<\tau})}{p_{\mathrm{a}}(\mathbf{x}_\tau, r_\tau | \mathbf{a}_{<\tau}, \mathbf{z}_{<\tau}, \mathbf{x}_{<\tau}, r_{<\tau})}}^{\text{mutual information}}\right]$$

planning bound

PLANNING

Computation Graph:



• log likelihood, log ratio

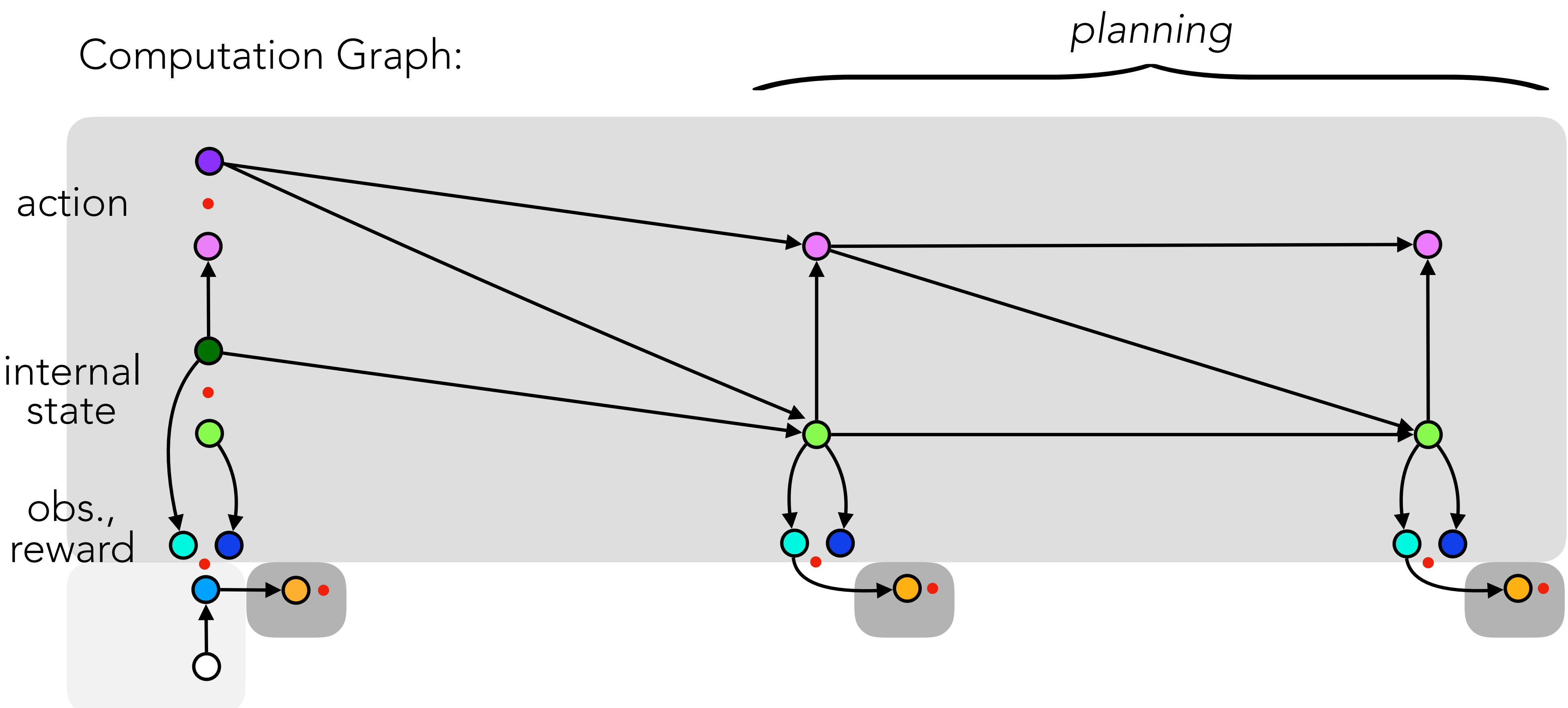
Env.
- $\mathbf{s}_t$ dynamics
- $\mathbf{x}_t, r_t$ emission

Optimality
- $\mathcal{O}_t$ cond. likelihood

Agent
- $\mathbf{z}_t$ prior
- $\mathbf{z}_t$ approx. post.
- $\mathbf{a}_t$ prior
- $\mathbf{a}_t$ approx. post.
- $\mathbf{x}_t, r_t$ cond. likelihood
- $\mathbf{x}_t, r_t$ marginal likelihood

PLANNING

Computation Graph:



• log likelihood, log ratio

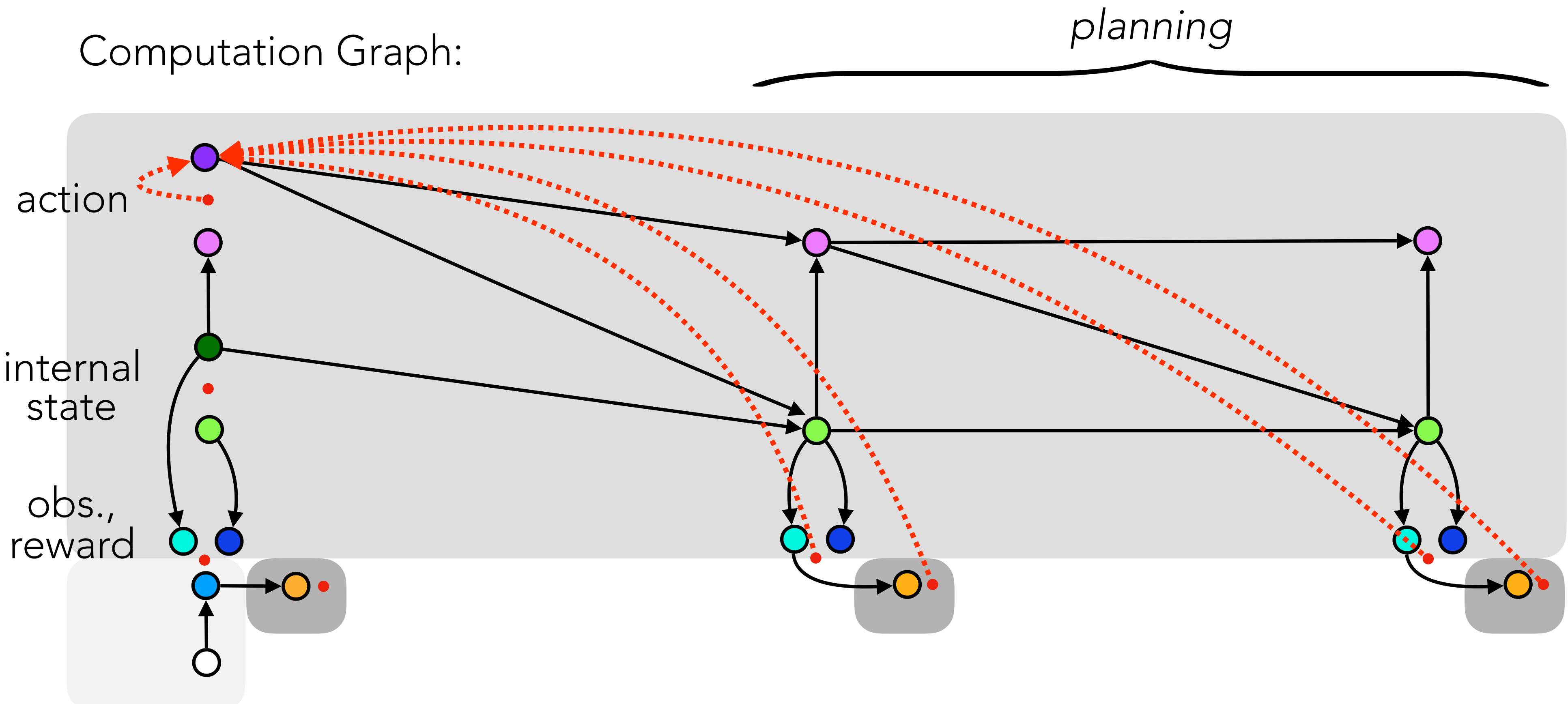
Env.
- ○ $\mathbf{s}_t$ dynamics
- ● $\mathbf{x}_t, r_t$ emission

Optimality
- ● $\mathcal{O}_t$ cond. likelihood

Agent
- ● $\mathbf{z}_t$ prior
- ● $\mathbf{z}_t$ approx. post.
- ● $\mathbf{a}_t$ prior
- ● $\mathbf{a}_t$ approx. post.
- ● $\mathbf{x}_t, r_t$ cond. likelihood
- ● $\mathbf{x}_t, r_t$ marginal likelihood

PLANNING

Computation Graph:



Action selection depends on:

MODEL-FREE PRIOR

$$p_{\mathrm{a}}(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})$$

MODEL-BASED LIKELIHOOD

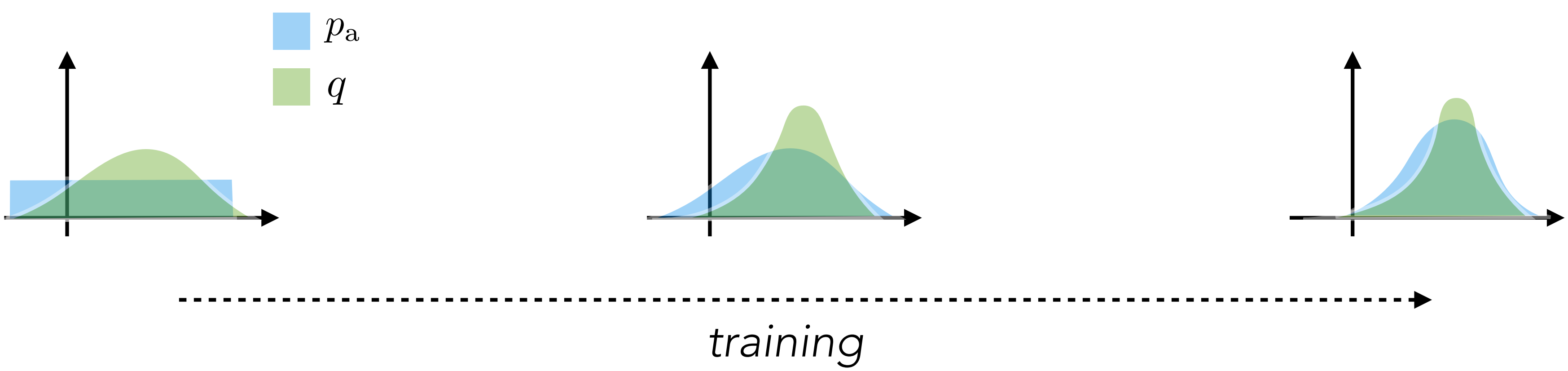
$$\widehat{\mathcal{L}}_{t+1:T}$$

PLANNING *DISTILLATION*

"**Habits** are sometimes said to be controlled by *antecedent stimuli*, whereas **goal-directed behavior** is said to be controlled by *its consequences*."

Actions and habits: the development of behavioural autonomy, Dickinson, 1985
Reinforcement learning: An introduction, Sutton & Barto, 2018

Learning $p_{\mathrm{a}}(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})$ will shift it toward $q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})$



Model-based planning will be "distilled" into a model-free policy, forming a *habit*.

MODELING THE ENVIRONMENT

GENERATIVE AGENT:

$$\mathcal{L} = \mathbb{E}_{\substack{\mathbf{s},\mathbf{x},r\sim p_{\mathrm{e}} \\ \mathbf{z},\mathbf{a}\sim q}} \left[\sum_{t=1}^{T} r_t + \log \frac{p_{\mathrm{a}}(\mathbf{x}_t, r_t | \mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{<t}, r_{<t})}{p_{\mathrm{a}}(\mathbf{x}_t, r_t | \mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})} - \log \frac{q(\mathbf{z}_t | \mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_{\mathrm{a}}(\mathbf{z}_t | \mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})} \right.$$
$$\left. - \log \frac{q(\mathbf{a}_t | \mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_{\mathrm{a}}(\mathbf{a}_t | \mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})} \right]$$

learning bound

$$\widehat{\mathcal{L}}_{t+1:T} = \mathbb{E}_{\mathbf{x},r,\mathbf{z},\mathbf{a}\sim p_{\mathrm{a}}} \left[\sum_{\tau=t+1}^{T} r_\tau + \log \frac{p_{\mathrm{a}}(\mathbf{x}_\tau, r_\tau | \mathbf{a}_{<\tau}, \mathbf{z}_{\leq \tau}, \mathbf{x}_{<\tau}, r_{<\tau})}{p_{\mathrm{a}}(\mathbf{x}_\tau, r_\tau | \mathbf{a}_{<\tau}, \mathbf{z}_{<\tau}, \mathbf{x}_{<\tau}, r_{<\tau})} \right]$$

inference (planning) bound

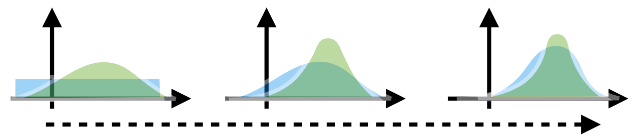
The likelihood ratio

learning: *encourages the agent to learn a task-relevant model*

inference: *biases planning toward less stochastic/uncertain outcomes*

RECAP: BENEFITS

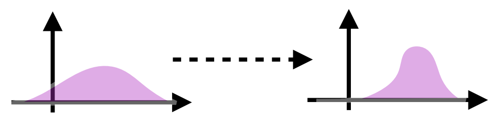
PLANNING DISTILLATION



convert model-based planning into a model-free policy

- fewer interactions during training
- after training, fast to act

MODELING THE ENVIRONMENT



estimate information gain / mutual information

- learn a more task-oriented internal state
- improve robustness during planning

joelouismarino.github.io

jmarino@caltech.edu