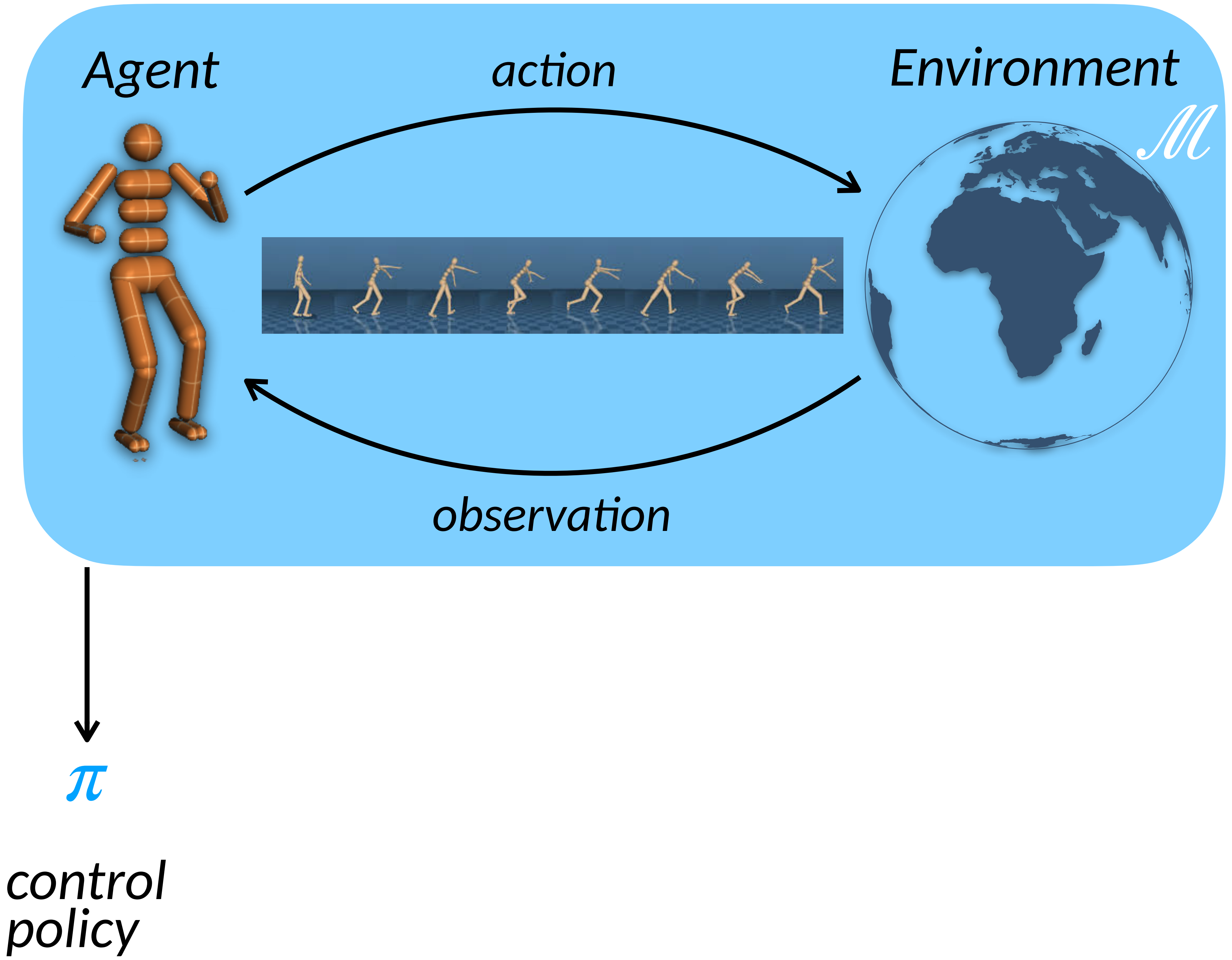
Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees

Florent Delgrange, Ann Nowé, Guillermo A. Pérez

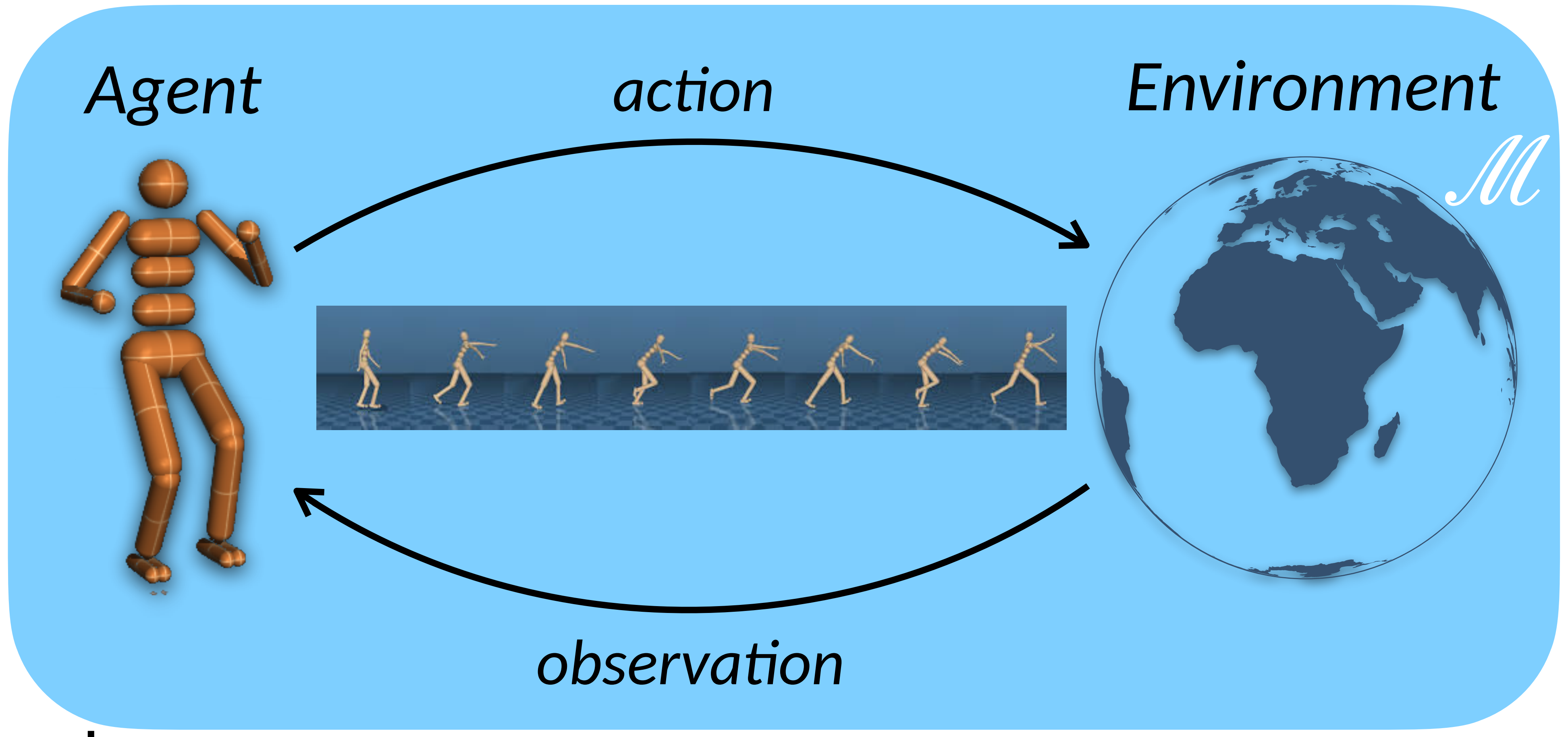
VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Universiteit Antwerpen

Reinforcement Learning

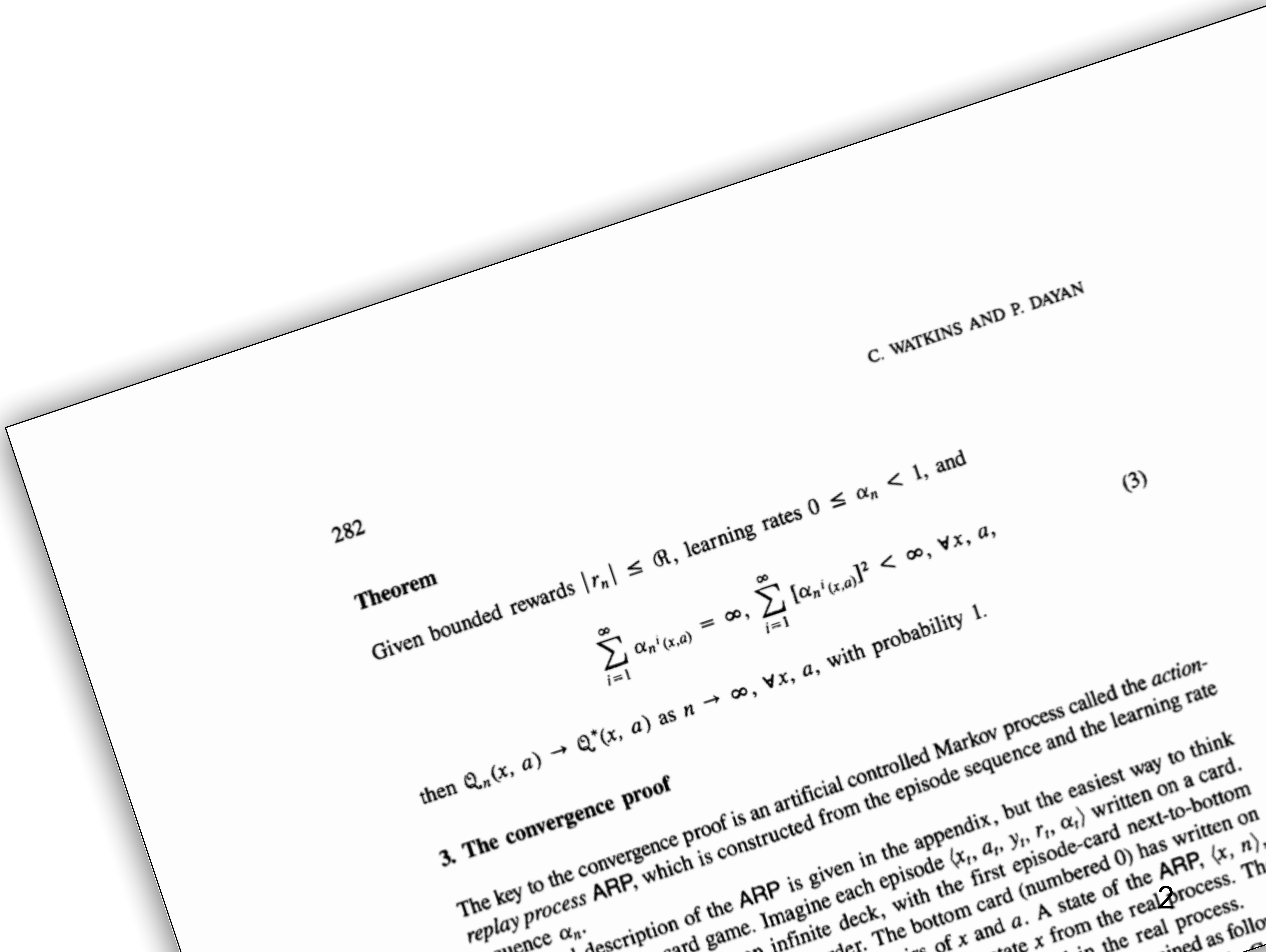


Reinforcement Learning



π

control policy



C. WATKINS AND P. DAYAN

282

Theorem

Given bounded rewards $|r_n| \le \mathcal{R}$, learning rates $0 \le \alpha_n < 1$, and

$$\sum_{i=1}^{\infty} \alpha_{n^i(x,a)} = \infty, \quad \sum_{i=1}^{\infty} [\alpha_{n^i(x,a)}]^2 < \infty, \forall x, a, \qquad (3)$$

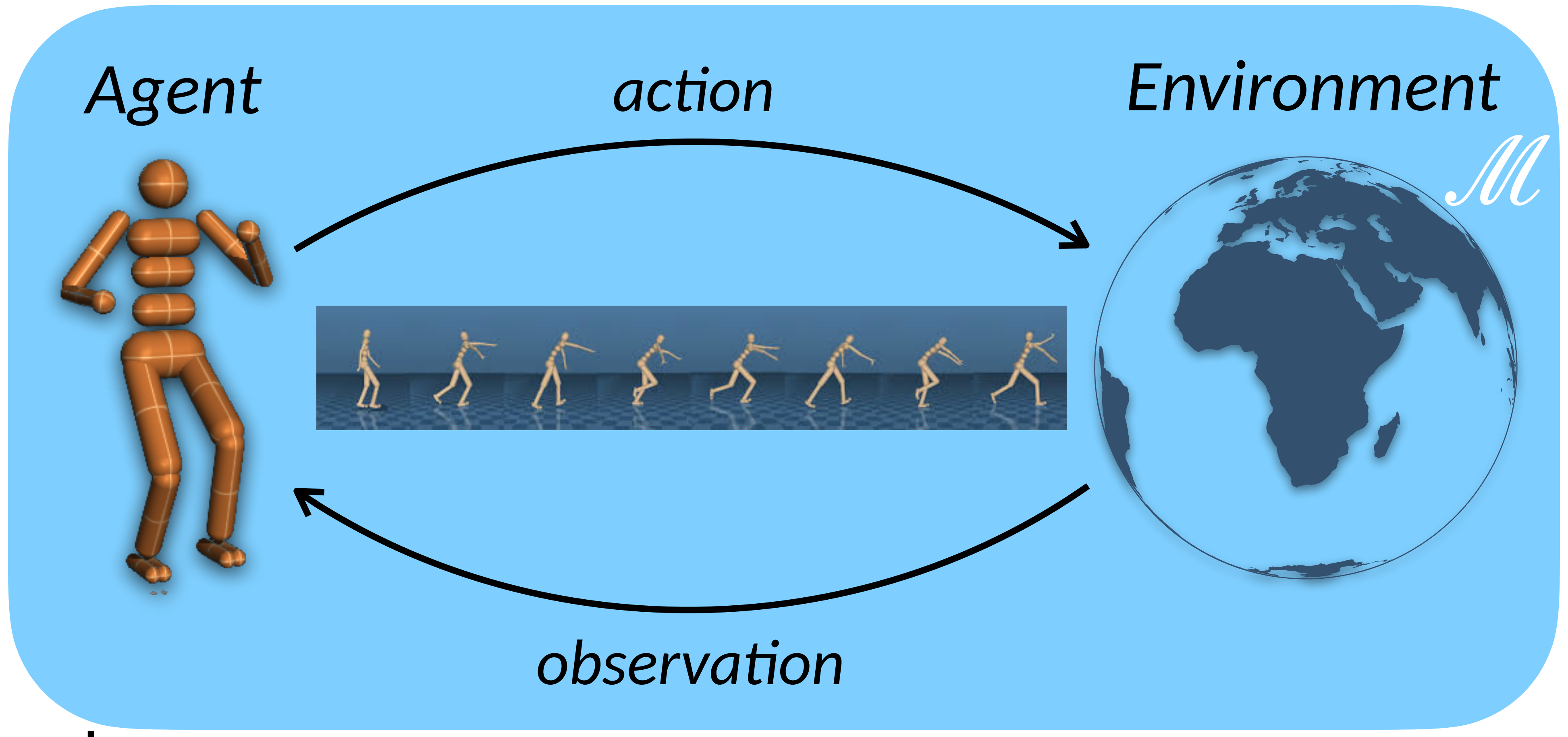
then $Q_n(x, a) \to Q^*(x, a)$ as $n \to \infty$, $\forall x, a$, with probability 1.

3. The convergence proof

The key to the convergence proof is an artificial controlled Markov process called the *action-replay process* ARP, which is constructed from the episode sequence and the learning rate sequence α_n.

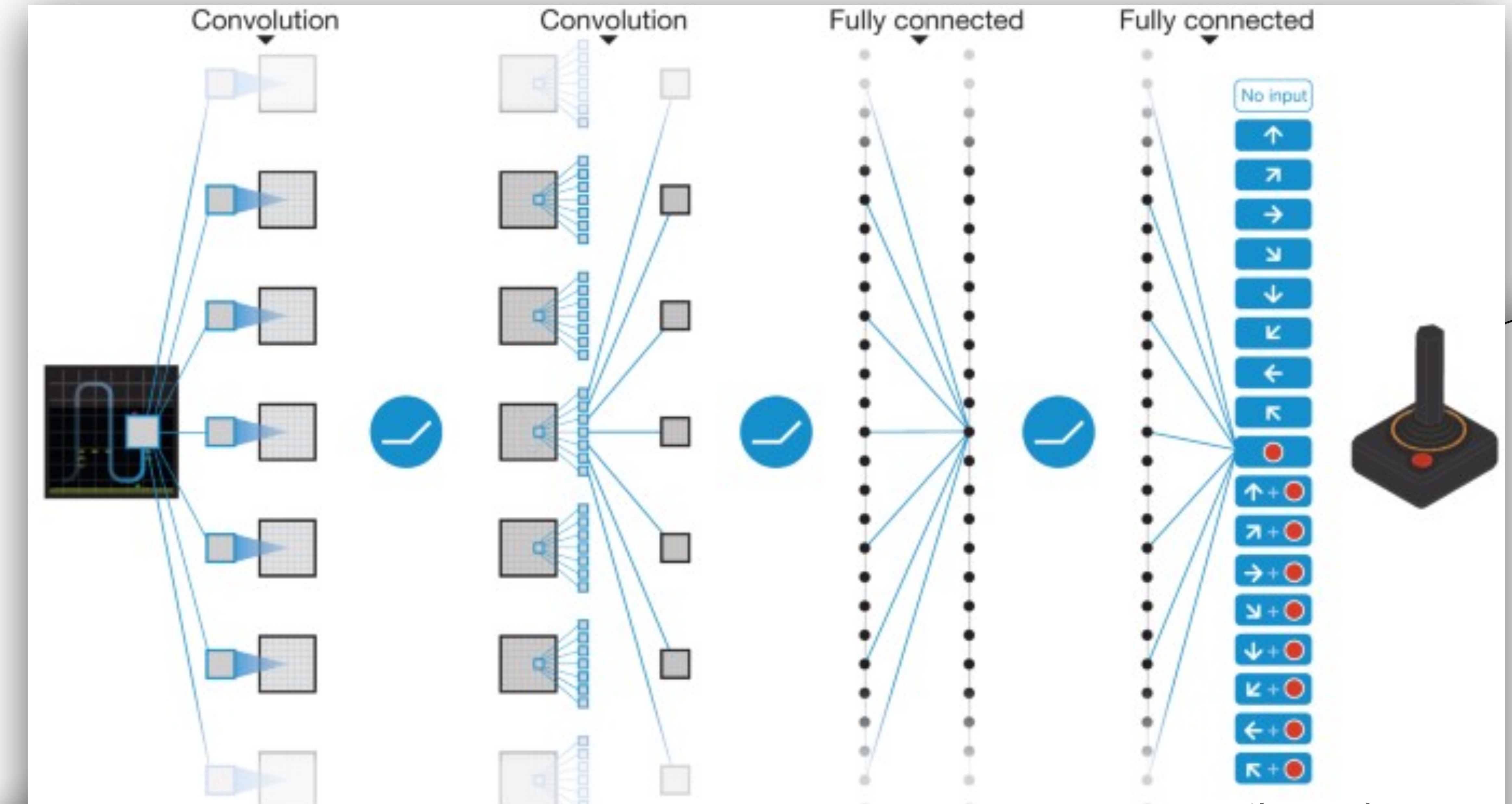
... description of the ARP is given in the appendix, but the easiest way to think ... card game. Imagine each episode $\langle x_t, a_t, y_t, r_t, \alpha_t \rangle$ written on a card. ... infinite deck, with the first episode-card next-to-bottom ... The bottom card (numbered 0) has written on ... of x and a. A state of the ARP, $\langle x, n \rangle$, ... state x from the real process. The ... the real process.

Reinforcement Learning

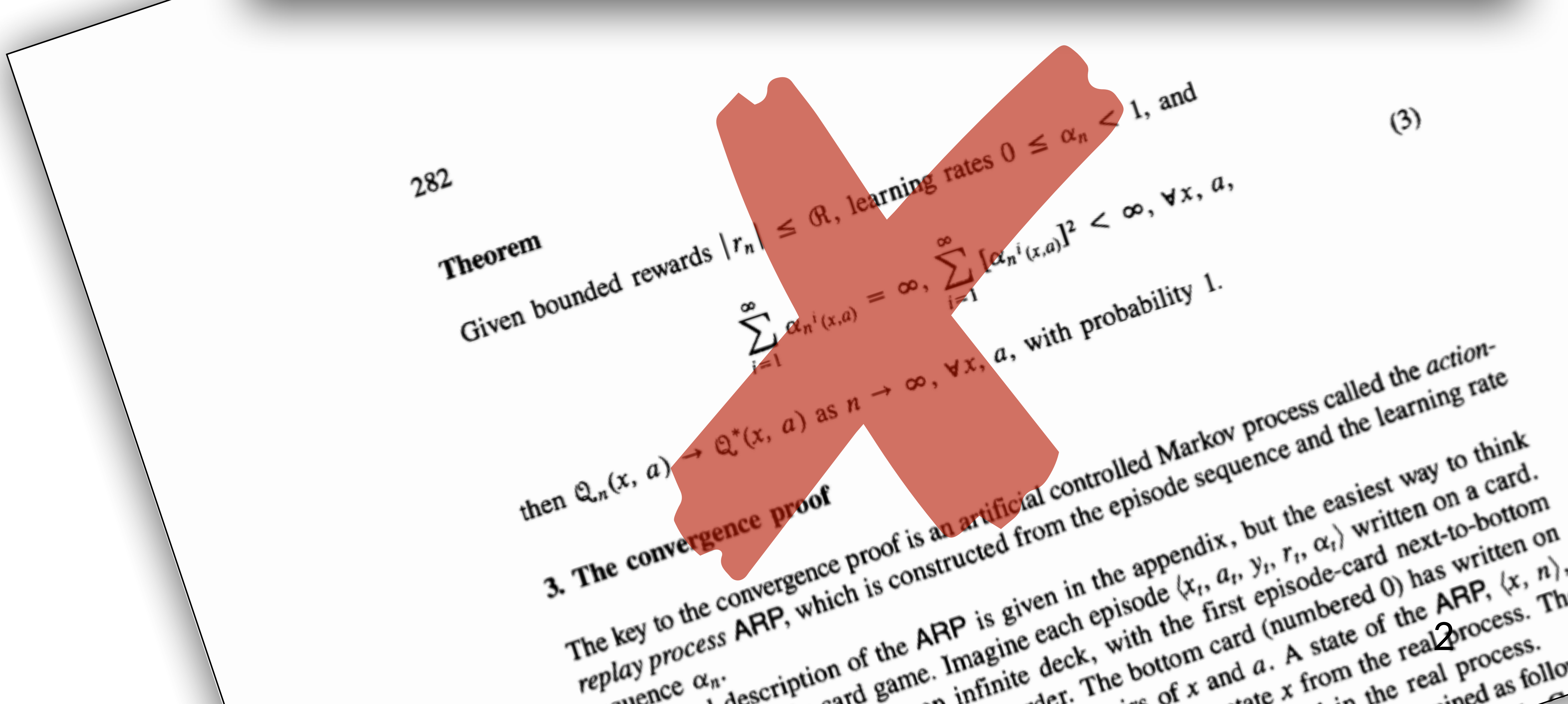


π

control policy

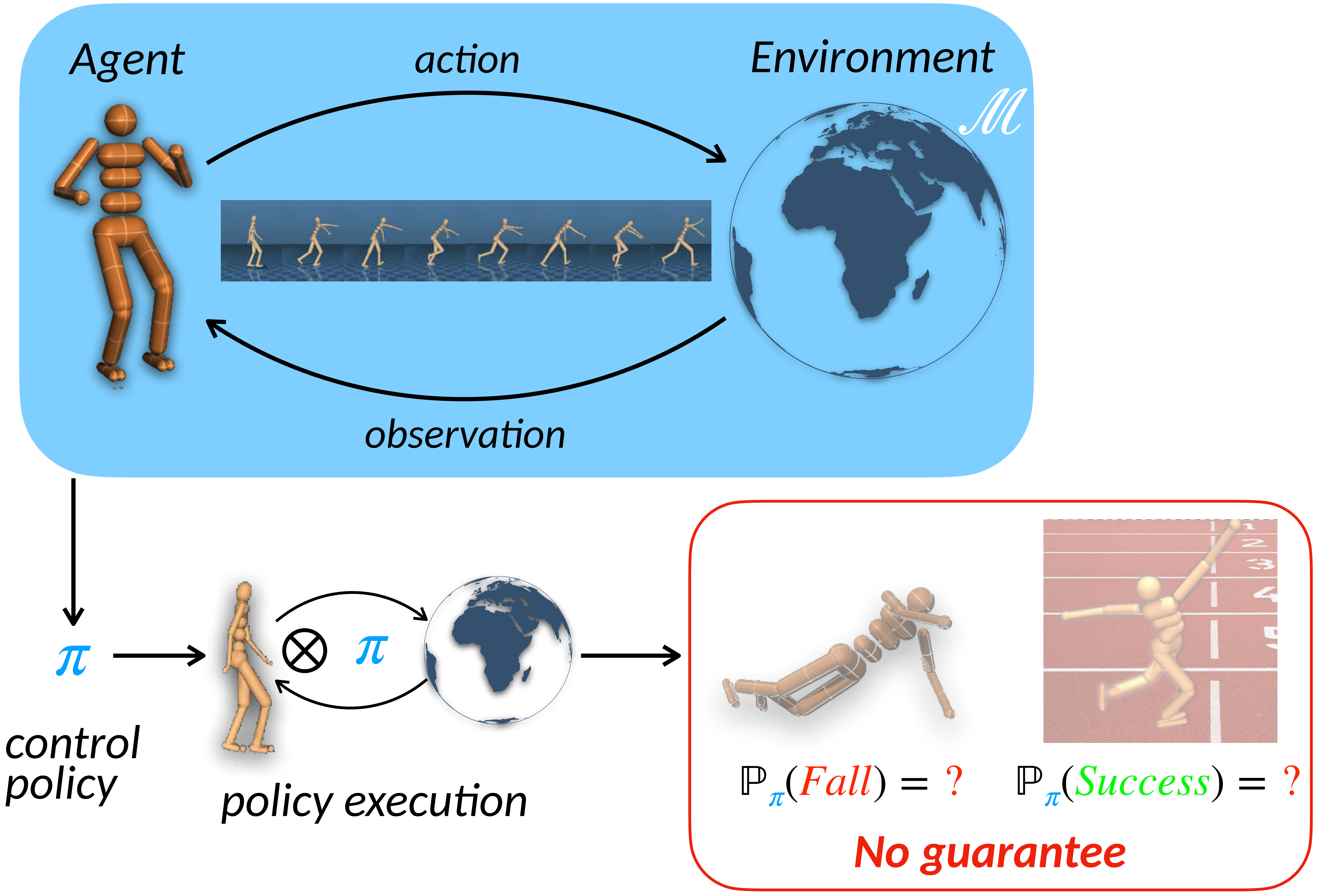


Mnih et. al, 2015



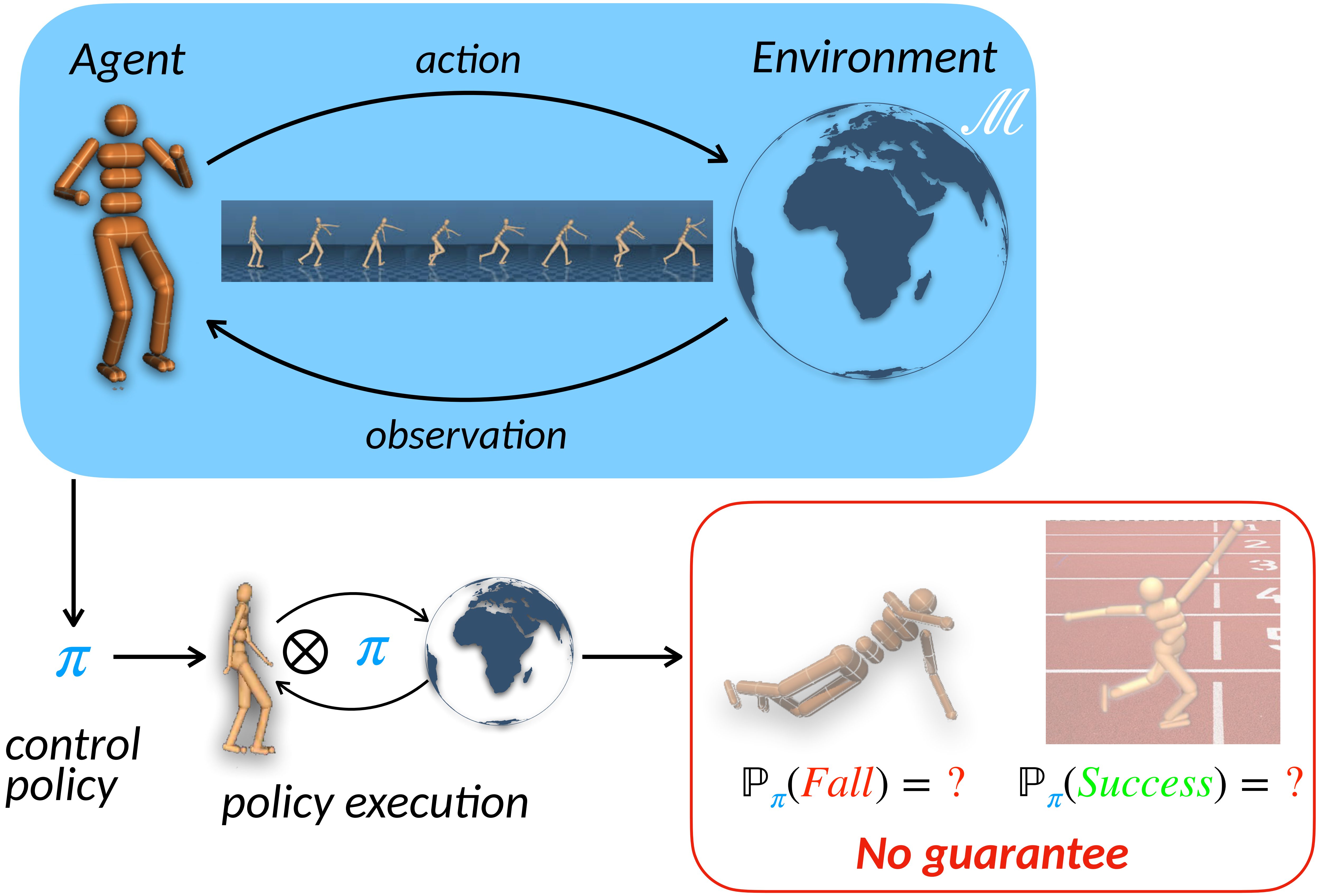
Overview

Reinforcement Learning



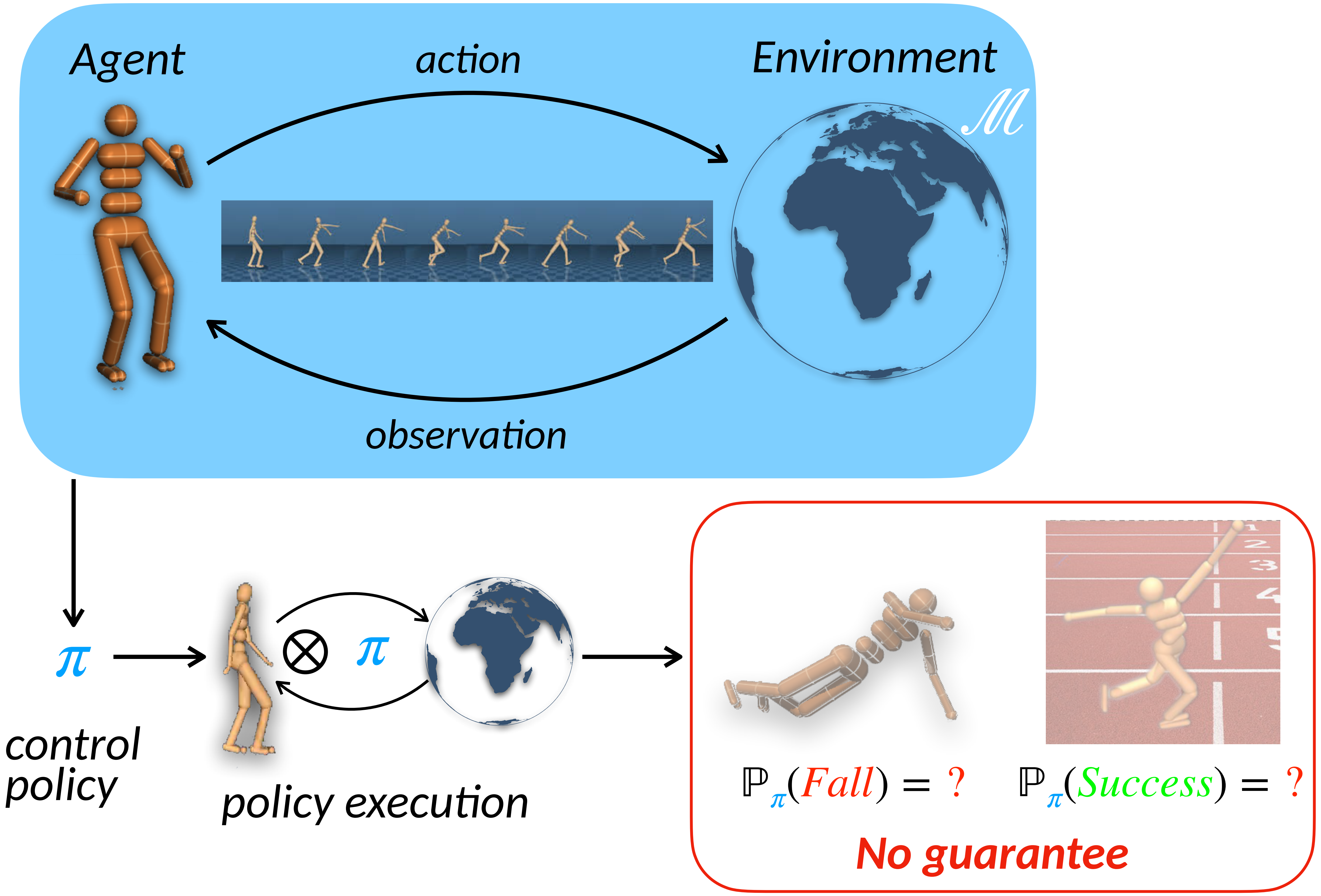
Overview

Reinforcement Learning



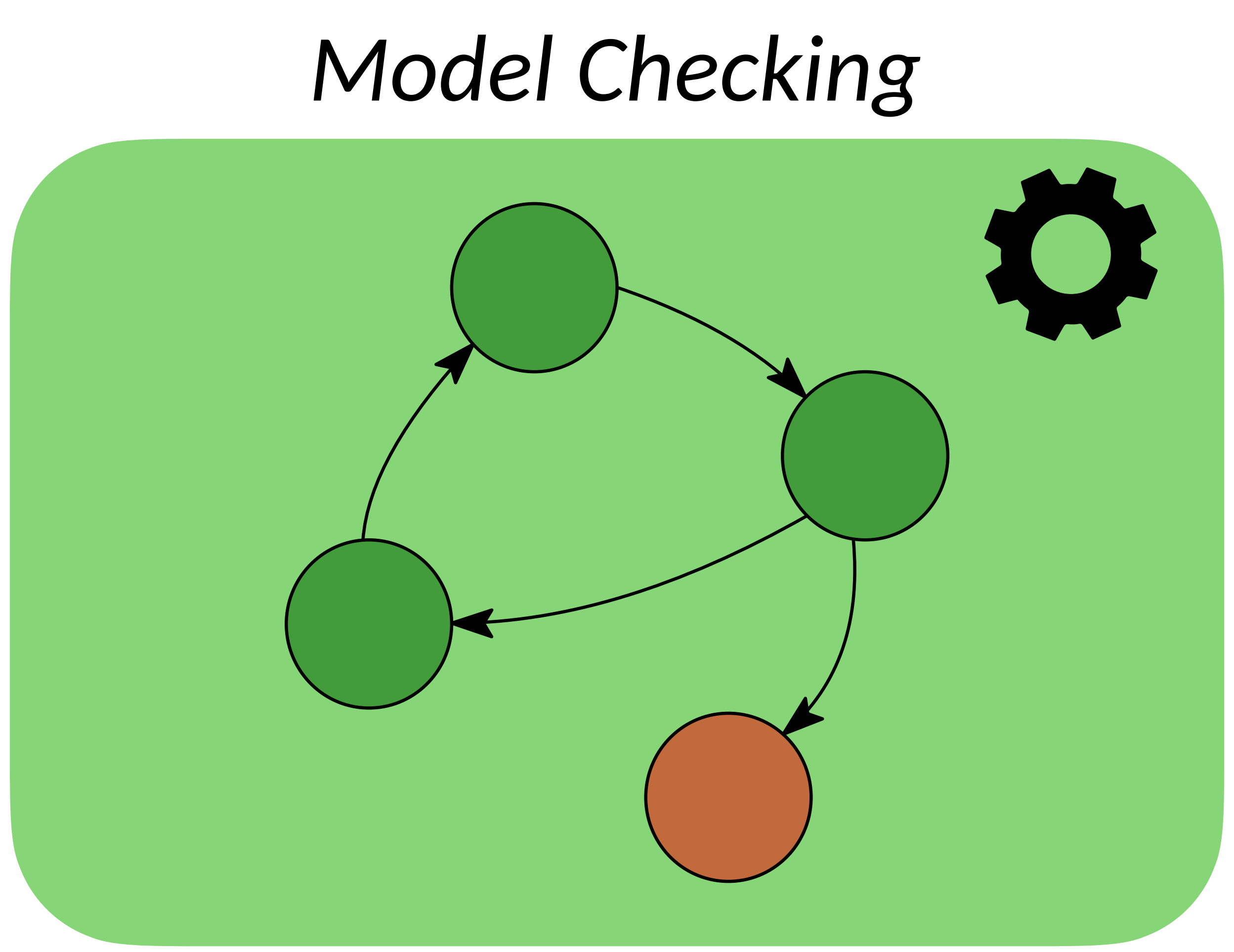
- Continuous state/action spaces
- Unknown environment

Reinforcement Learning



- Continuous state/action spaces
- Unknown environment

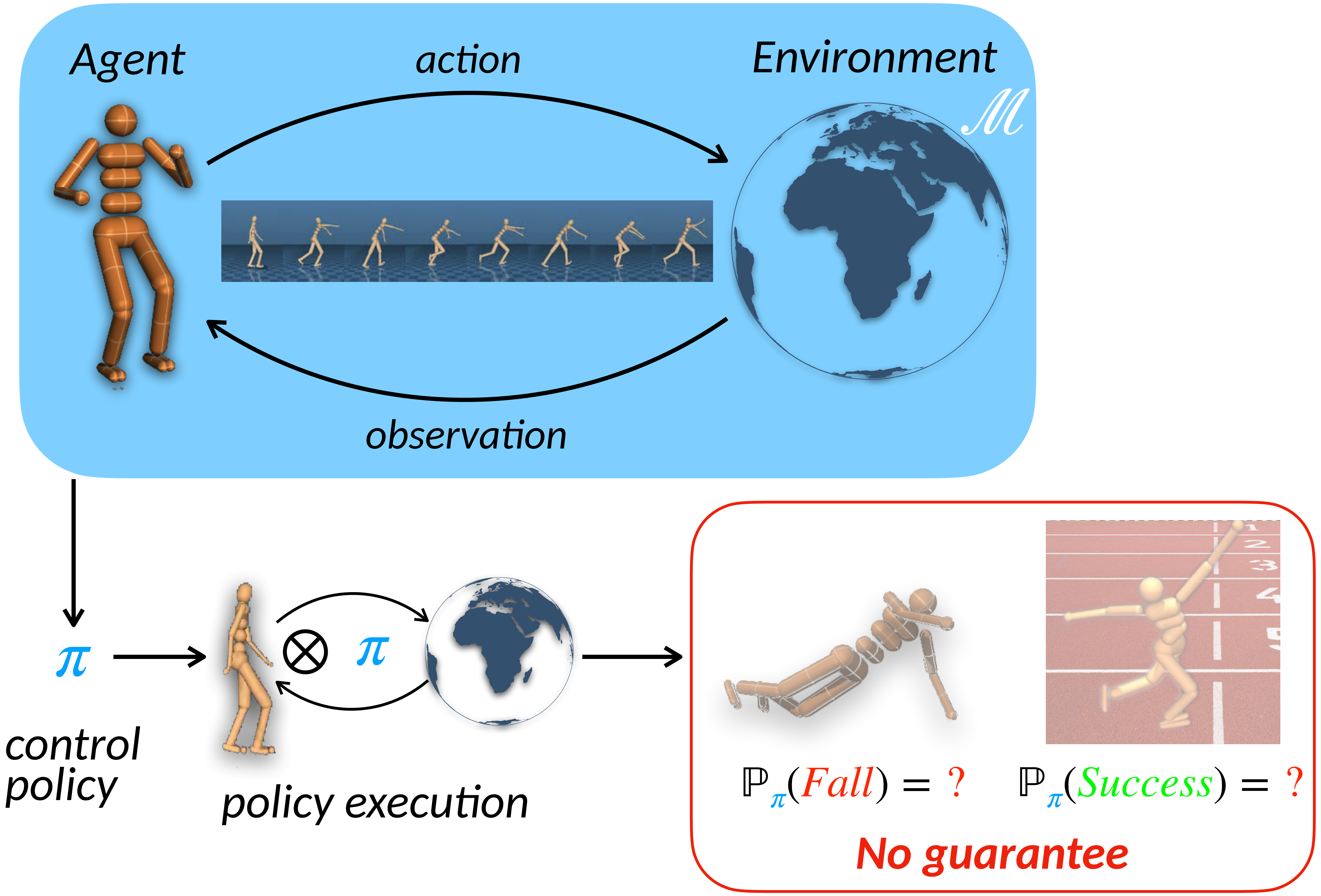
Formal Guarantees



- Full knowledge of the model of the interaction

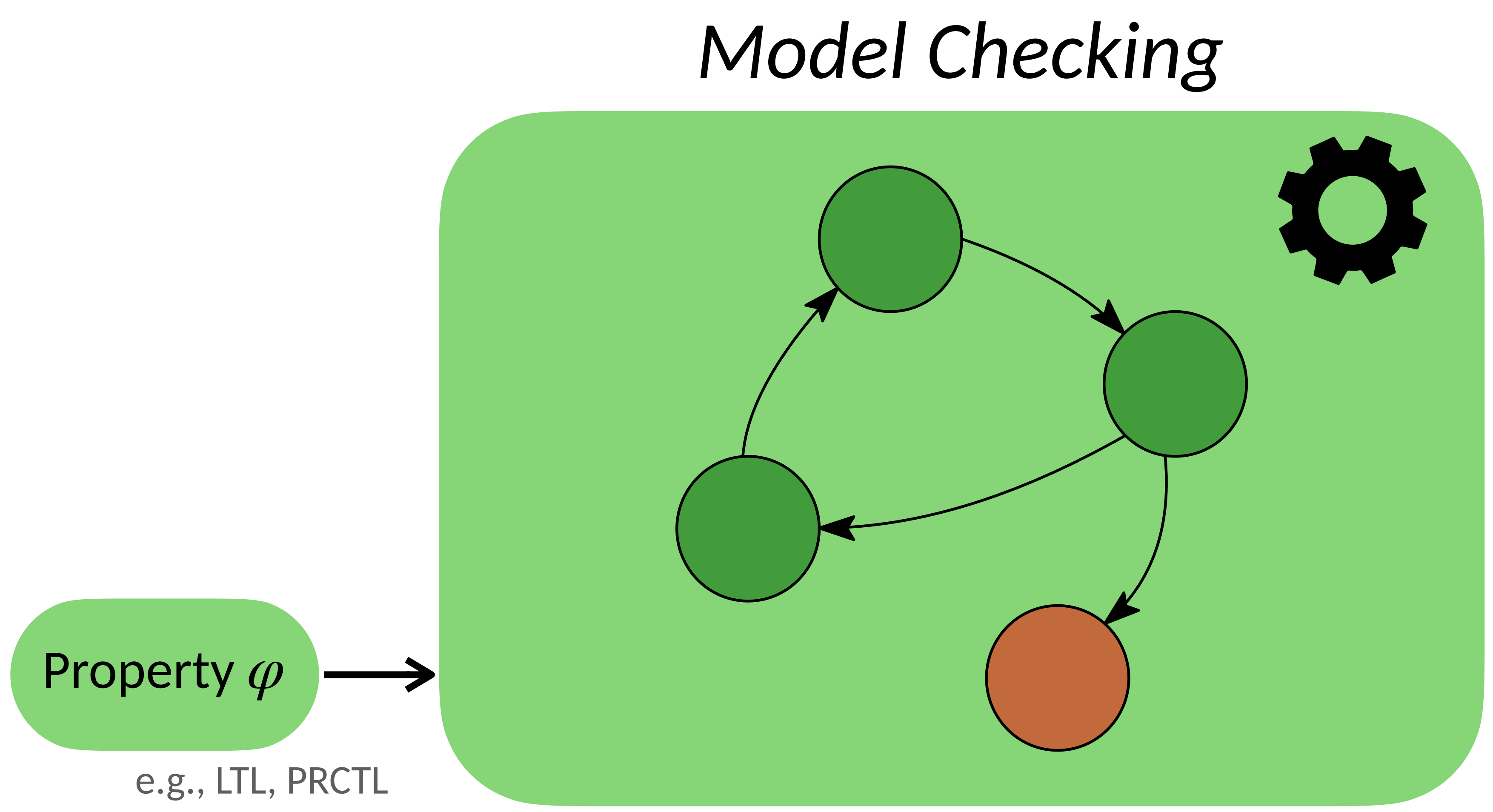
Overview

Reinforcement Learning



- Continuous state/action spaces
- Unknown environment

Formal Guarantees



- Full knowledge of the model of the interaction

Overview

Reinforcement Learning

Formal Guarantees

- Continuous state/action spaces
- Unknown environment

- Full knowledge of the model of the interaction

Overview

Reinforcement Learning

Formal Guarantees

- Continuous state/action spaces
- Unknown environment

- Full knowledge of the model of the interaction
- Exhaustive exploration of the model
- Sensitive to the state space explosion problem

Overview

Reinforcement Learning Policies with Formal Guarantees

- Continuous state/action spaces
- Unknown environment

- Full knowledge of the model of the interaction
- Exhaustive exploration of the model
- Sensitive to the state space explosion problem

Latent Space Model

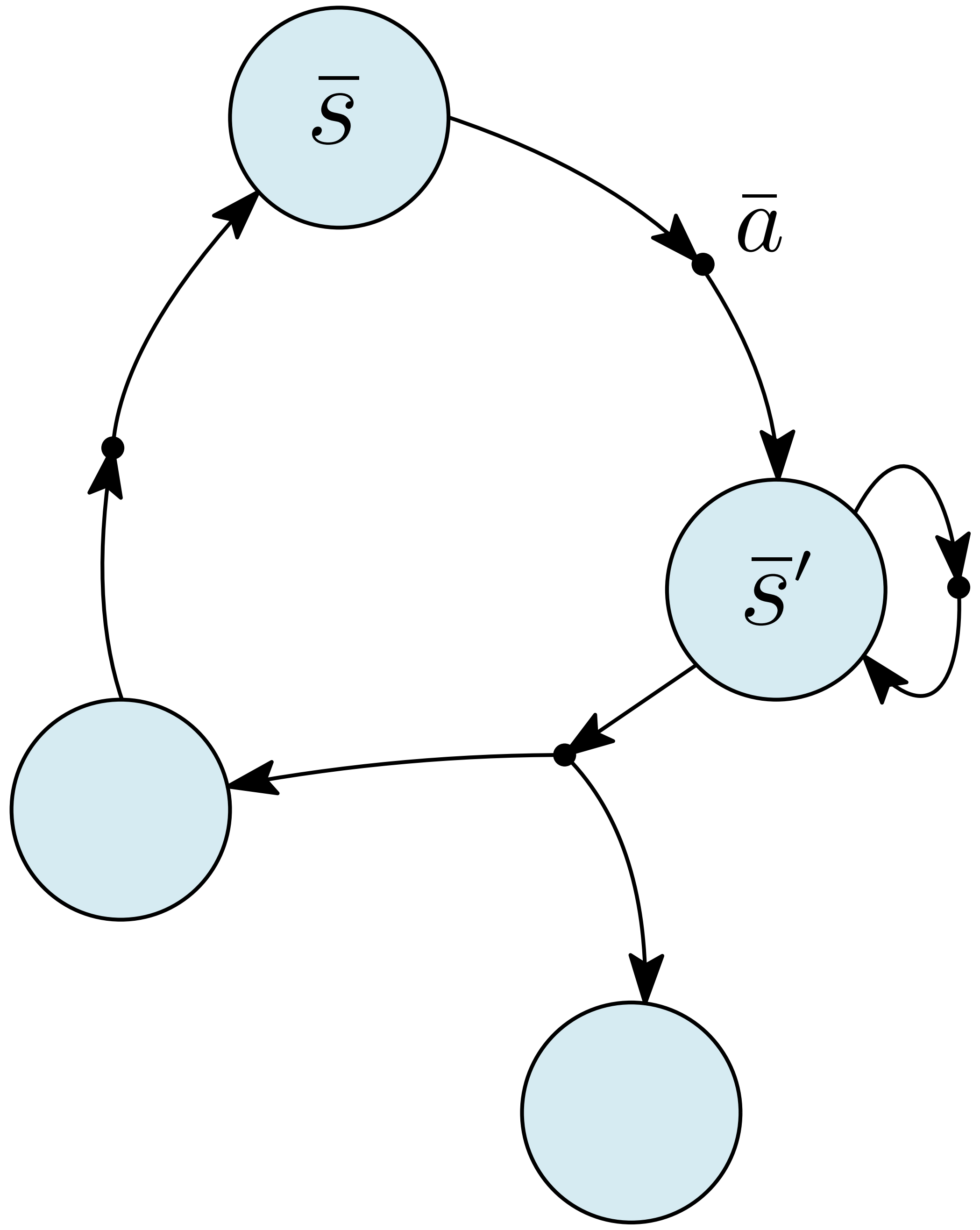
equivalent?

state embedding

$\phi\colon \mathcal{S} \to \bar{\mathcal{S}},\ s \mapsto \bar{s}$

action embedding

$\psi : \bar{\mathcal{S}} \times \bar{\mathcal{A}} \to \mathcal{A},\ \langle \bar{s}, \bar{a} \rangle \mapsto a$



Continuous-spaces MDP

Discrete latent MDP

$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$

$\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

Latent Space Model

$B \in \mathcal{S}^2$ is a ***stochastic bisimulation*** iff for all $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}, T/B$

$$\ell(s_1) = \ell(s_2) \qquad \mathcal{R}(s_1, a) = \mathcal{R}(s_2, a) \qquad \text{and} \qquad \mathbf{P}(T \mid s_1, a) = \mathbf{P}(T \mid s_2, a)$$

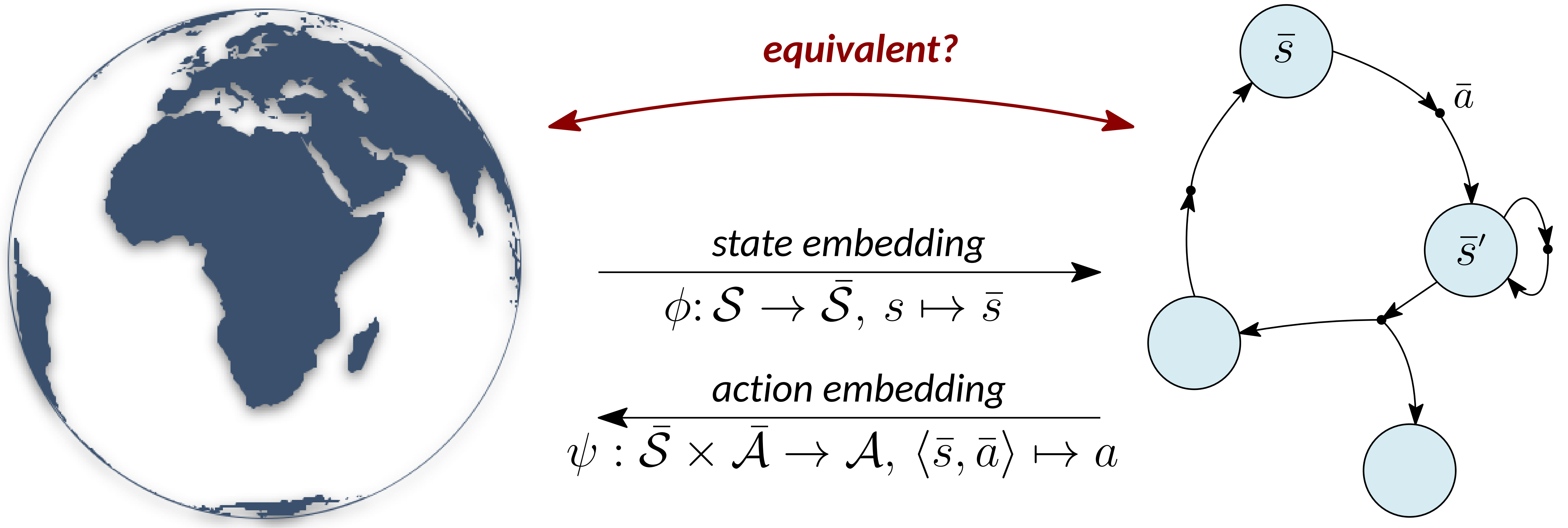
Largest: $\sim$

(Larsen and Skou 1989; Givan, Dean, and Greig 2003)

- Behavioral equivalence

➡ Trajectory, value, and optimal policy equivalence

➡ Agents ***behave in the same way*** in bisimilar models



Continuous-spaces MDP

Discrete latent MDP

$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$

$\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

Latent Space Model

$B \in \mathcal{S}^2$ is a ~~stochastic bisimulation iff~~ for all $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}, T/B$

$$\ell(s_1) = \ell(s_2) \qquad \mathcal{R}(s_1, a) \neq \mathcal{R}(s_2, a) + \epsilon \quad \text{and} \quad \mathbf{P}(T \mid s_1, a) \neq \mathbf{P}(T \mid s_2, a) + \epsilon$$

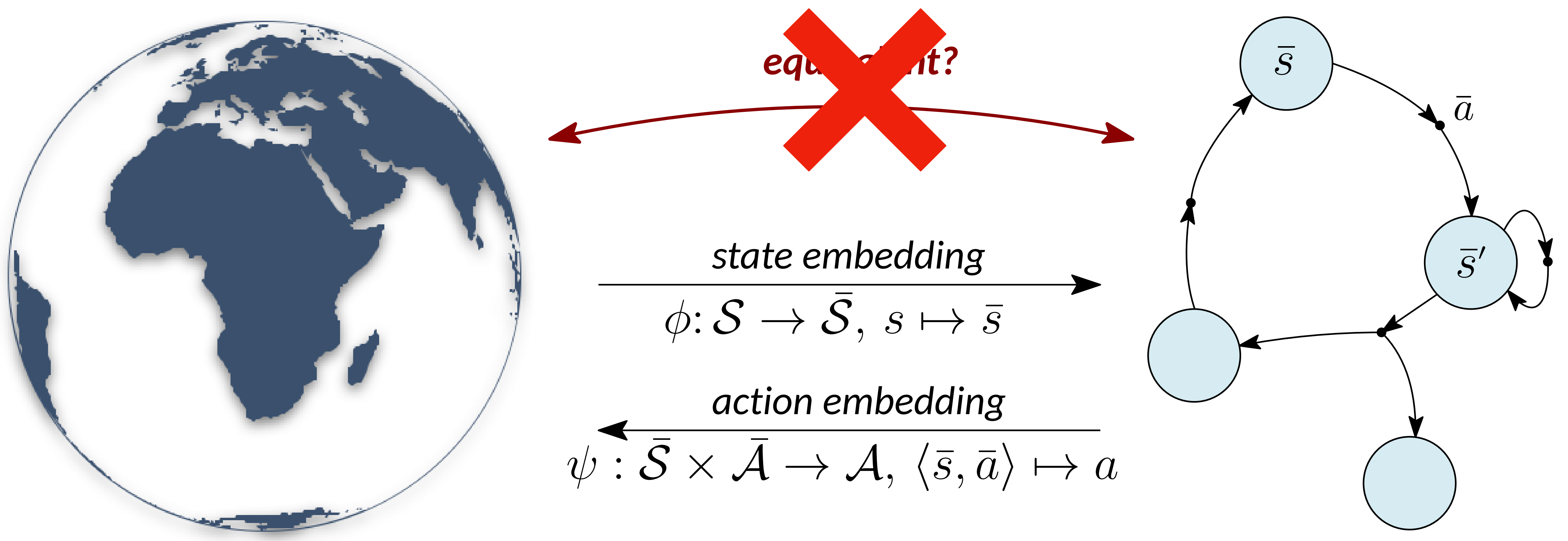
Largest: $\sim$

(Larsen and Skou 1989; Givan, Dean, and Greig 2003)

- Behavioral equivalence

➡ Trajectory, value, and optimal policy equivalence

➡ Agents ***behave in the same way*** in bisimilar models



Continuous-spaces MDP

Discrete latent MDP

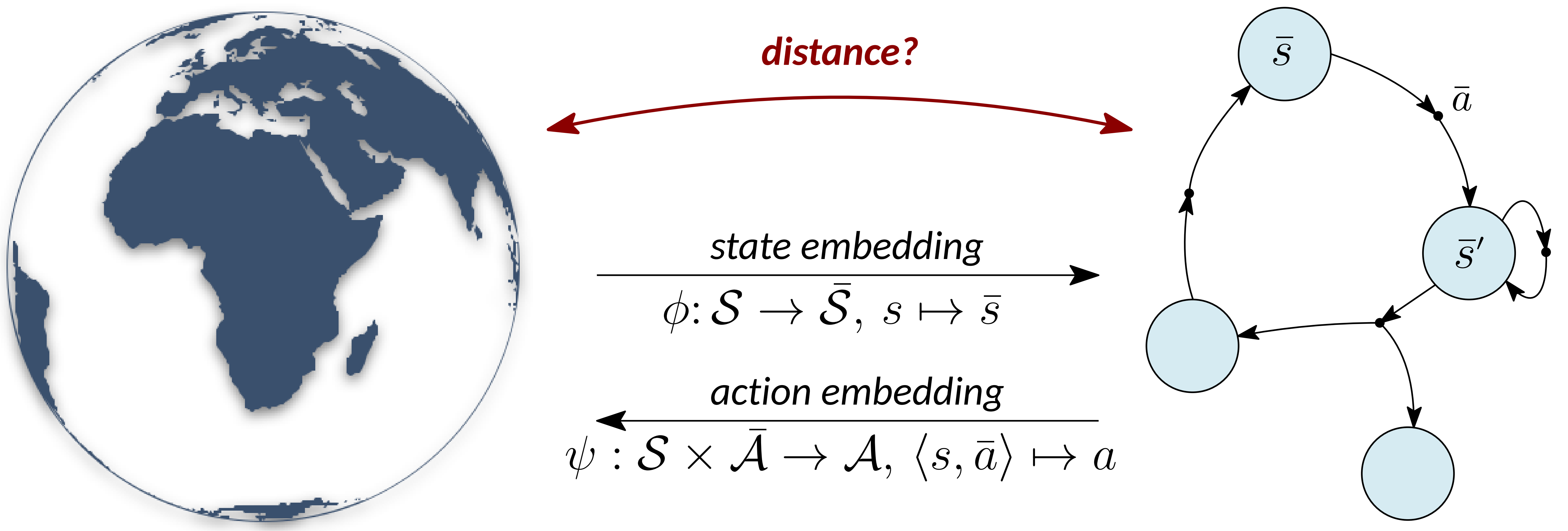
$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$

$\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

◉ *All or nothing:* two states *nearly identical* with slight numerical difference ϵ are $\neq$

Bisimulation distance

Continuous-spaces MDP | *Discrete latent* MDP



$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$ | $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

- For policy π, $\gamma \in [0,1[$, and formal logic $\mathscr{L}$:
  - ➡ ***Bisimulation distance:*** *largest behavioral difference* *(Desharnais et. al, 2004)*

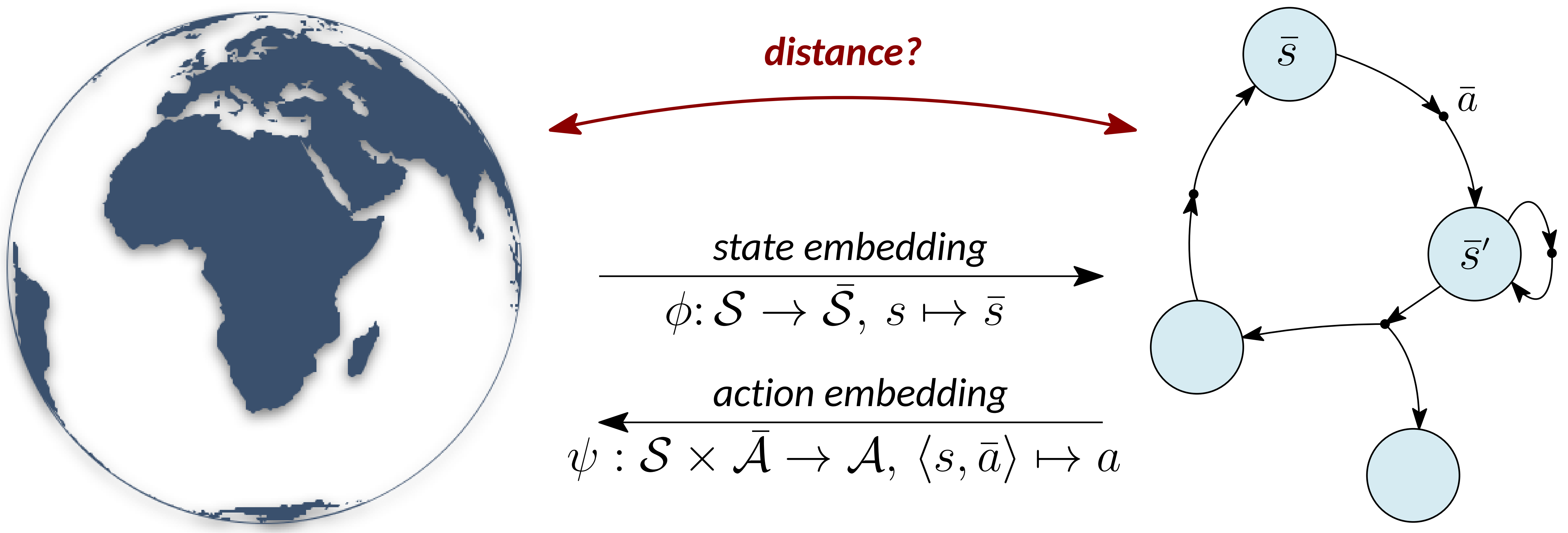
$$\tilde{d}_\pi (s_1, s_2) = \sup_{V \in \mathscr{F}_\gamma^{\mathscr{L}}(\pi)} \left| V_\pi(s_1) - V_\pi(s_2) \right| \quad \forall s_1, s_2 \in \mathcal{S}$$

where $\mathscr{F}_\gamma^{\mathscr{L}}(\pi)$ *is a logical family of functional expressions defining the semantics of* $\mathscr{L}$

➡ ***Kernel is bisimilarity:*** $\tilde{d}_\pi (s_1, s_2) = 0 \iff s_1 \sim s_2$

Bisimulation distance

Continuous-spaces MDP — *Discrete latent* MDP



$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$ $\qquad$ $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

- For policy π, $\gamma \in [0,1[$, and formal logic $\mathscr{L}$:
  - ➡ ***Bisimulation distance:*** *largest behavioral difference (Desharnais et. al, 2004)*

$$\tilde{d}_{\pi}(s_1, s_2) = \sup_{V \in \mathscr{F}_{\gamma}^{\mathscr{L}}(\pi)} \left| V_{\pi}(s_1) - V_{\pi}(s_2) \right| \quad \forall s_1, s_2 \in \mathcal{S}$$

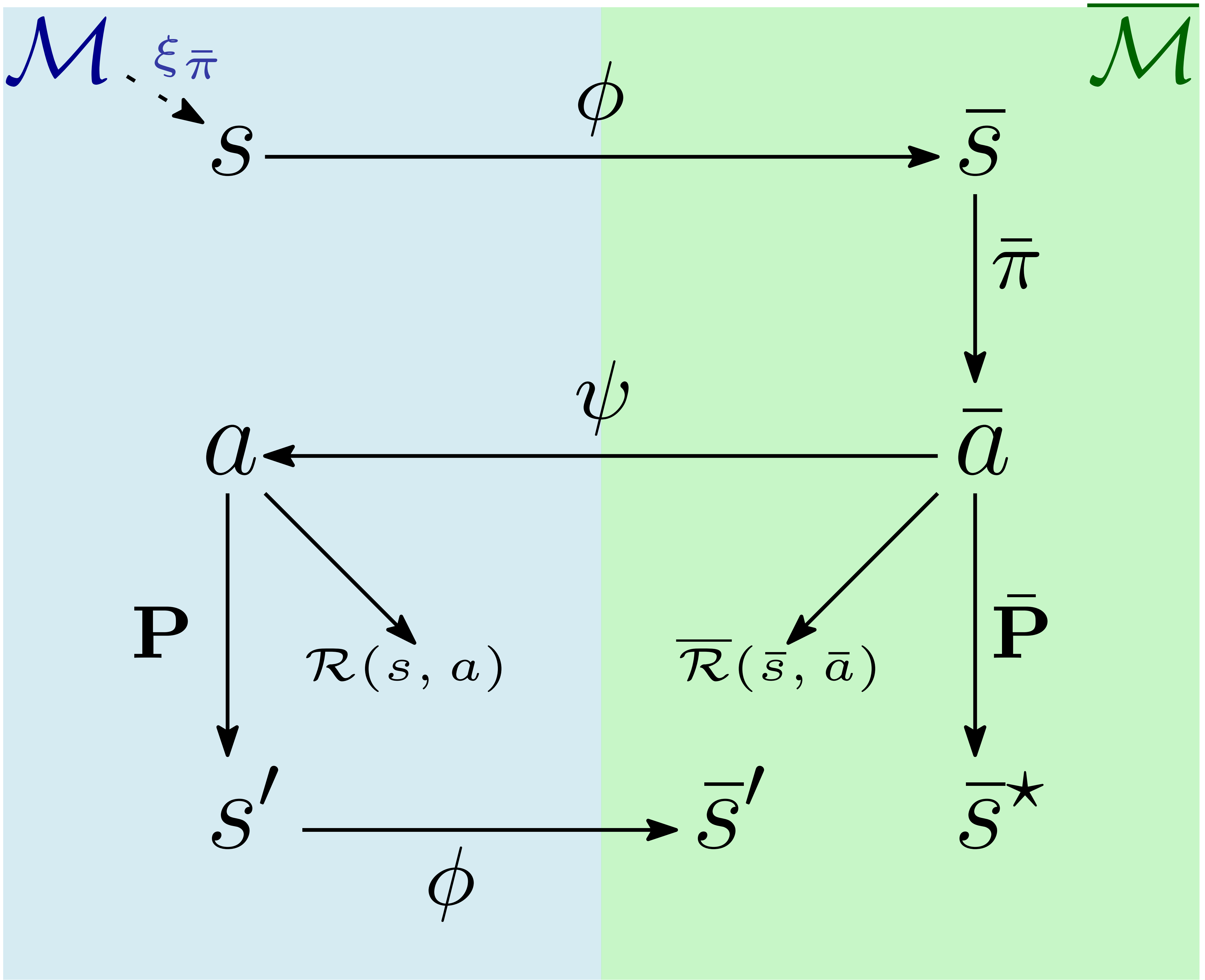
We need a policy that can be executed (separately) in the original and latent models

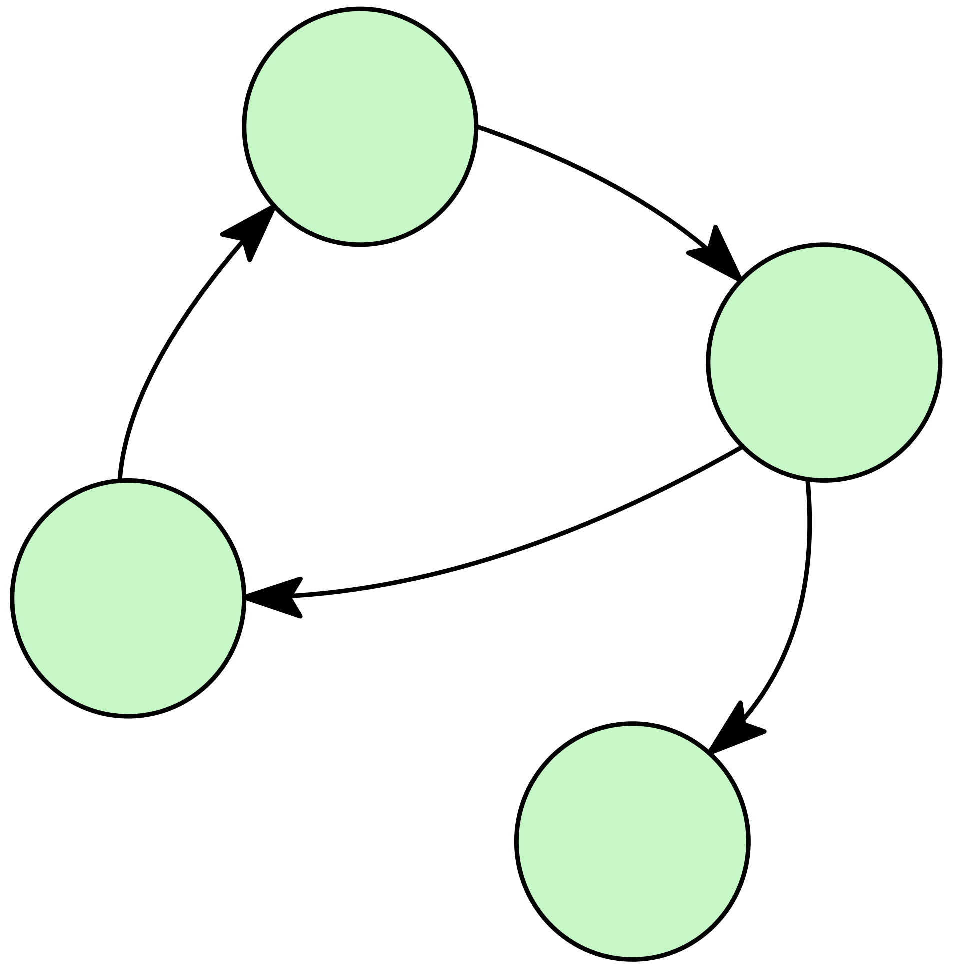
where $\mathscr{F}_{\gamma}^{\mathscr{L}}(\pi)$ *is a logical family of functional expressions defining the semantics of* $\mathscr{L}$

➡ ***Kernel is bisimilarity:*** $\tilde{d}_{\pi}(s_1, s_2) = 0 \iff s_1 \sim s_2$

Latent Flow

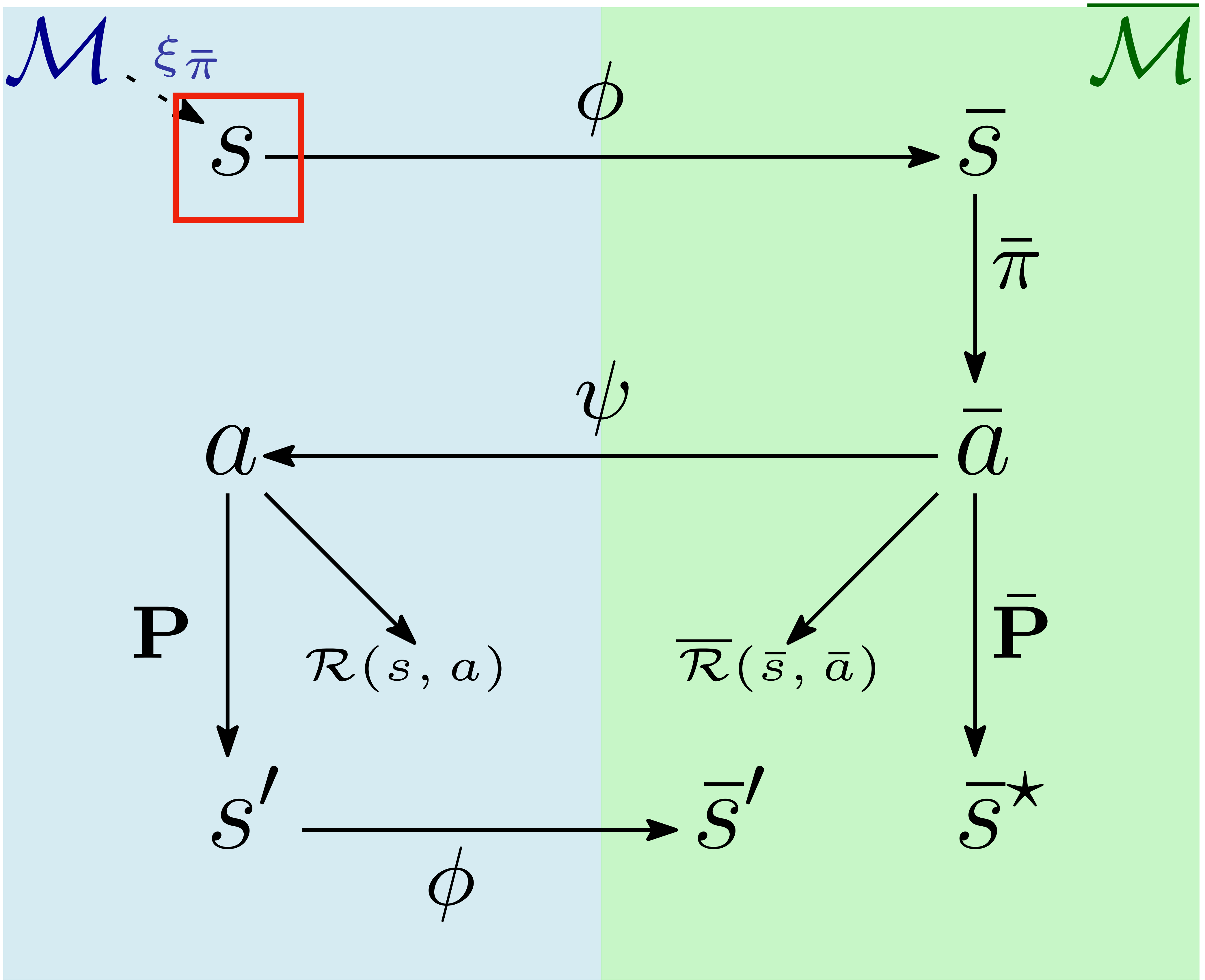
Execution of a latent policy $\bar{\pi}$ in the original model

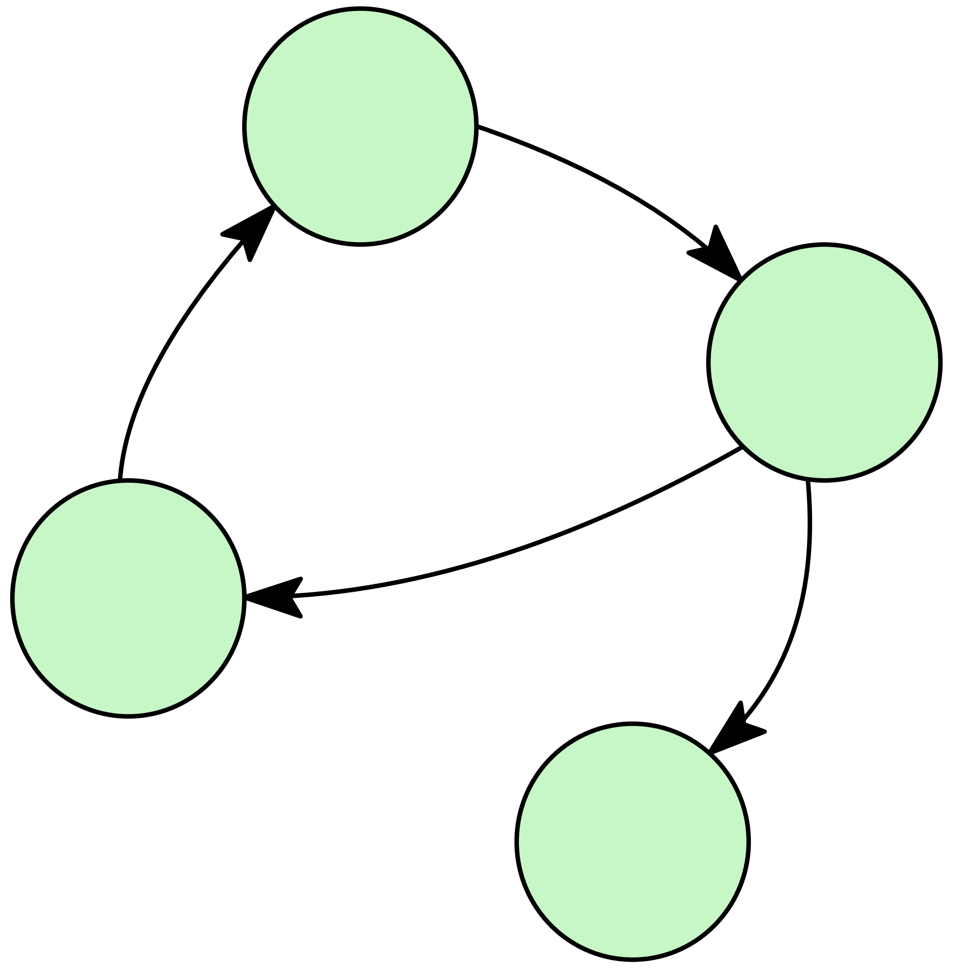




Latent Flow

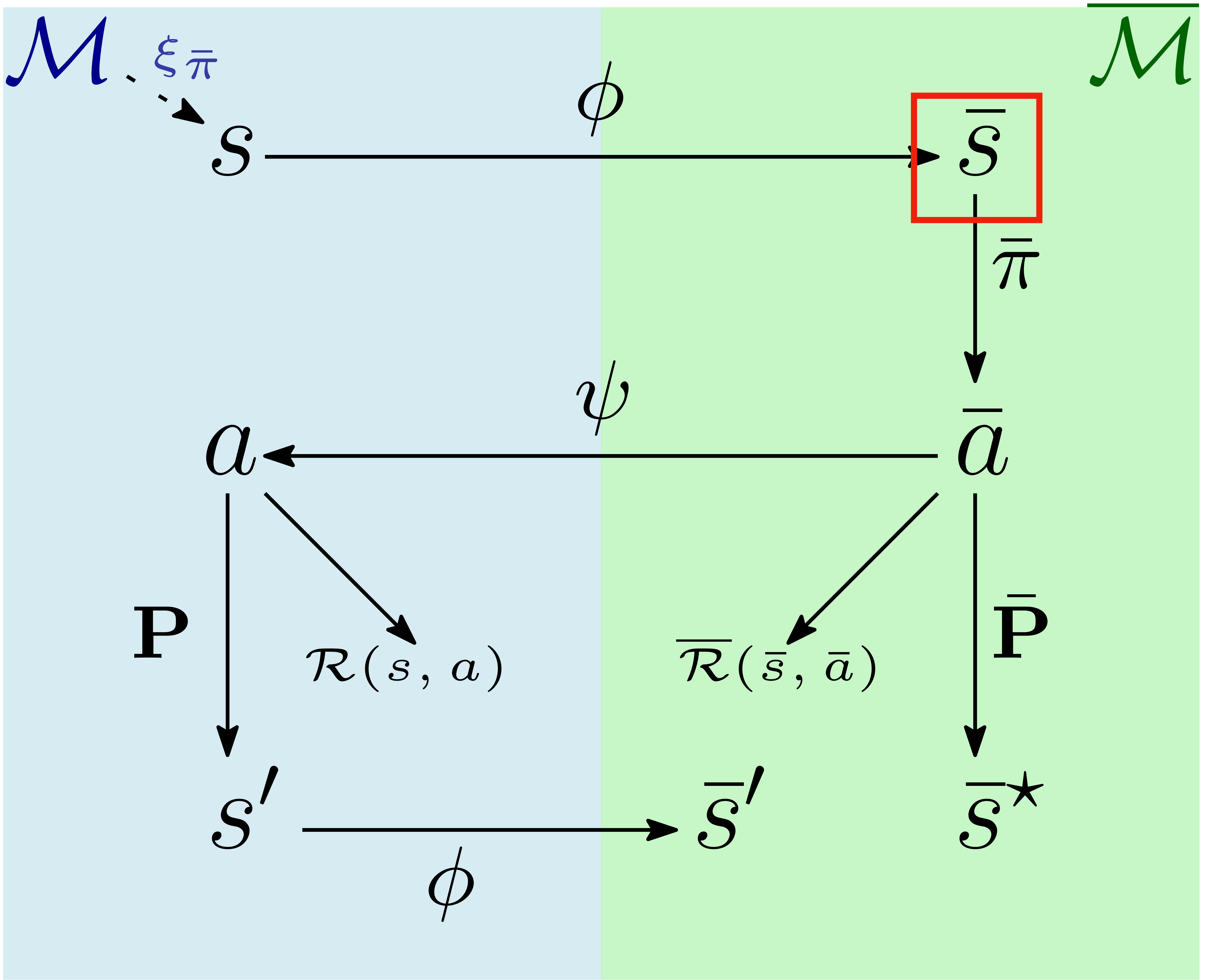
Execution of a latent policy $\bar{\pi}$ in the original model

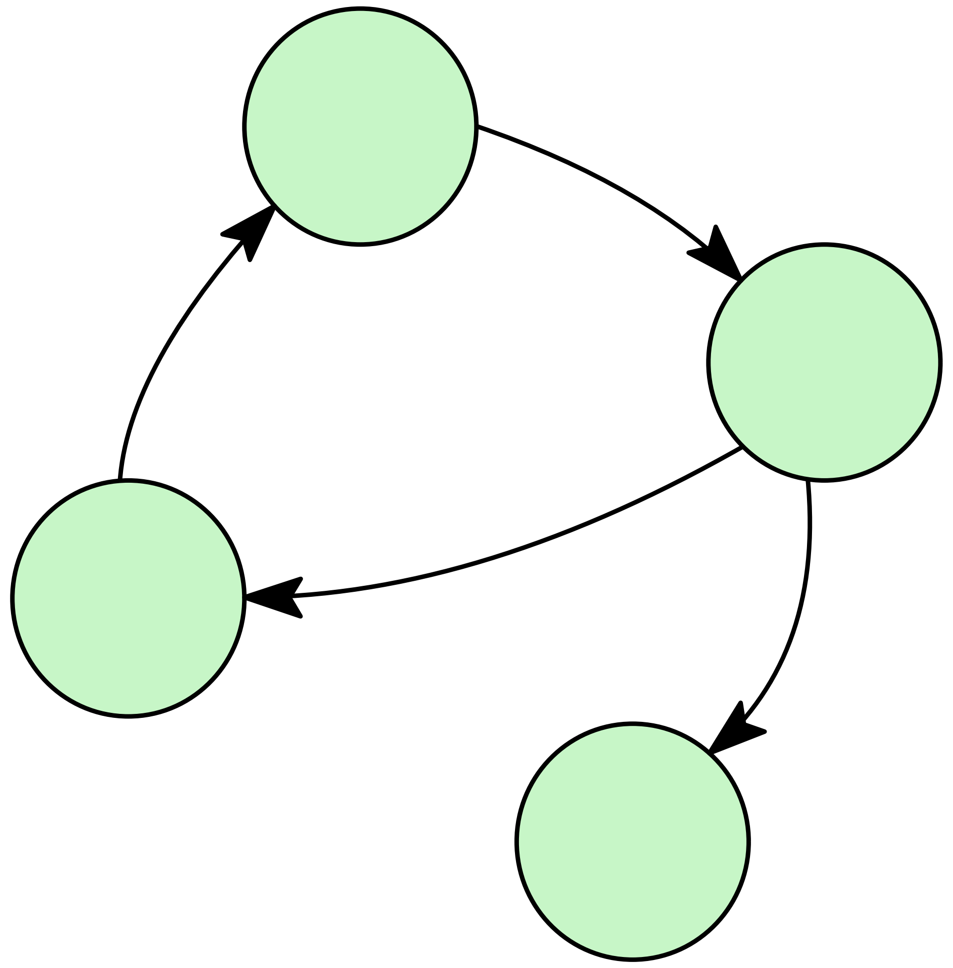




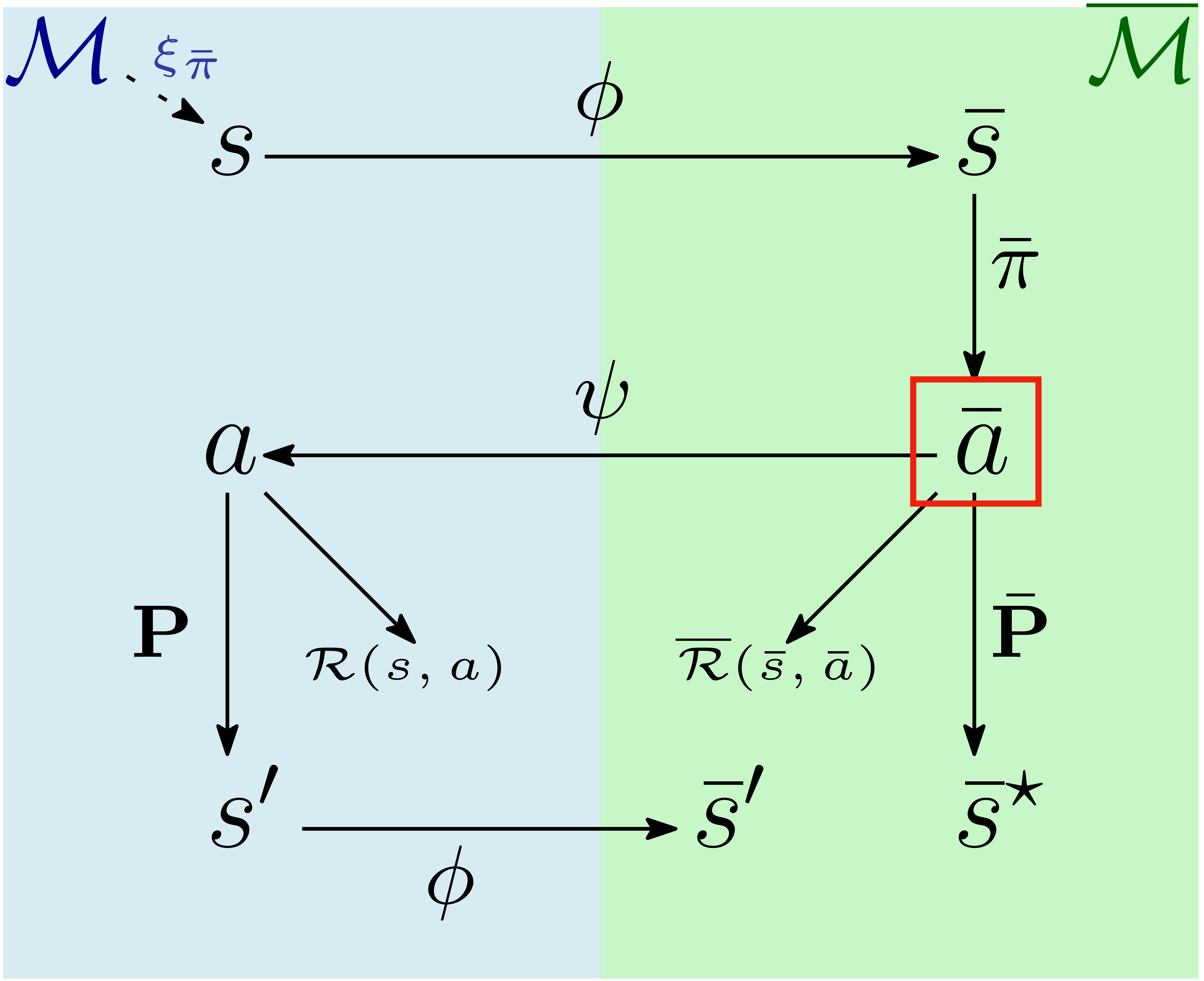
Latent Flow

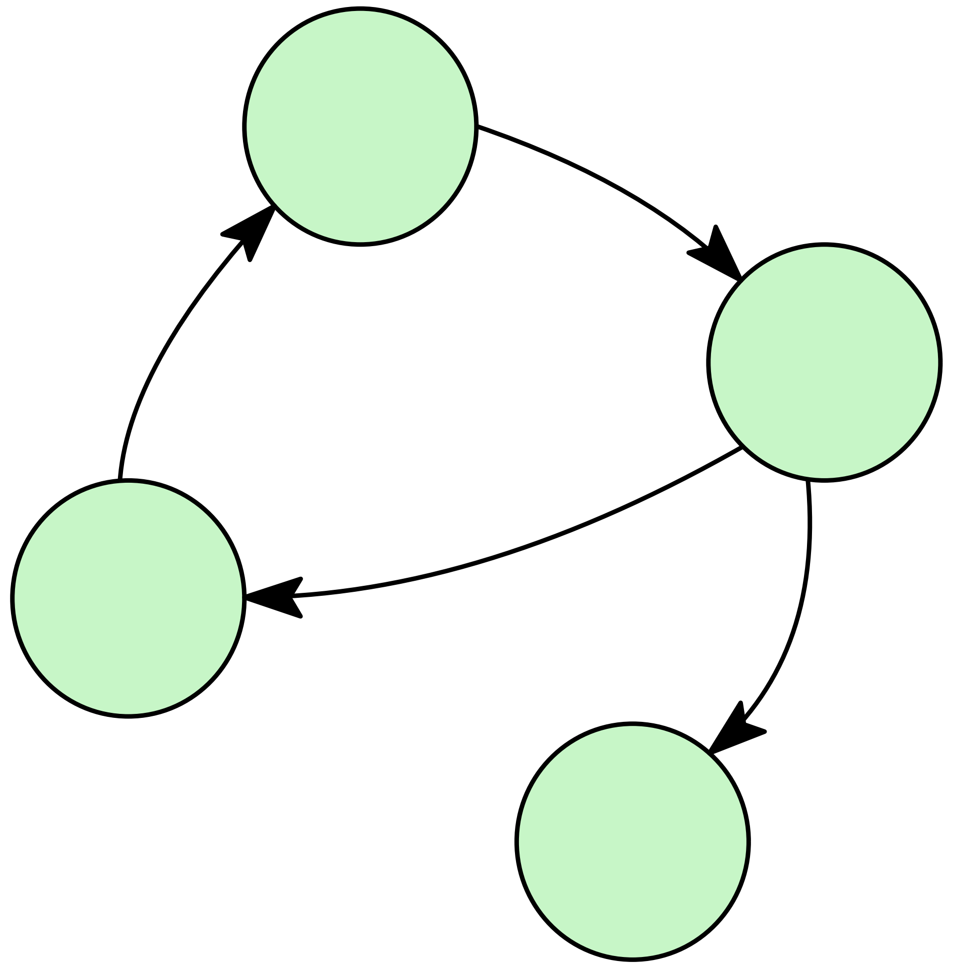
Execution of a latent policy $\bar{\pi}$ in the original model



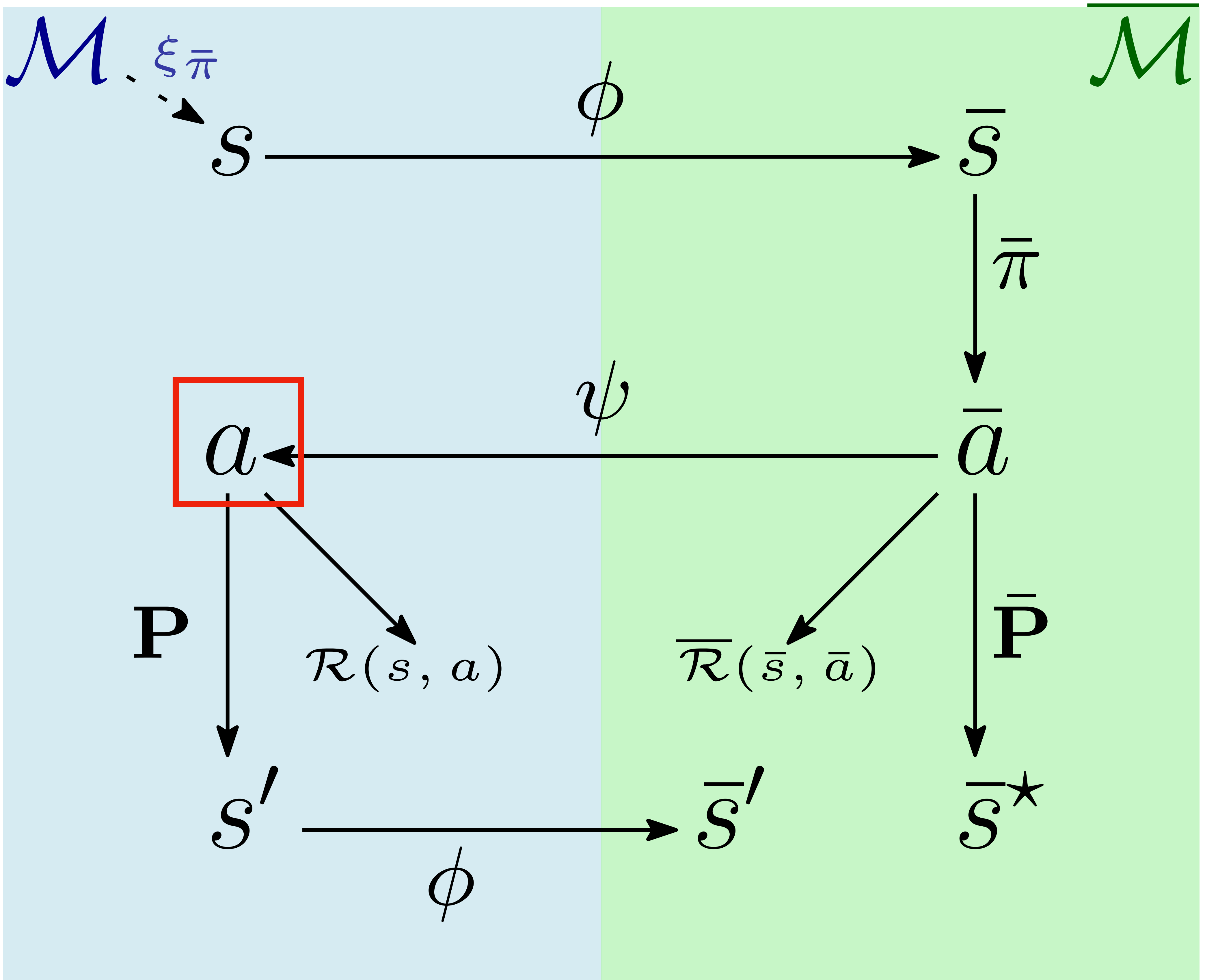


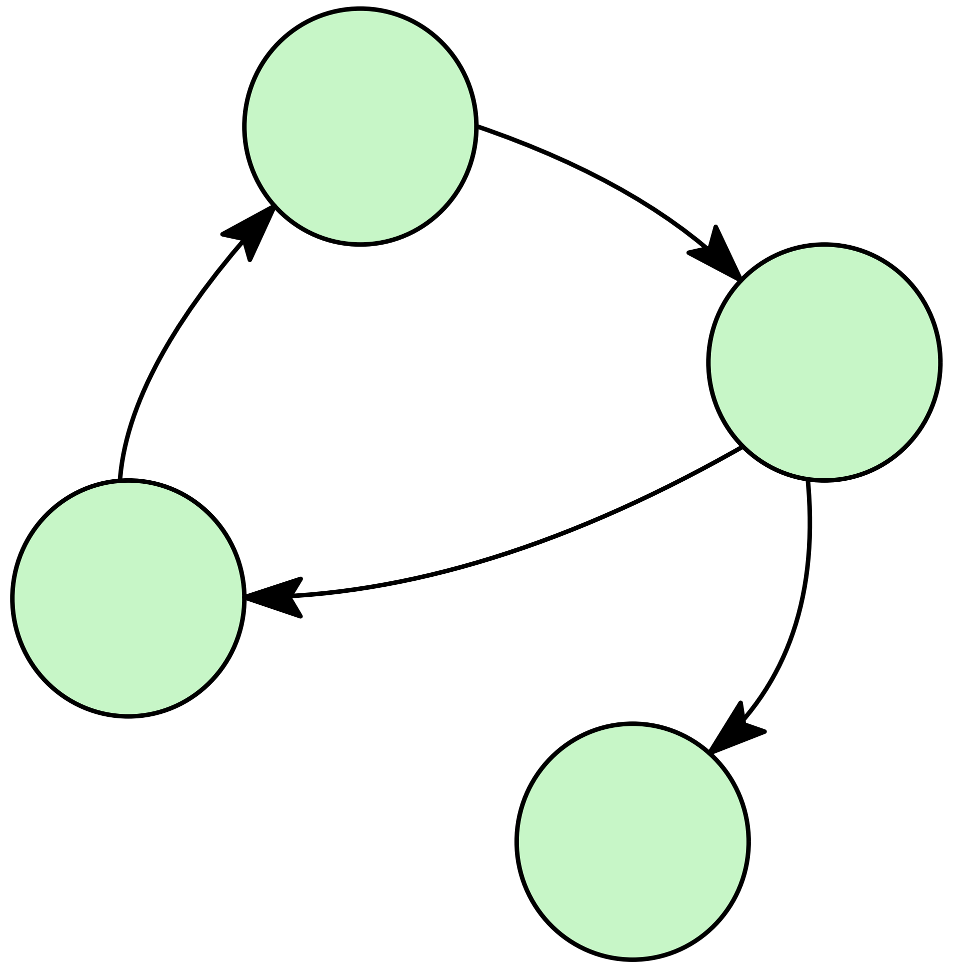
Execution of a latent policy $\bar{\pi}$ in the original model





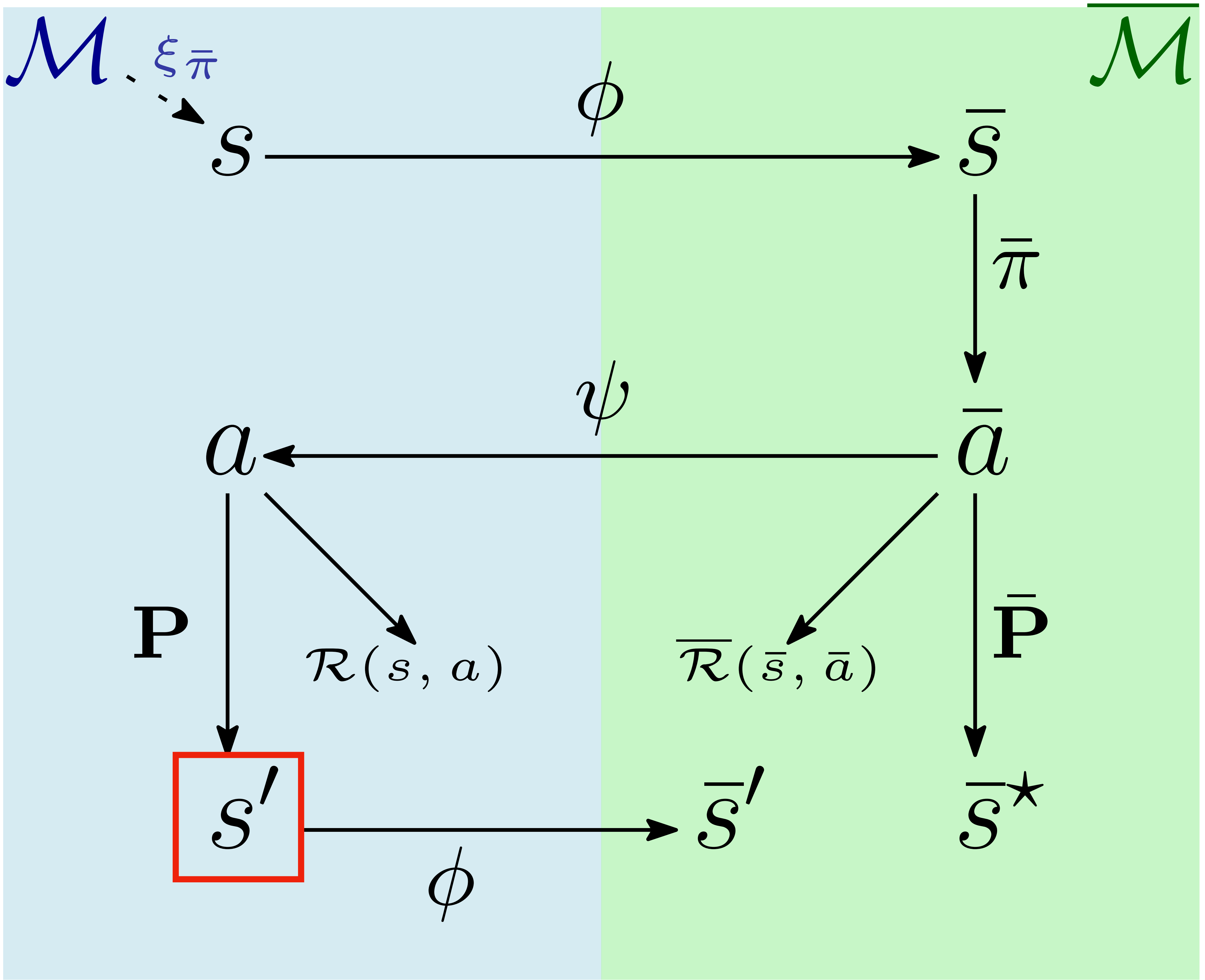
Execution of a latent policy $\bar{\pi}$ in the original model

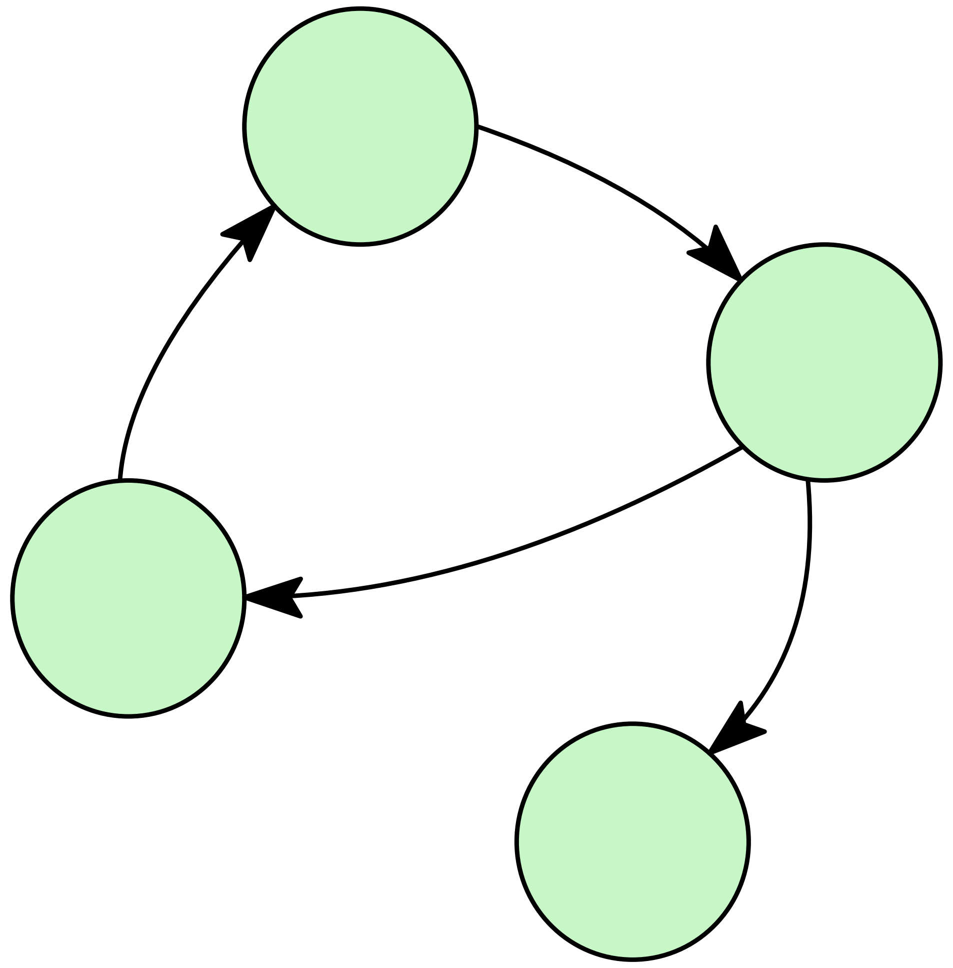




Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model

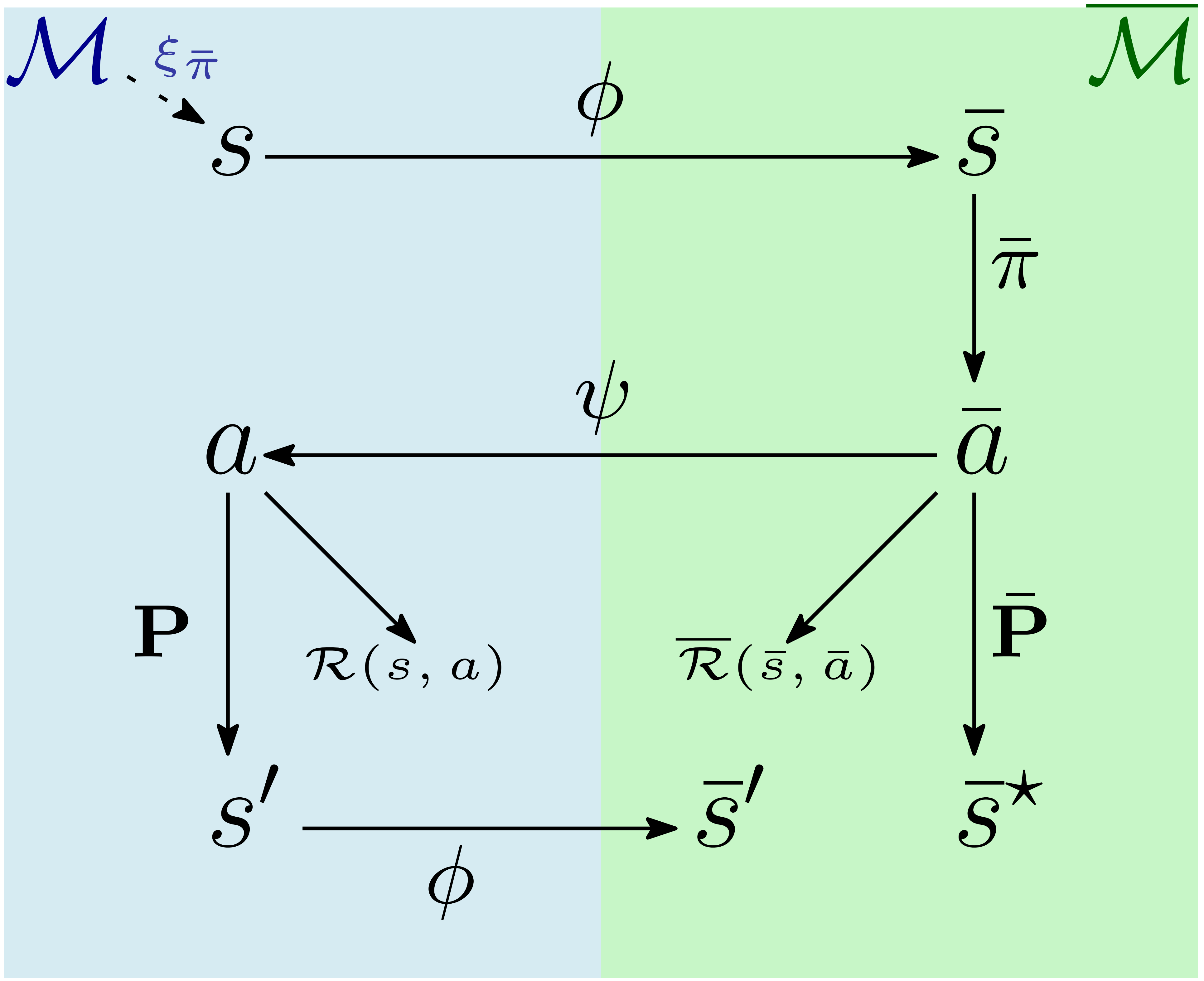




Latent Flow

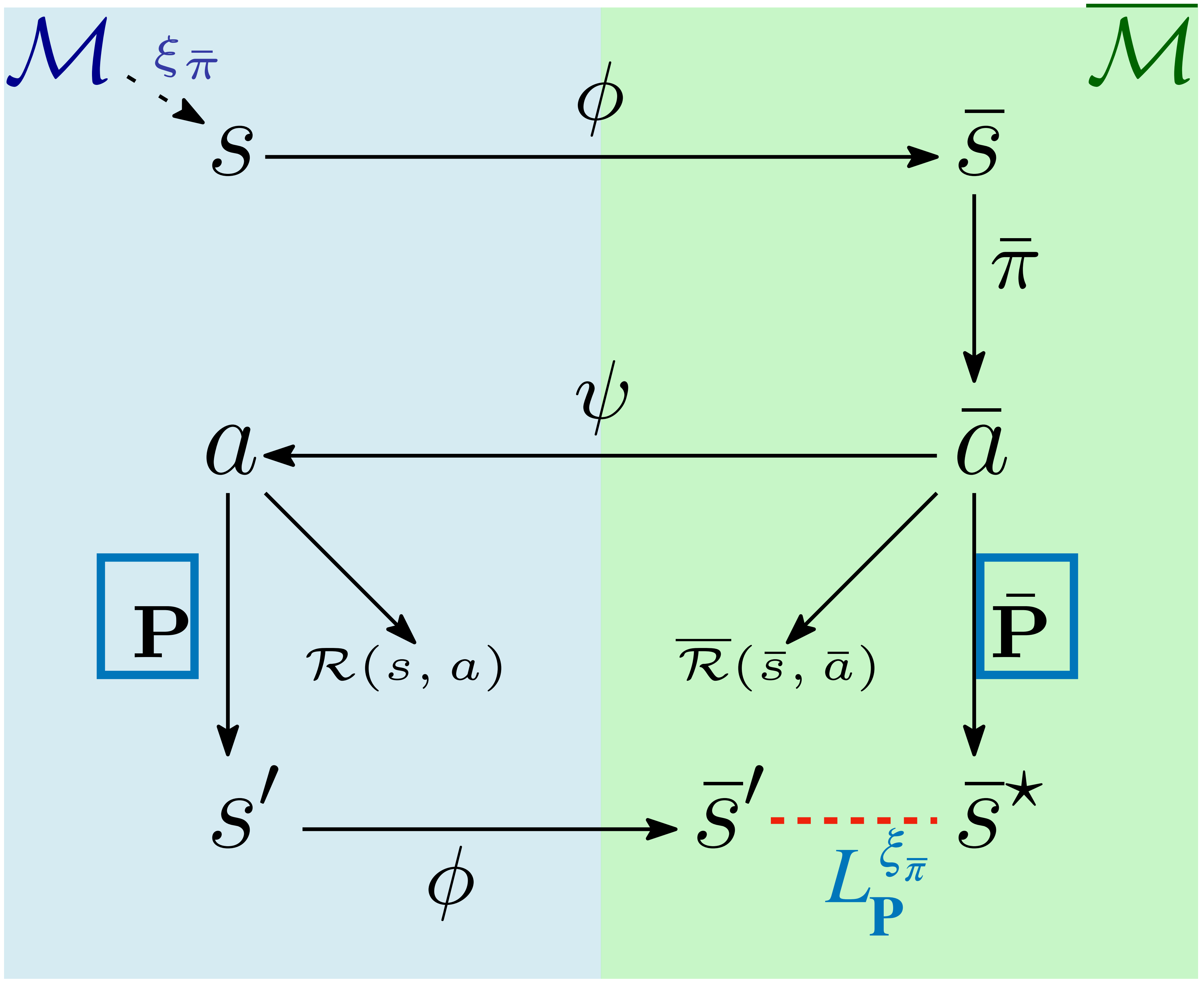
Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**

- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$



Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**

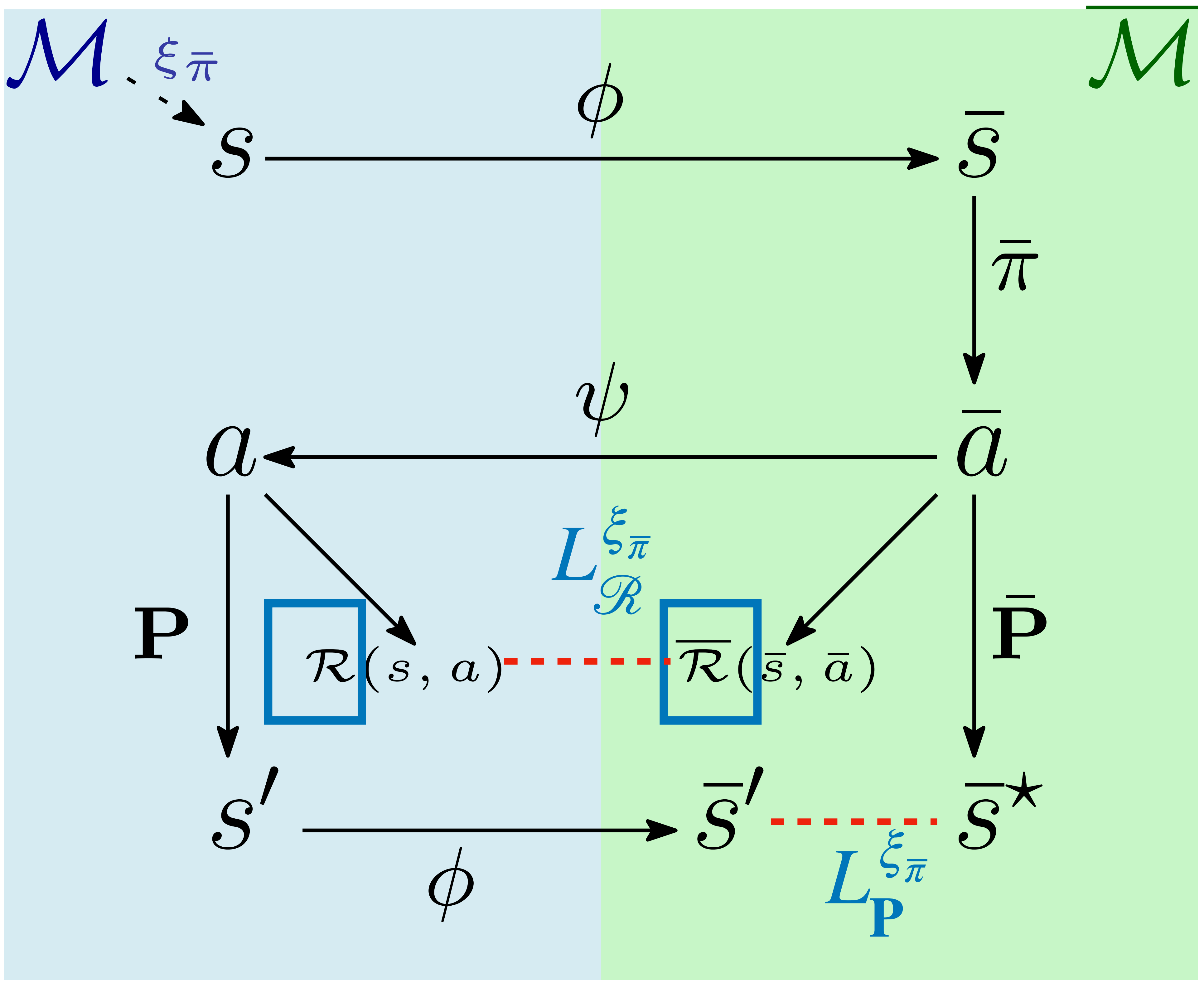


- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{S}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



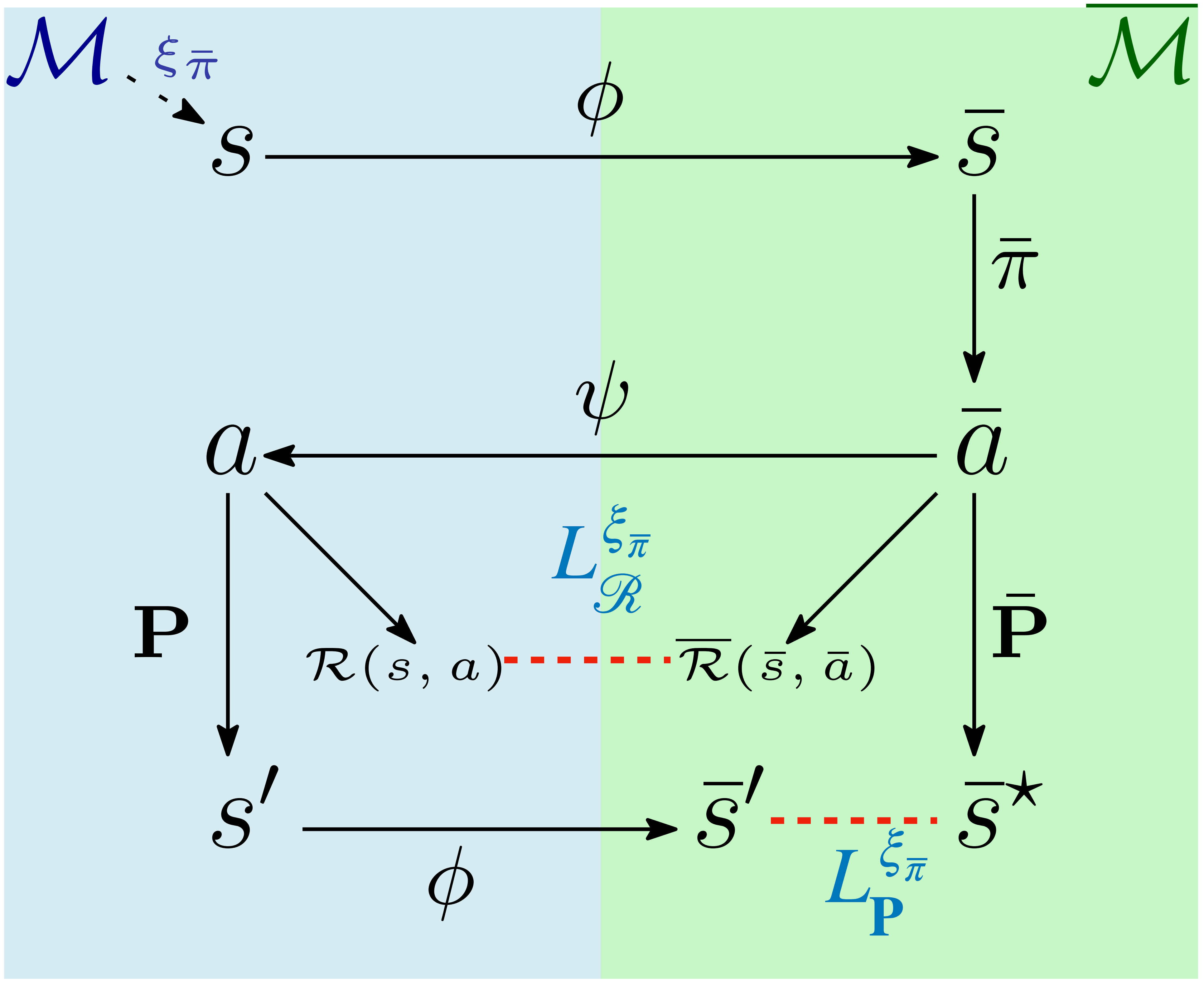
- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

$$L^{\xi_{\bar{\pi}}}_{\mathbf{P}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{S}}}\left(\phi\mathbf{P}\left(\,\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\,\cdot \mid \phi(s), \bar{a}\right)\right)$$

$$L^{\xi_{\bar{\pi}}}_{\mathcal{R}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} \left|\mathcal{R}\left(s, \bar{a}\right) - \overline{\mathcal{R}}\left(\phi(s), \bar{a}\right)\right|$$

Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{S}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

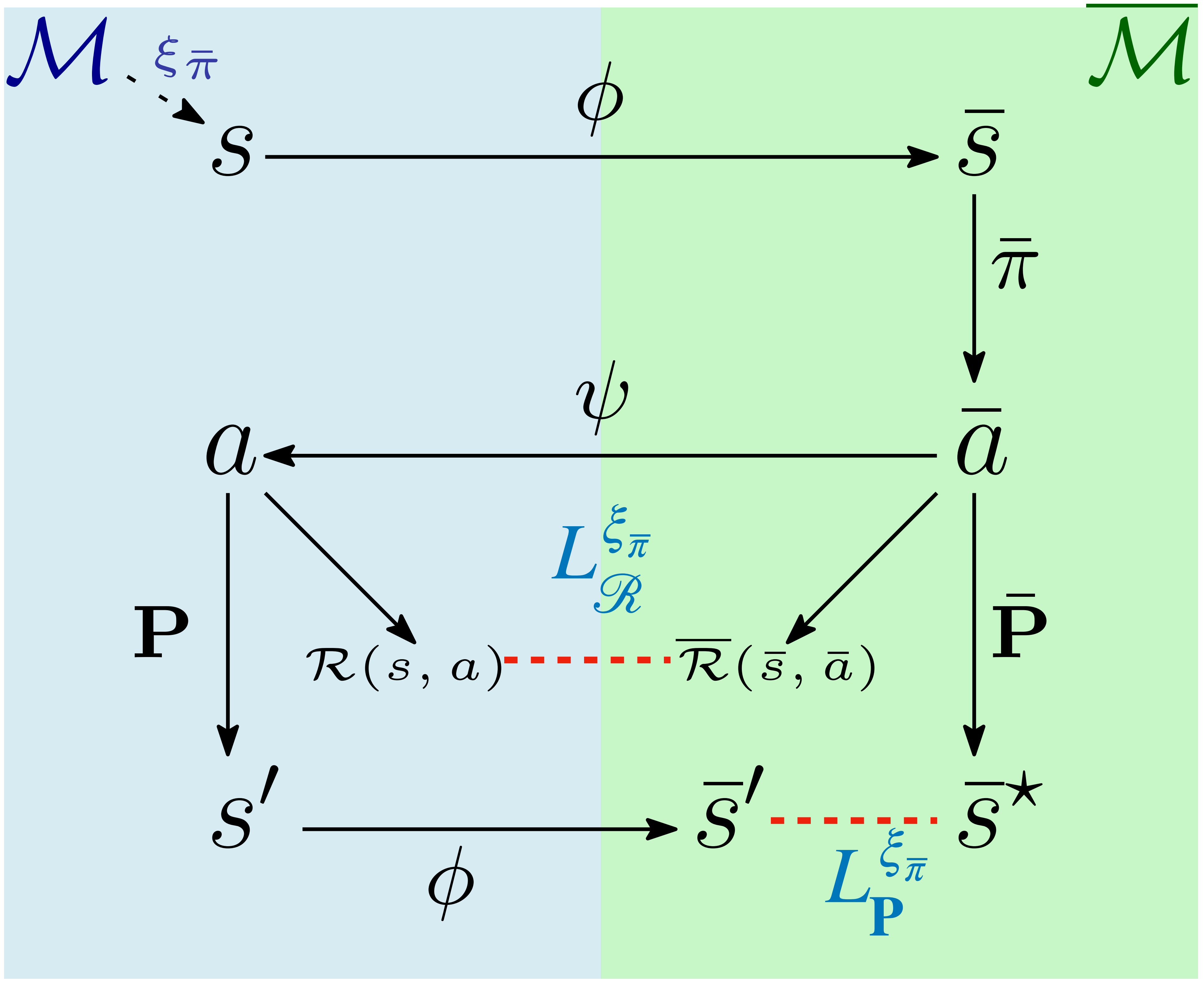
$$L_{\mathcal{R}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} \left|\mathcal{R}\left(s, \bar{a}\right) - \overline{\mathcal{R}}\left(\phi(s), \bar{a}\right)\right|$$

- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$
- ***Representation quality:*** for all $s_1, s_2 \in \mathcal{S}$ such that $\phi(s_1) = \phi(s_2)$

$$\tilde{d}_{\bar{\pi}}\left(s_1, s_2\right) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}\right) \cdot \left(\xi_{\bar{\pi}}^{-1}\left(s_1\right) + \xi_{\bar{\pi}}^{-1}\left(s_2\right)\right)$$

Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{S}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

$$L_{\mathcal{R}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} \left|\mathcal{R}\left(s, \bar{a}\right) - \overline{\mathcal{R}}\left(\phi(s), \bar{a}\right)\right|$$

- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$
- ***Representation quality:*** for all $s_1, s_2 \in \mathcal{S}$ such that $\phi(s_1) = \phi(s_2)$

$$\tilde{d}_{\bar{\pi}}\left(s_1, s_2\right) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}\right) \cdot \left(\xi_{\bar{\pi}}^{-1}\left(s_1\right) + \xi_{\bar{\pi}}^{-1}\left(s_2\right)\right)$$

- **PAC scheme from samples**: let *trace* $\langle s_{0:T}, \bar{a}_{0:T-1}, r_{0:T-1}\rangle \sim \xi_{\bar{\pi}}$, $\epsilon, \delta \in]0,1[$ and

$T \geq \left\lceil \dfrac{-\log\left(\delta/4\right)}{2\epsilon^2} \right\rceil$:

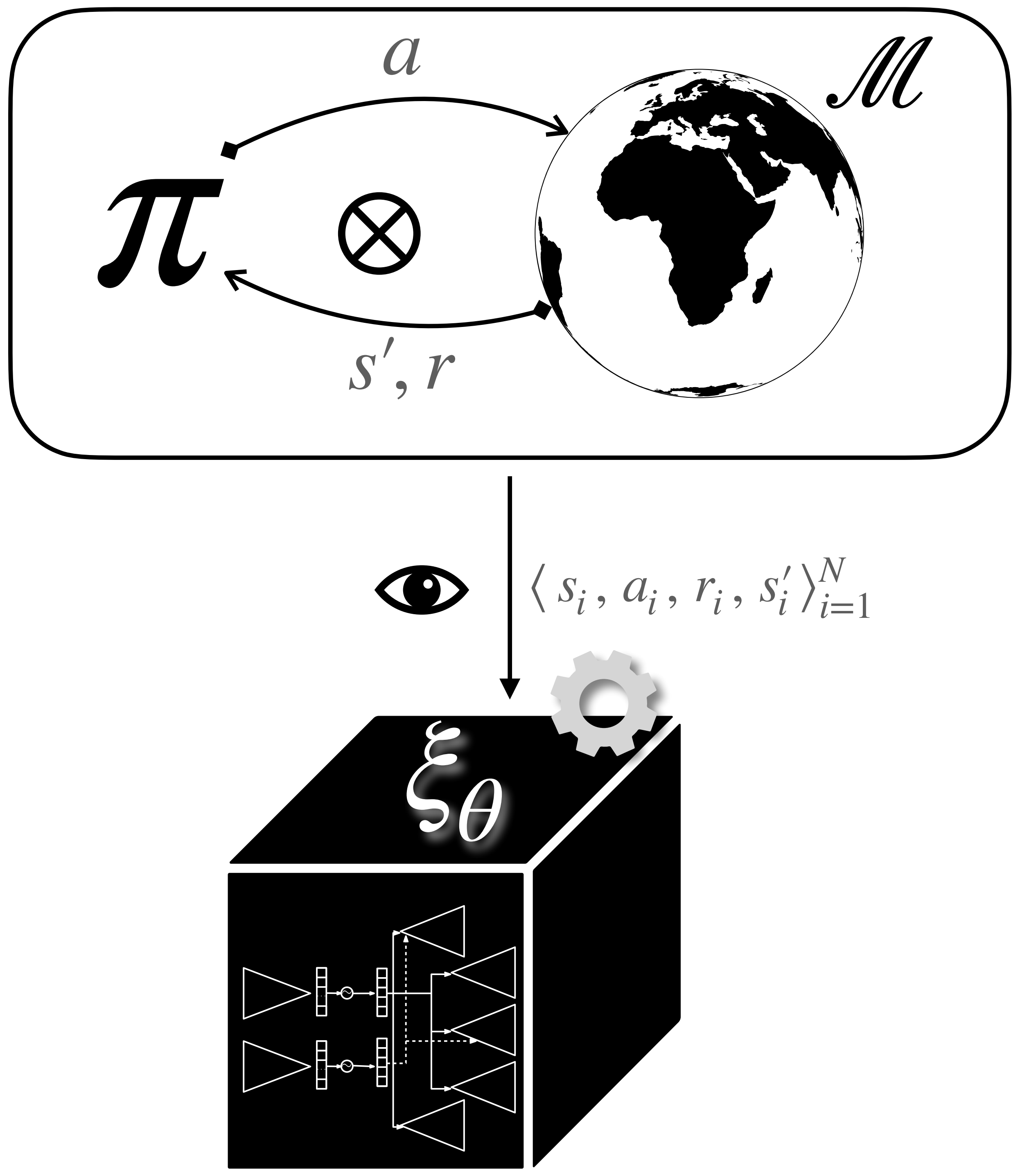
$$\hat{L}_{\mathcal{R}}^{\xi_{\bar{\pi}}} = \frac{1}{T}\sum_{t=0}^{T-1}\left|r_t - \overline{\mathcal{R}}\left(\phi(s_t), \bar{a}_t\right)\right| \quad \text{and} \quad \hat{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \frac{1}{T}\sum_{t=0}^{T-1}\left[1 - \overline{\mathbf{P}}\left(\phi(s_{t+1}) \mid \phi(s_t), \bar{a}_t\right)\right]$$

Then, $\left|L_{\mathcal{R}}^{\xi_{\bar{\pi}}} - \hat{L}_{\mathcal{R}}^{\xi_{\bar{\pi}}}\right| \leq \epsilon$ and $\left|\dot{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}} - \hat{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}}\right| \leq \epsilon$ **with probability** $1-\delta$

Learning the Latent Space Model

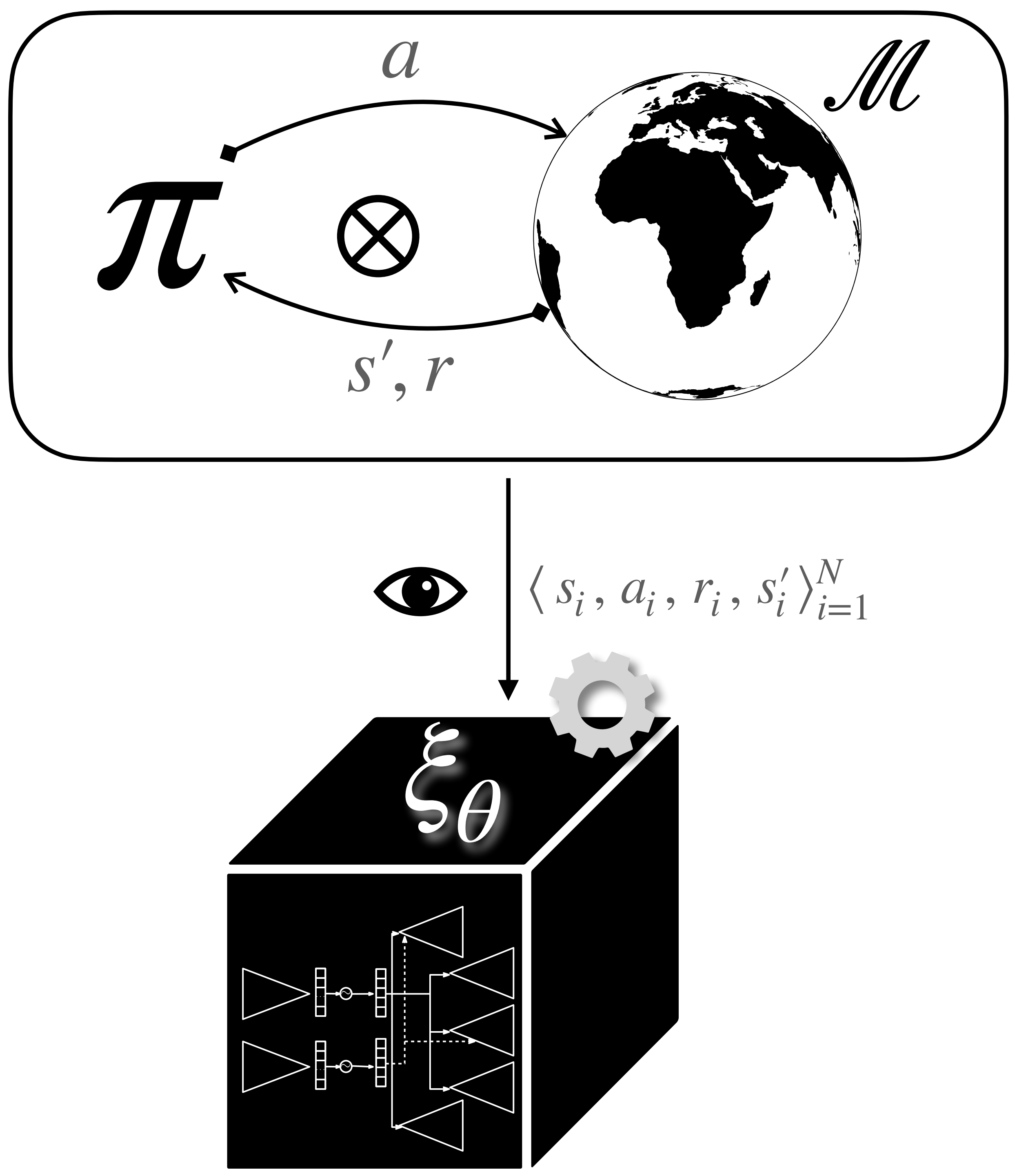
Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$



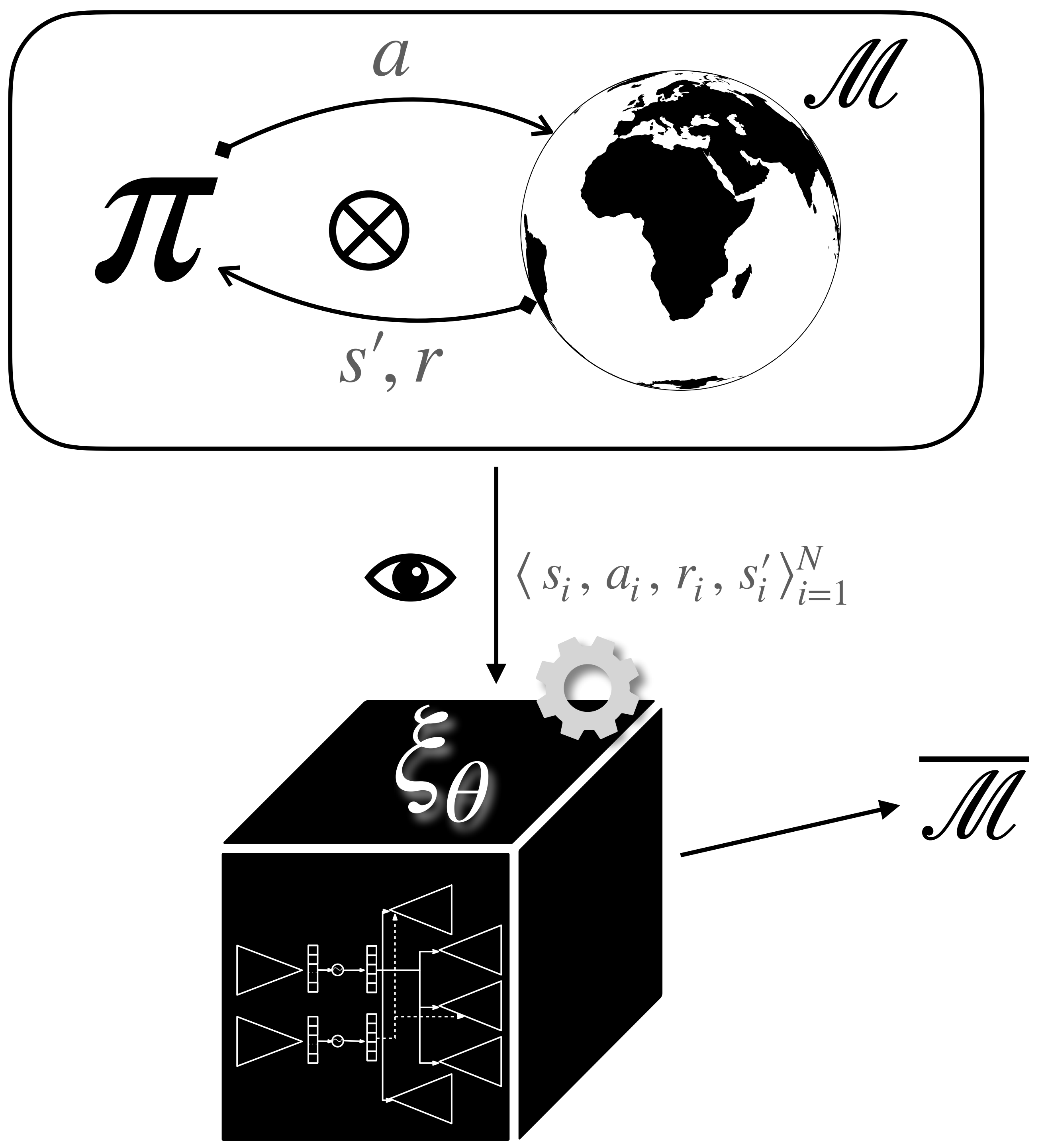
Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:



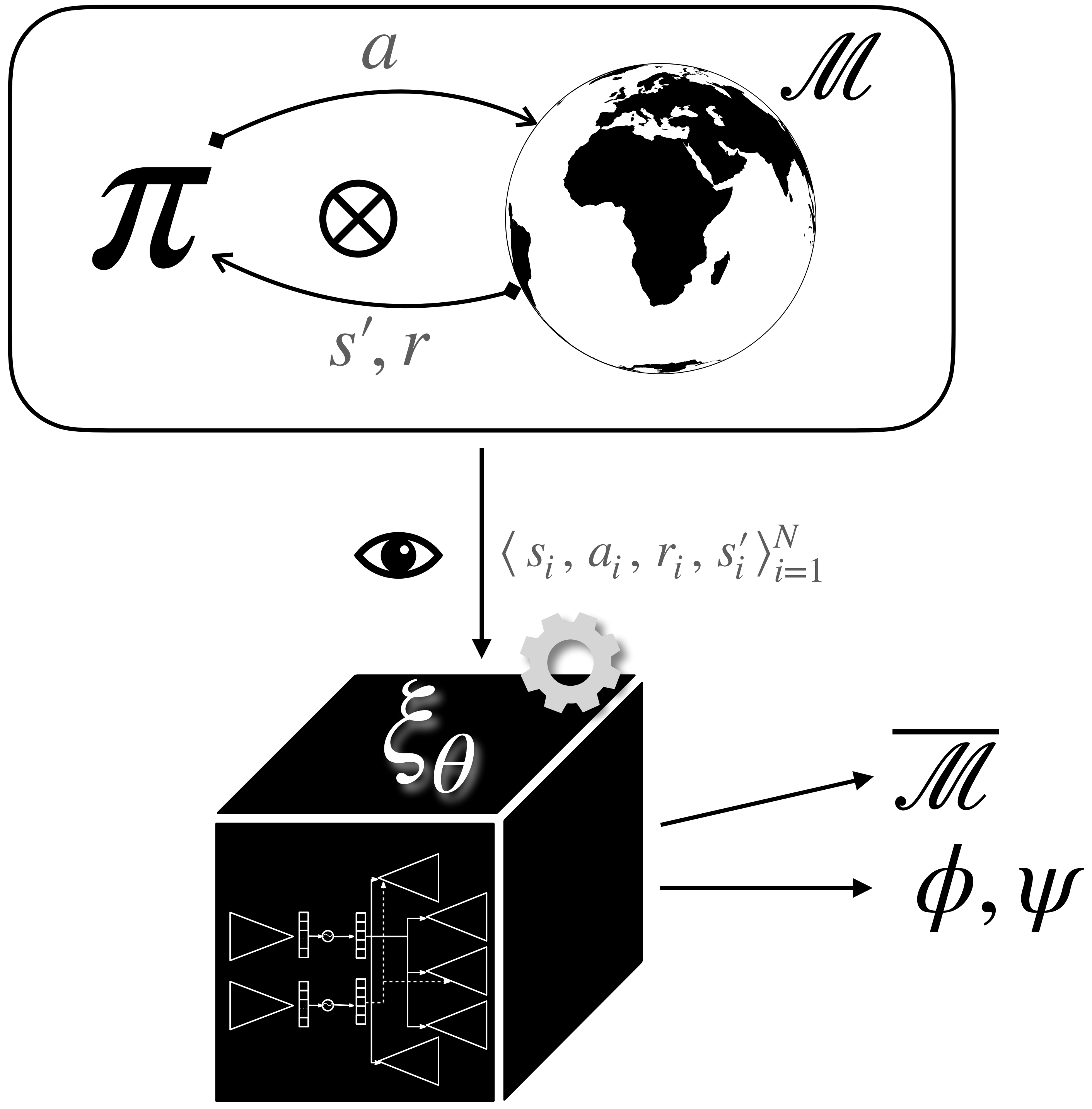
Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$



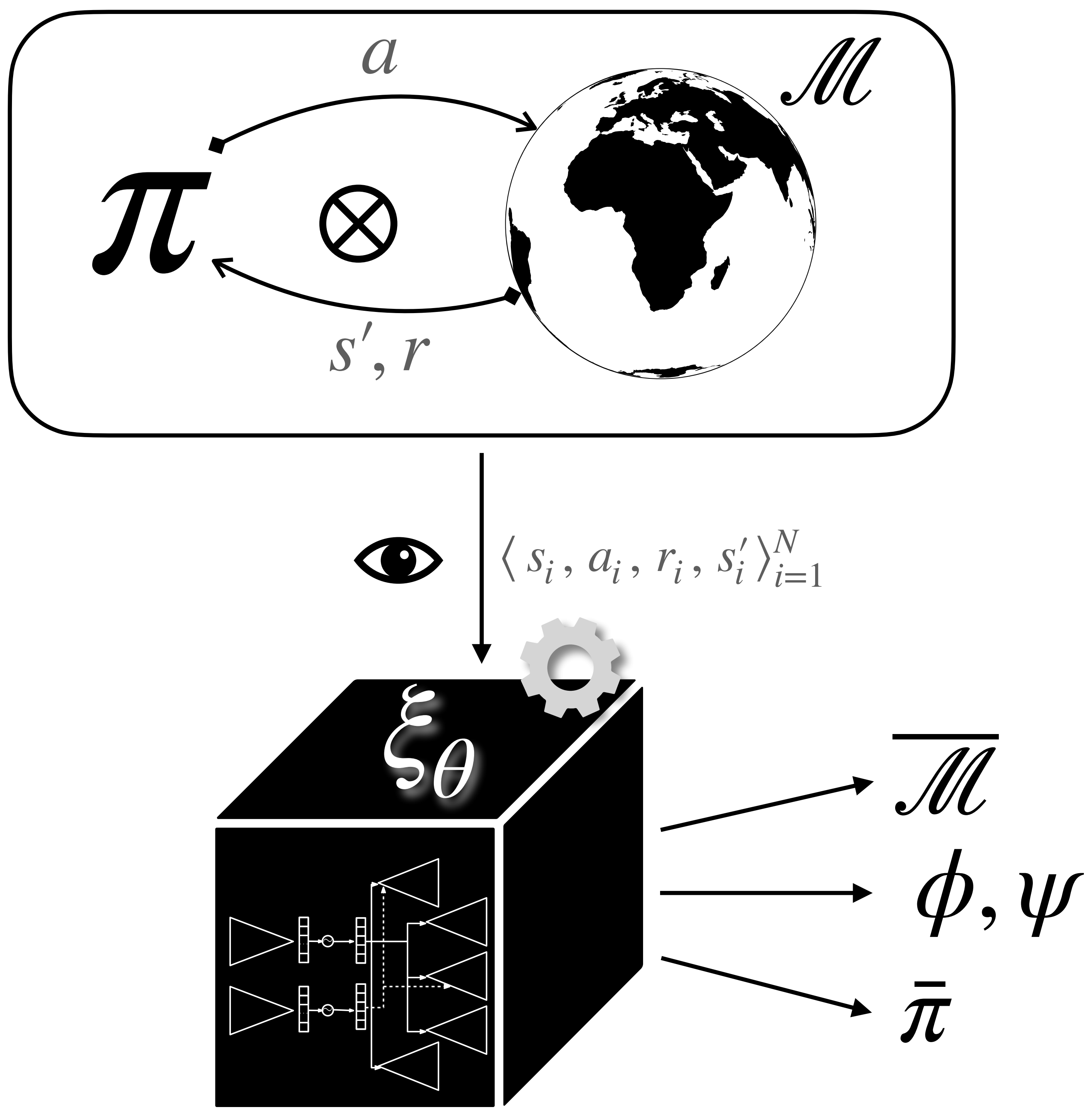
Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ



Learning the Latent Space Model

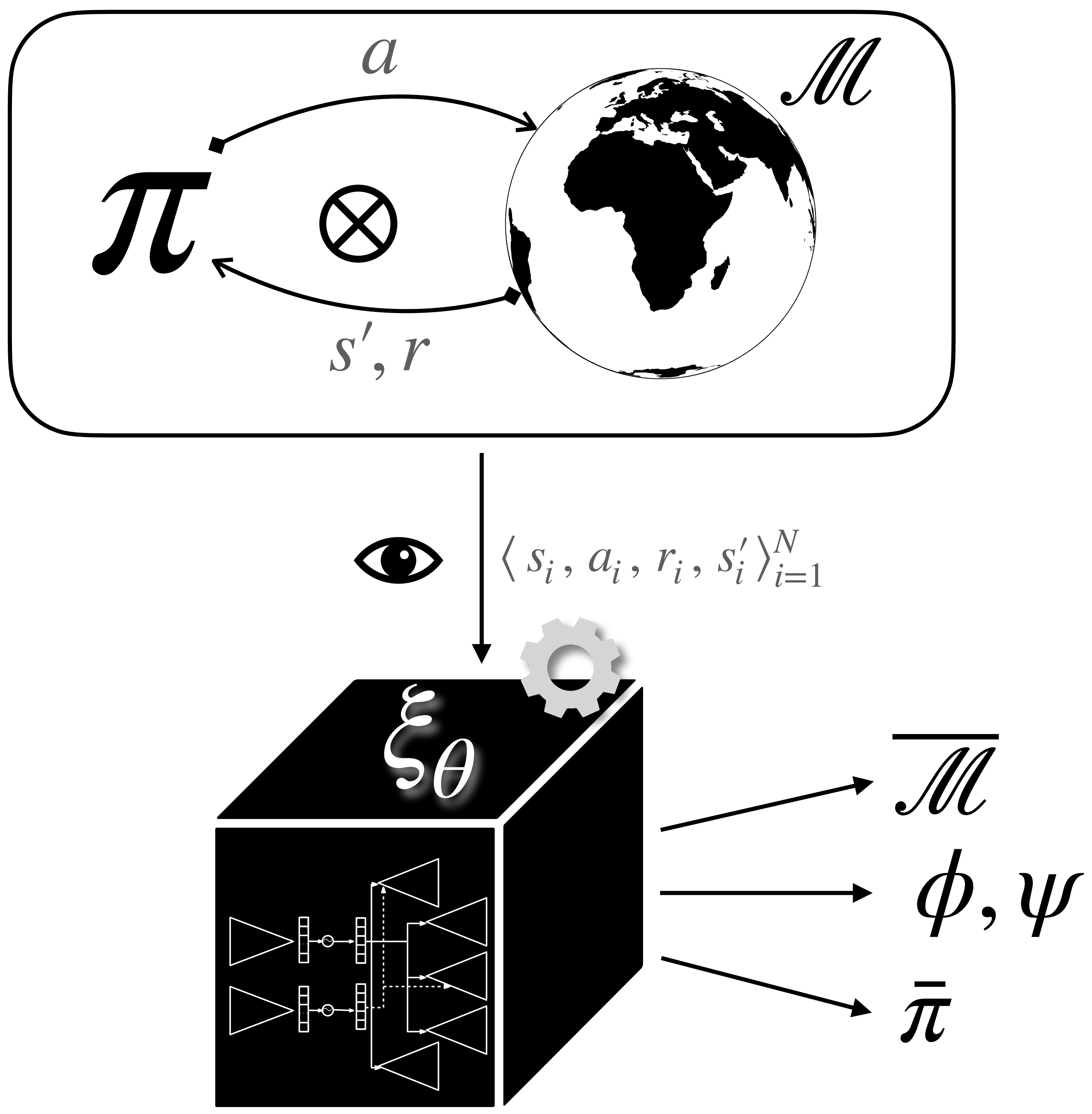
- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π



Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ

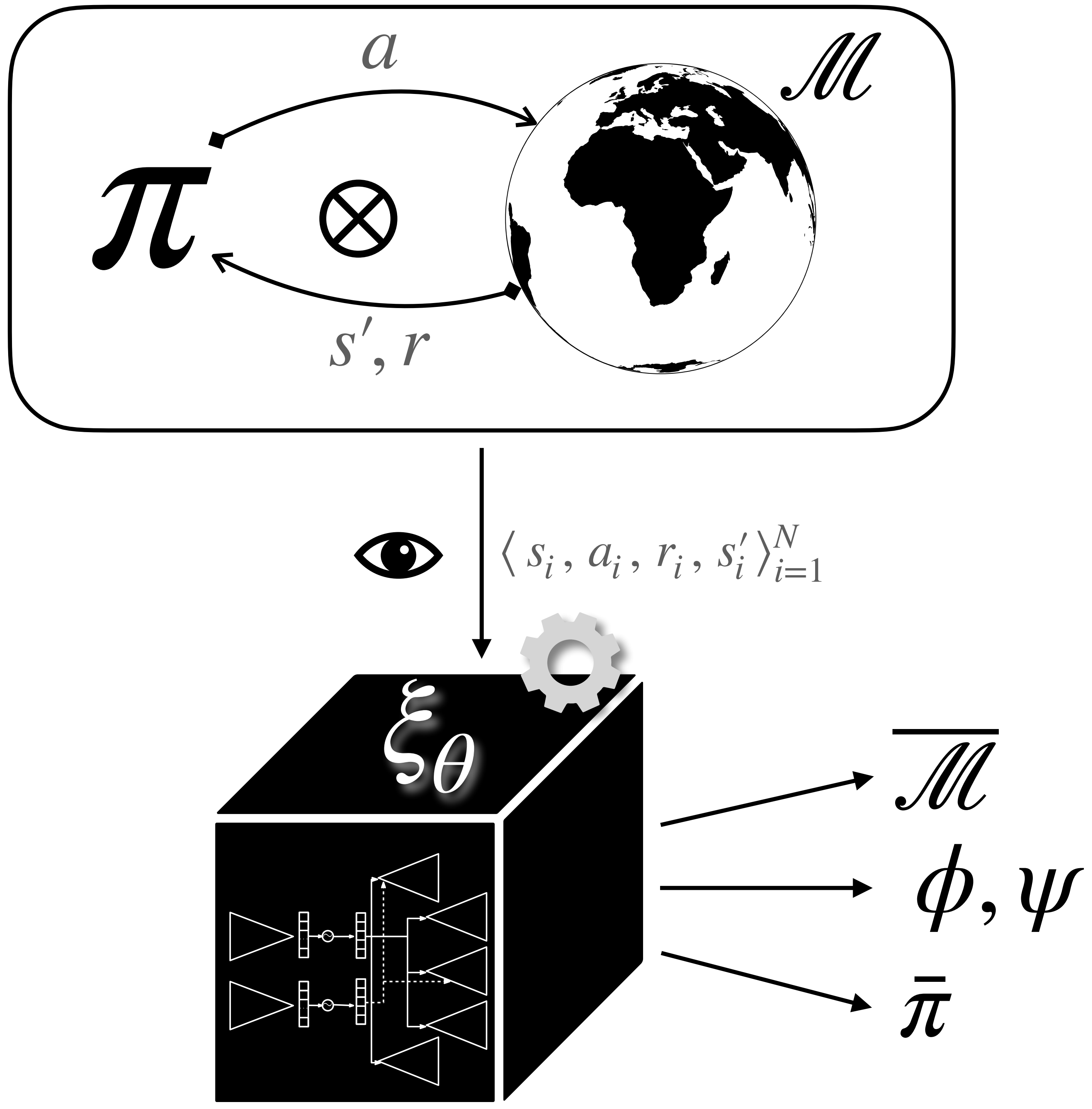
$$\min_\theta D\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$



Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ

$$\min_{\theta} D_{KL}\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

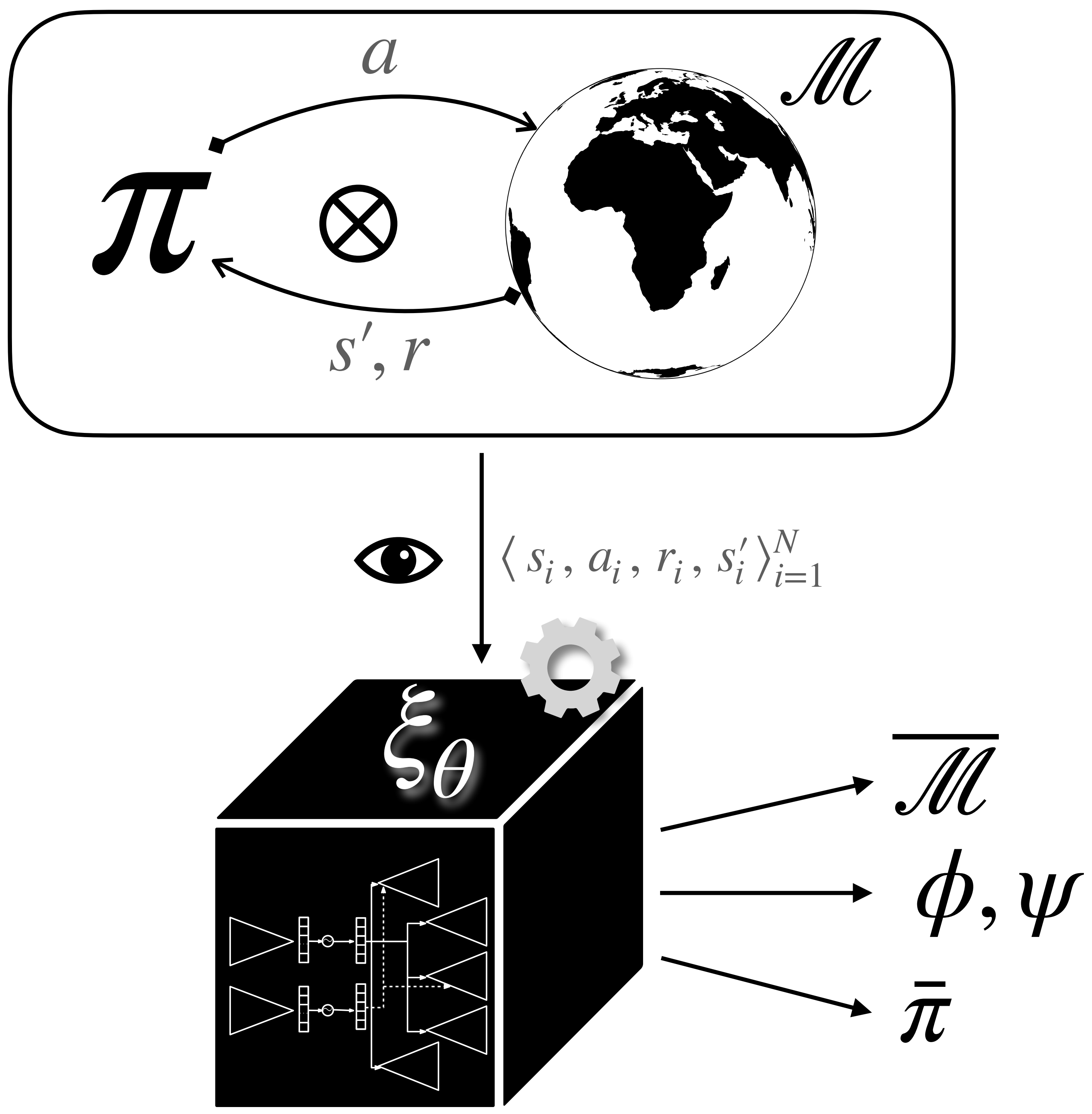


- Choose the *Kullback-Leibler divergence*

$$D_{KL}\left(P, Q\right) = \mathbb{E}_{x \sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π



- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ

$$\min_{\theta} D_{KL}\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

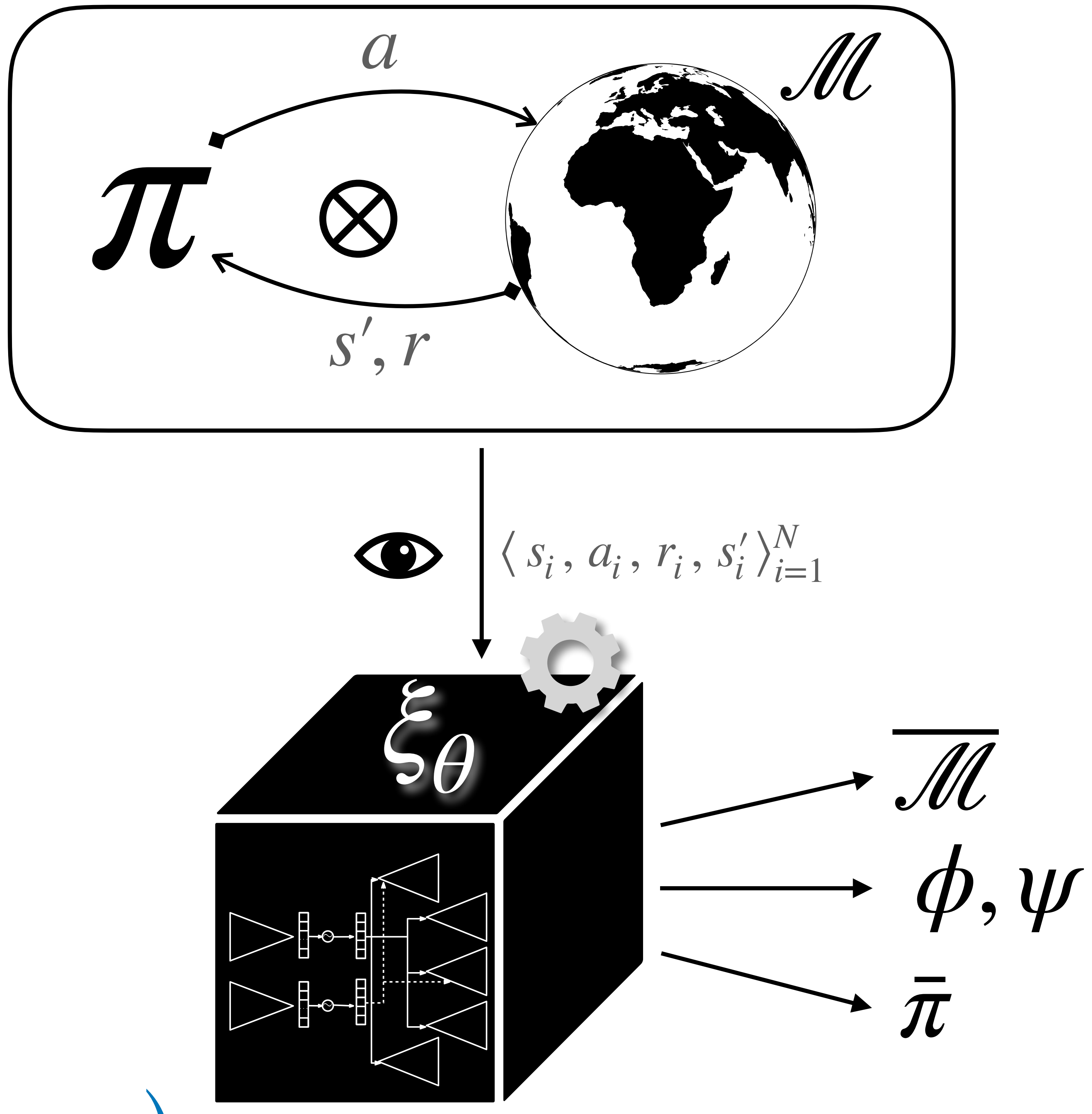
$$\equiv \max_{\theta} \mathbb{E}_{\tau \sim \mathcal{M} \otimes \pi}\left[\log \xi_\theta(\tau)\right]$$

- Choose the *Kullback-Leibler divergence*

$$D_{KL}\left(P, Q\right) = \mathbb{E}_{x \sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ



$$\min_\theta D_{KL}\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

$$\equiv \max_\theta \mathbb{E}_{\tau \sim \mathcal{M} \otimes \pi}\left[\log \xi_\theta(\tau)\right] \geq \max_{\iota, \theta} ELBO\left(\overline{\mathcal{M}}_\theta, \phi_\iota, \psi_\theta\right)$$

(Kingma & Welling, 2014; Hoffman et al., 2013)

- Choose the *Kullback-Leibler divergence*

$$D_{KL}\left(P, Q\right) = \mathbb{E}_{x \sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

$$\max_{\iota,\theta} ELBO\left(\overline{\mathscr{M}}_{\theta}, \phi_{\iota}, \psi_{\theta}\right)$$

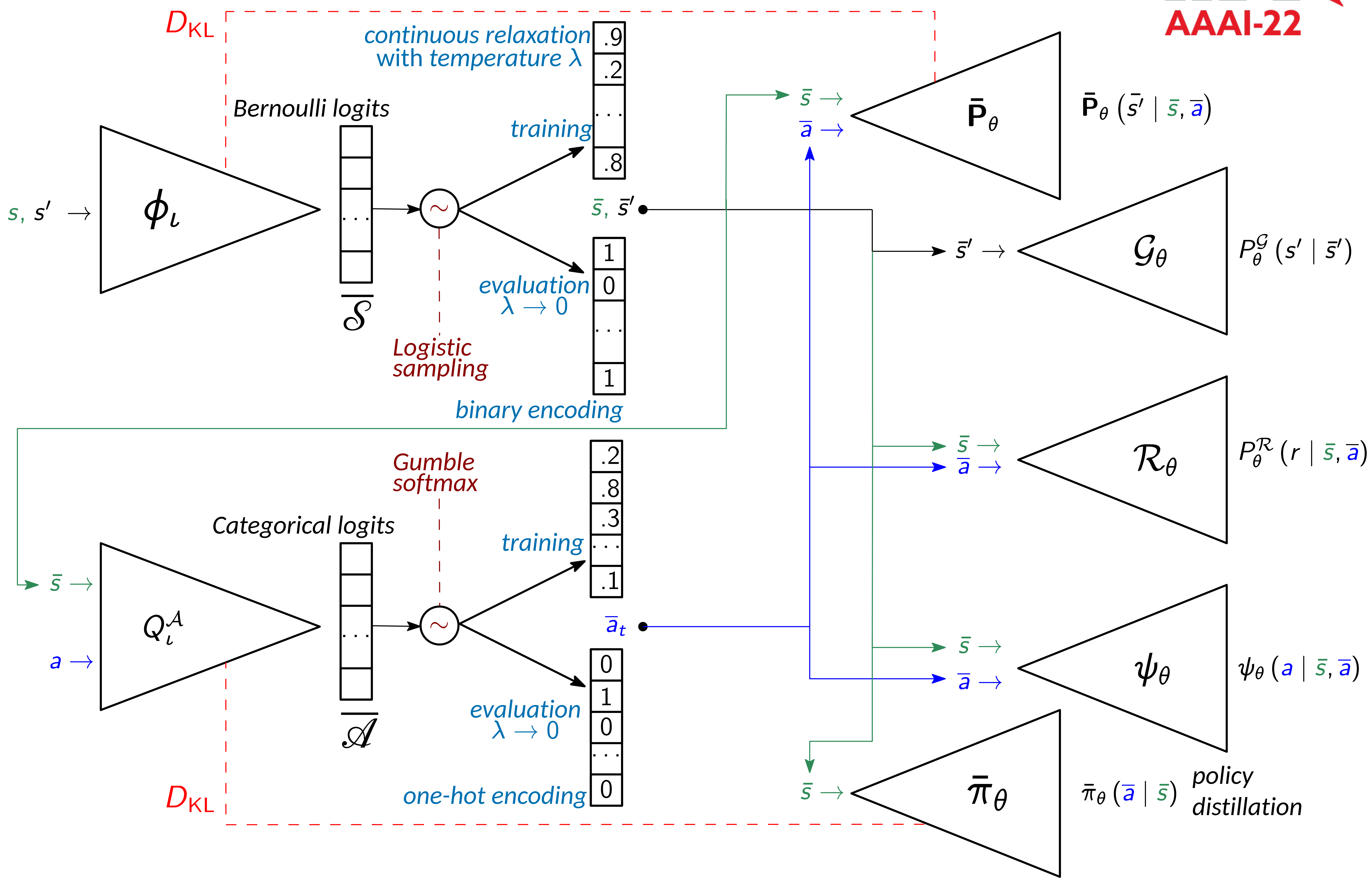
$$\max_{\iota,\theta} ELBO\left(\overline{\mathscr{M}}_{\theta}, \phi_{\iota}, \psi_{\theta}\right) = -\min_{\iota,\theta}\left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

Variational Markov Decision Process

$$\max_{\iota,\theta} ELBO\left(\overline{\mathcal{M}}_{\theta}, \phi_{\iota}, \psi_{\theta}\right) = -\min_{\iota,\theta}\left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

$$\mathbf{D}_{\iota,\theta} = -\mathop{\mathbb{E}}_{\substack{s,a,r,s'\sim\xi_\pi \\ \bar{s},\bar{s}'\sim\phi_\iota(\cdot|s,s') \\ \bar{a}\sim Q^{\mathcal{A}}_\iota(\cdot|\bar{s},a)}}\left[\log P^{\mathcal{G}}_\theta(s' \mid \bar{s}') + \log\psi_\theta(a \mid \bar{s},\bar{a}) + \log P^{\mathcal{R}}_\theta(r \mid \bar{s},\bar{a})\right]$$

$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s'\sim\xi_\pi \\ \bar{s}\sim\phi_\iota(\cdot|s) \\ \bar{a}\sim Q^{\mathcal{A}}_\iota(\cdot|\bar{s},a)}}\left[D_{\mathrm{KL}}\left(\phi_\iota(\cdot \mid s') \,\|\, \overline{\mathbf{P}}_\theta(\cdot \mid \bar{s},\bar{a})\right) + D_{\mathrm{KL}}\left(Q^{\mathcal{A}}_\iota(\cdot \mid \bar{s},a) \,\|\, \bar{\pi}_\theta(\cdot \mid \bar{s})\right)\right]$$



Variational Markov Decision Process

$$\max_{\iota,\theta} \; ELBO\left(\overline{\mathcal{M}}_\theta, \phi_\iota, \psi_\theta\right) = -\min_{\iota,\theta} \left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

$$\mathbf{D}_{\iota,\theta} = -\mathop{\mathbb{E}}_{\substack{s,a,r,s'\sim\xi_\pi \\ \bar{s},\bar{s}'\sim\phi_\iota(\cdot\mid s,s') \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot\mid\bar{s},a)}} \left[\log P_\theta^{\mathcal{G}}(s'\mid\bar{s}') + \log\psi_\theta(a\mid\bar{s},\bar{a}) + \log P_\theta^{\mathcal{R}}(r\mid\bar{s},\bar{a})\right]$$

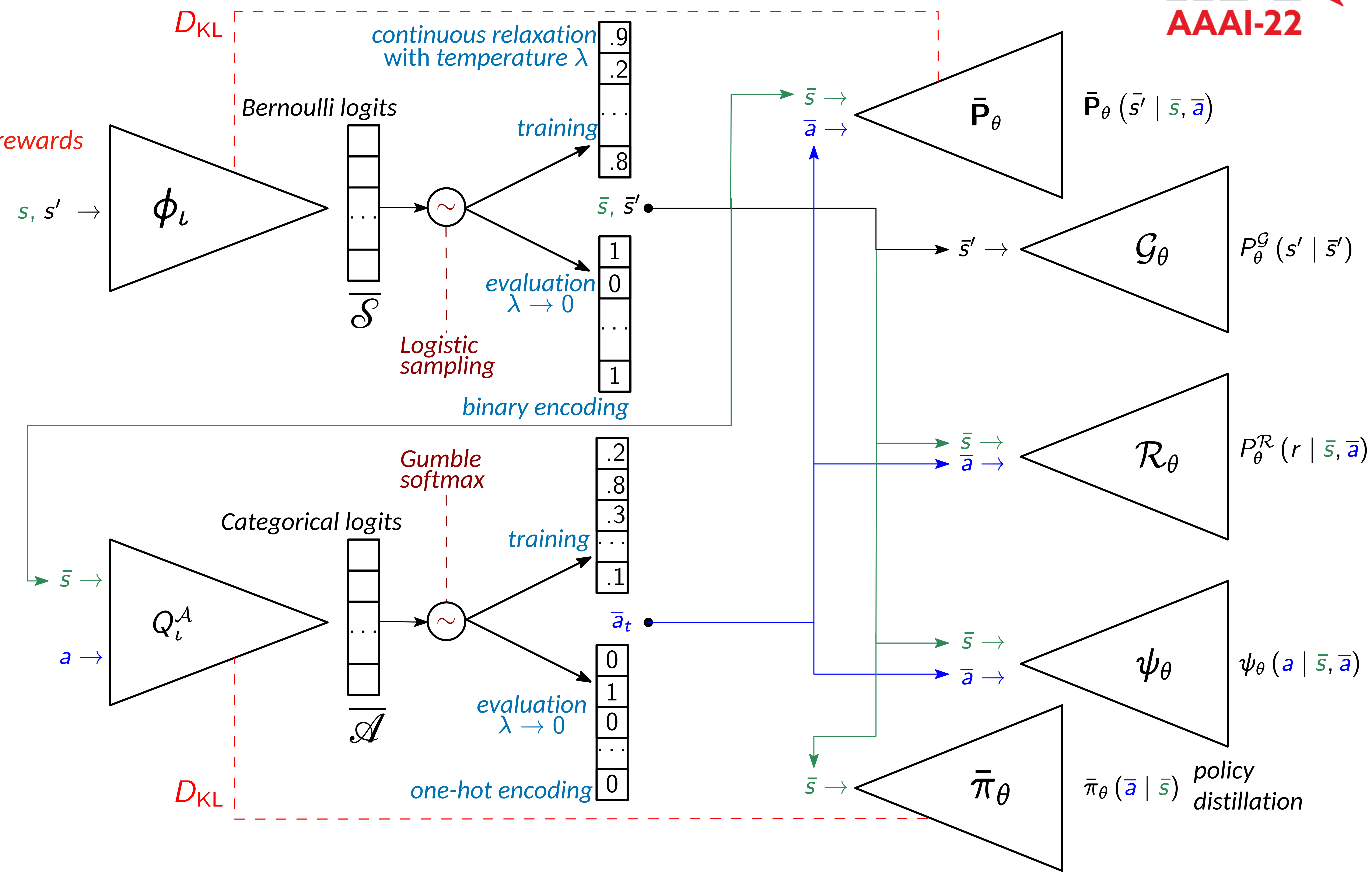
Log-likelihood of rewards

$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s'\sim\xi_\pi \\ \bar{s}\sim\phi_\iota(\cdot\mid s) \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot\mid\bar{s},a)}} \left[D_{KL}\left(\phi_\iota(\cdot\mid s') \parallel \bar{\mathbf{P}}_\theta(\cdot\mid\bar{s},\bar{a})\right) + D_{\mathrm{KL}}\left(Q_\iota^{\mathcal{A}}(\cdot\mid\bar{s},a) \parallel \bar{\pi}_\theta(\cdot\mid\bar{s})\right)\right]$$

Variational version of local transition loss

$$L_{\mathbf{P}}^{\xi_\pi} = \mathbb{E}_{s,\bar{a}\sim\xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}(\cdot\mid s,\bar{a}), \bar{\mathbf{P}}(\cdot\mid\phi(s),\bar{a})\right)$$
$$\leq \mathbb{E}_{s,\bar{a},s'\sim\xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi(\cdot\mid s'), \bar{\mathbf{P}}(\cdot\mid\phi(s),\bar{a})\right)$$

- Variational proxies to local losses



Variational Markov Decision Process

$$\max_{\iota,\theta} \; ELBO\left(\overline{\mathcal{M}}_{\theta}, \phi_{\iota}, \psi_{\theta}\right) = -\min_{\iota,\theta} \left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

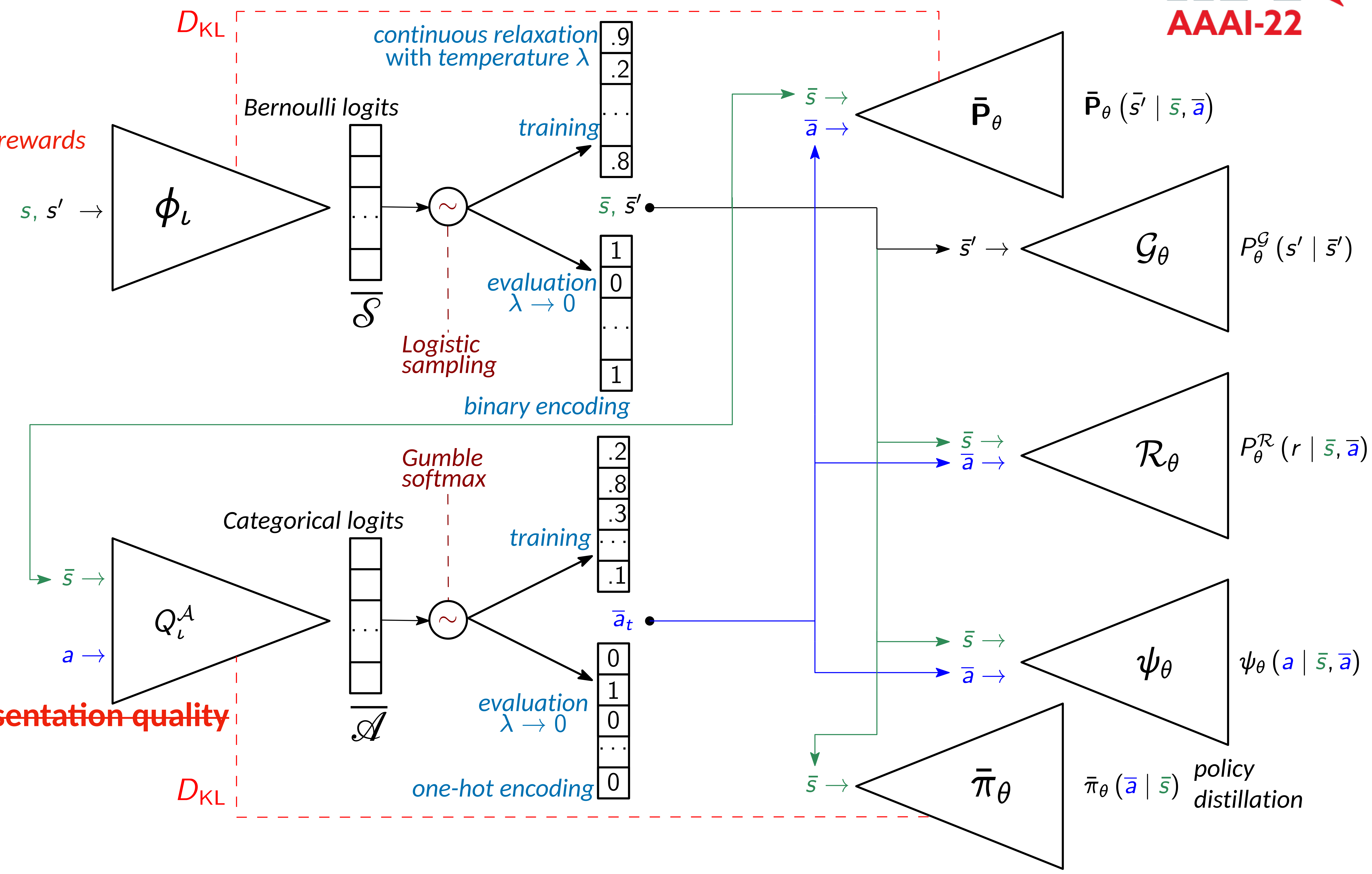
$$\mathbf{D}_{\iota,\theta} = -\mathop{\mathbb{E}}_{\substack{s,a,r,s' \sim \xi_{\pi} \\ \bar{s},\bar{s}' \sim \phi_{\iota}(\cdot \mid s,s') \\ \bar{a} \sim Q_{\iota}^{\mathcal{A}}(\cdot \mid \bar{s},a)}} \left[\log P_{\theta}^{\mathcal{G}}(s' \mid \bar{s}') + \log \psi_{\theta}(a \mid \bar{s},\bar{a}) + \log P_{\theta}^{\mathcal{R}}(r \mid \bar{s},\bar{a})\right]$$

Log-likelihood of rewards

$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s' \sim \xi_{\pi} \\ \bar{s} \sim \phi_{\iota}(\cdot \mid s) \\ \bar{a} \sim Q_{\iota}^{\mathcal{A}}(\cdot \mid \bar{s},a)}} \left[D_{KL}\left(\phi_{\iota}(\cdot \mid s') \,\|\, \overline{\mathbf{P}}_{\theta}(\cdot \mid \bar{s},\bar{a})\right) + D_{\mathrm{KL}}\left(Q_{\iota}^{\mathcal{A}}(\cdot \mid \bar{s},a) \,\|\, \bar{\pi}_{\theta}(\cdot \mid \bar{s})\right)\right]$$

Variational version of local transition loss

$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a} \sim \xi_{\bar{\pi}}} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}(\cdot \mid s,\bar{a}), \overline{\mathbf{P}}(\cdot \mid \phi(s),\bar{a})\right)$$
$$\leq \mathbb{E}_{s,\bar{a},s' \sim \xi_{\bar{\pi}}} W_{d_{\bar{\mathcal{S}}}}\left(\phi(\cdot \mid s'), \overline{\mathbf{P}}(\cdot \mid \phi(s),\bar{a})\right)$$



- Variational proxies to local losses
  ➡ **No learning guarantee: ~~abstraction quality~~, ~~representation quality~~**

Variational Markov Decision Process

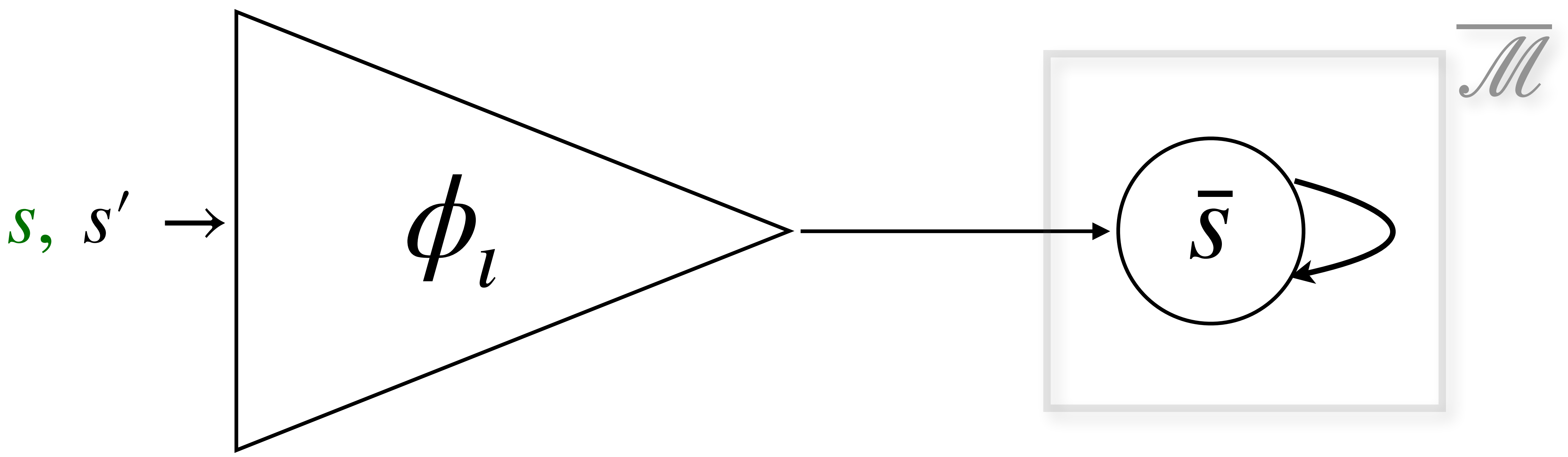
$$\max_{\iota,\theta} \; ELBO\left(\overline{\mathcal{M}}_\theta, \phi_\iota, \psi_\theta\right) = -\min_{\iota,\theta}\left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

$$\mathbf{D}_{\iota,\theta} = -\mathop{\mathbb{E}}_{\substack{s,a,r,s'\sim\xi_\pi \\ \bar{s},\bar{s}'\sim\phi_\iota(\cdot|s,s') \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot|\bar{s},a)}}\left[\log P_\theta^{\mathcal{G}}(s' \mid \bar{s}') + \log \psi_\theta(a \mid \bar{s}, \bar{a}) + \log P_\theta^{\mathcal{R}}(r \mid \bar{s}, \bar{a})\right]$$

$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s'\sim\xi_\pi \\ \bar{s}\sim\phi_\iota(\cdot|s) \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot|\bar{s},a)}}\left[D_{KL}\left(\phi_\iota(\cdot \mid s') \parallel \overline{\mathbf{P}}_\theta(\cdot \mid \bar{s}, \bar{a})\right) + D_{\mathrm{KL}}\left(Q_\iota^{\mathcal{A}}(\cdot \mid \bar{s}, a) \parallel \bar{\pi}_\theta(\cdot \mid \bar{s})\right)\right]$$

Variational version of local transition loss

$$L_{\mathbf{P}}^{\xi_\pi} = \mathbb{E}_{s,\bar{a}\sim\xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}(\cdot \mid s, \bar{a}), \overline{\mathbf{P}}(\cdot \mid \phi(s), \bar{a})\right)$$
$$\leq \mathbb{E}_{s,\bar{a},s'\sim\xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi(\cdot \mid s'), \overline{\mathbf{P}}(\cdot \mid \phi(s), \bar{a})\right)$$



- Variational proxies to local losses
  - ➡ **No learning guarantee: ~~abstraction quality~~, ~~representation quality~~**
  - ➡ **Mode collapse**

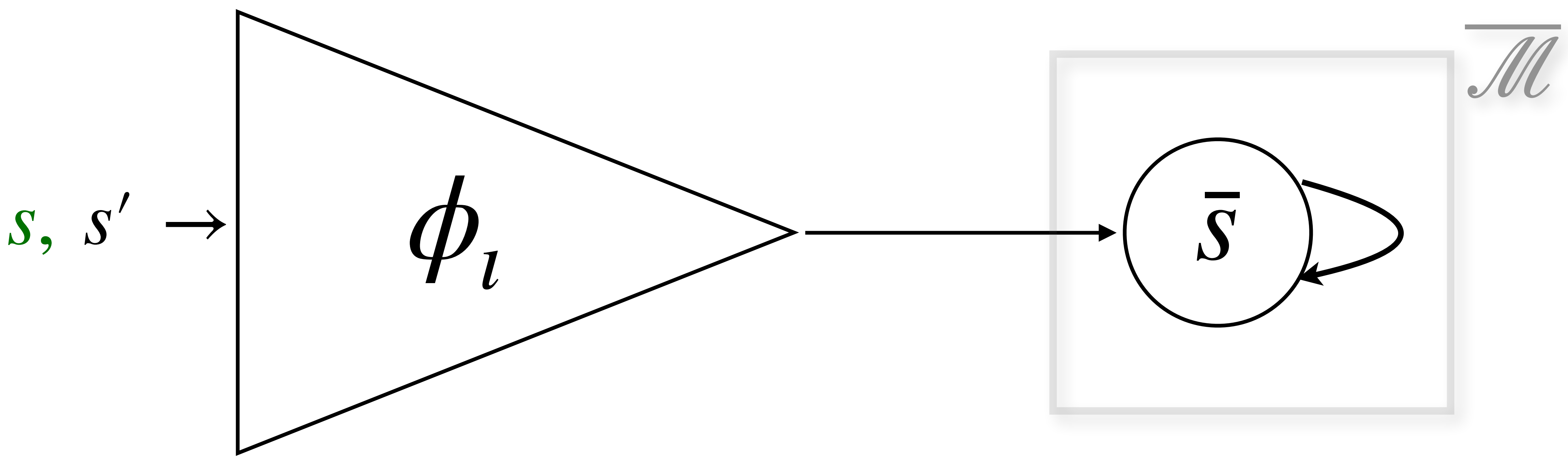
Variational Markov Decision Process

$$\max_{\iota,\theta} \; ELBO\left(\overline{\mathcal{M}}_\theta, \phi_\iota, \psi_\theta\right) = -\min_{\iota,\theta}\left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

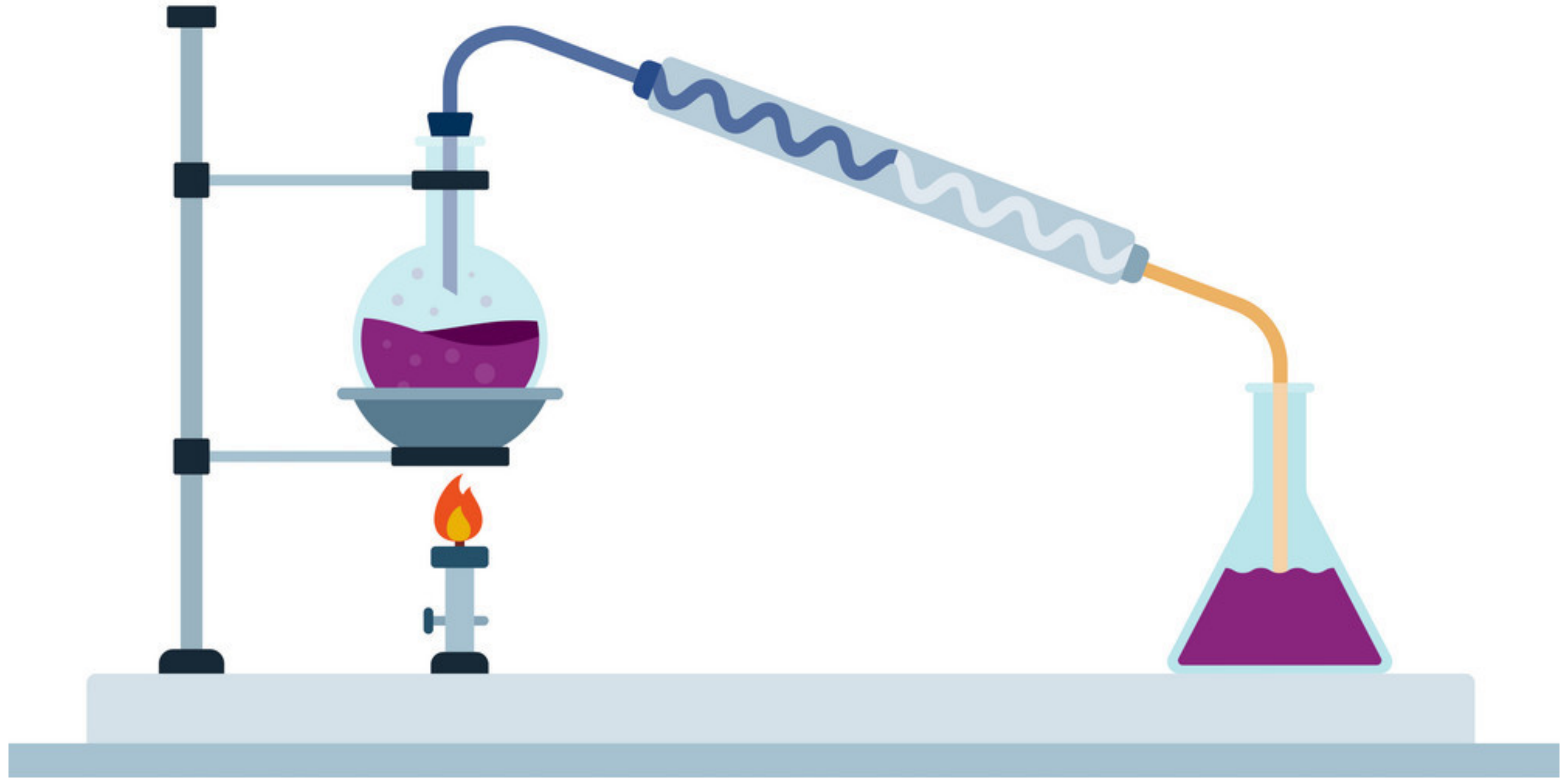
$$\mathbf{D}_{\iota,\theta} = -\mathop{\mathbb{E}}_{\substack{s,a,r,s' \sim \xi_\pi \\ \bar{s},\bar{s}' \sim \phi_\iota(\cdot \mid s,s') \\ \bar{a} \sim Q^{\mathcal{A}}_\iota(\cdot \mid \bar{s},a)}} \left[\log P^{\mathcal{G}}_\theta(s' \mid \bar{s}') + \log \psi_\theta(a \mid \bar{s},\bar{a}) + \log P^{\mathcal{R}}_\theta(r \mid \bar{s},\bar{a})\right]$$

$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s' \sim \xi_\pi \\ \bar{s} \sim \phi_\iota(\cdot \mid s) \\ \bar{a} \sim Q^{\mathcal{A}}_\iota(\cdot \mid \bar{s},a)}} \left[D_{KL}\left(\phi_\iota(\cdot \mid s') \parallel \overline{\mathbf{P}}_\theta(\cdot \mid \bar{s},\bar{a})\right) + D_{\mathrm{KL}}\left(Q^{\mathcal{A}}_\iota(\cdot \mid \bar{s},a) \parallel \bar{\pi}_\theta(\cdot \mid \bar{s})\right)\right]$$

$$L^{\xi_\pi}_{\mathbf{P}} = \mathbb{E}_{s,\bar{a} \sim \xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi \mathbf{P}(\cdot \mid s,\bar{a}), \overline{\mathbf{P}}(\cdot \mid \phi(s),\bar{a})\right)$$
$$\leq \mathbb{E}_{s,\bar{a},s' \sim \xi_\pi} W_{d_{\bar{\mathcal{S}}}}\left(\phi(\cdot \mid s'), \overline{\mathbf{P}}(\cdot \mid \phi(s),\bar{a})\right)$$



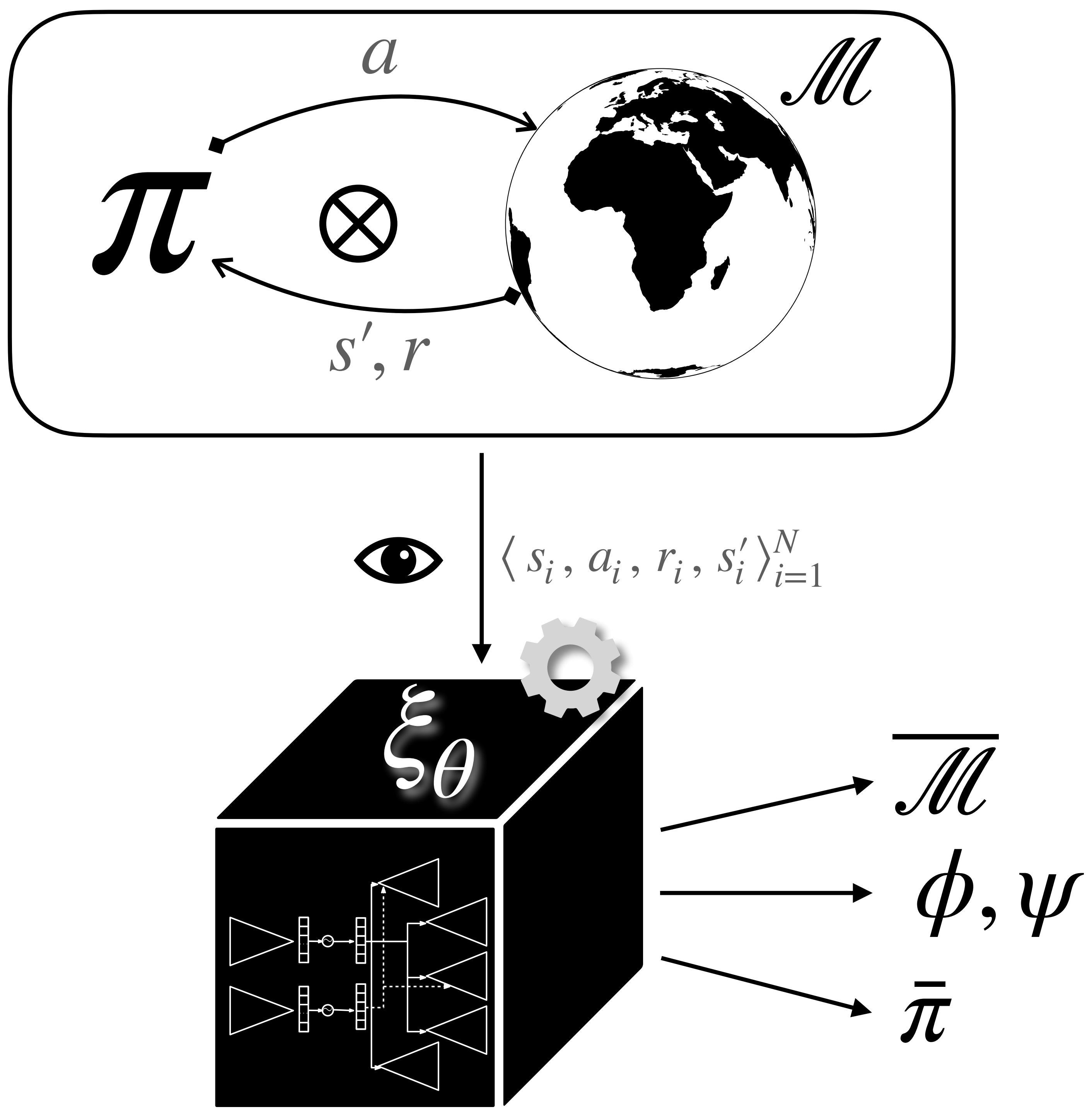
- Variational proxies to local losses
  - ➡ **No learning guarantee: ~~abstraction quality~~, ~~representation quality~~**
  - ➡ **Mode collapse**
  - ➡ **Fix: annealing scheme, extra entropy regularization term, prioritized experience replay, …**



Learning the Latent Space Model

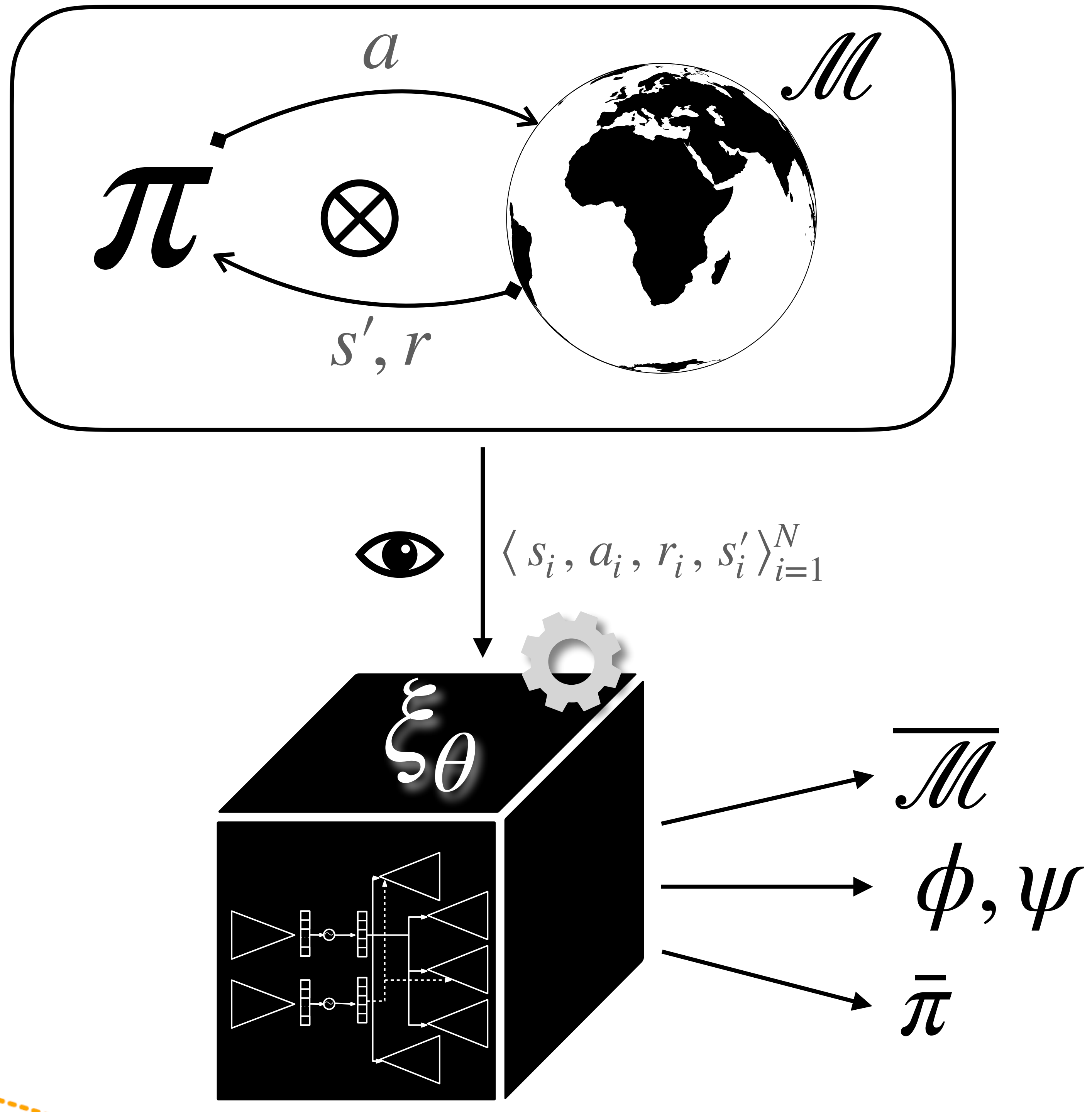
- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ

$$\min_{\theta} D\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

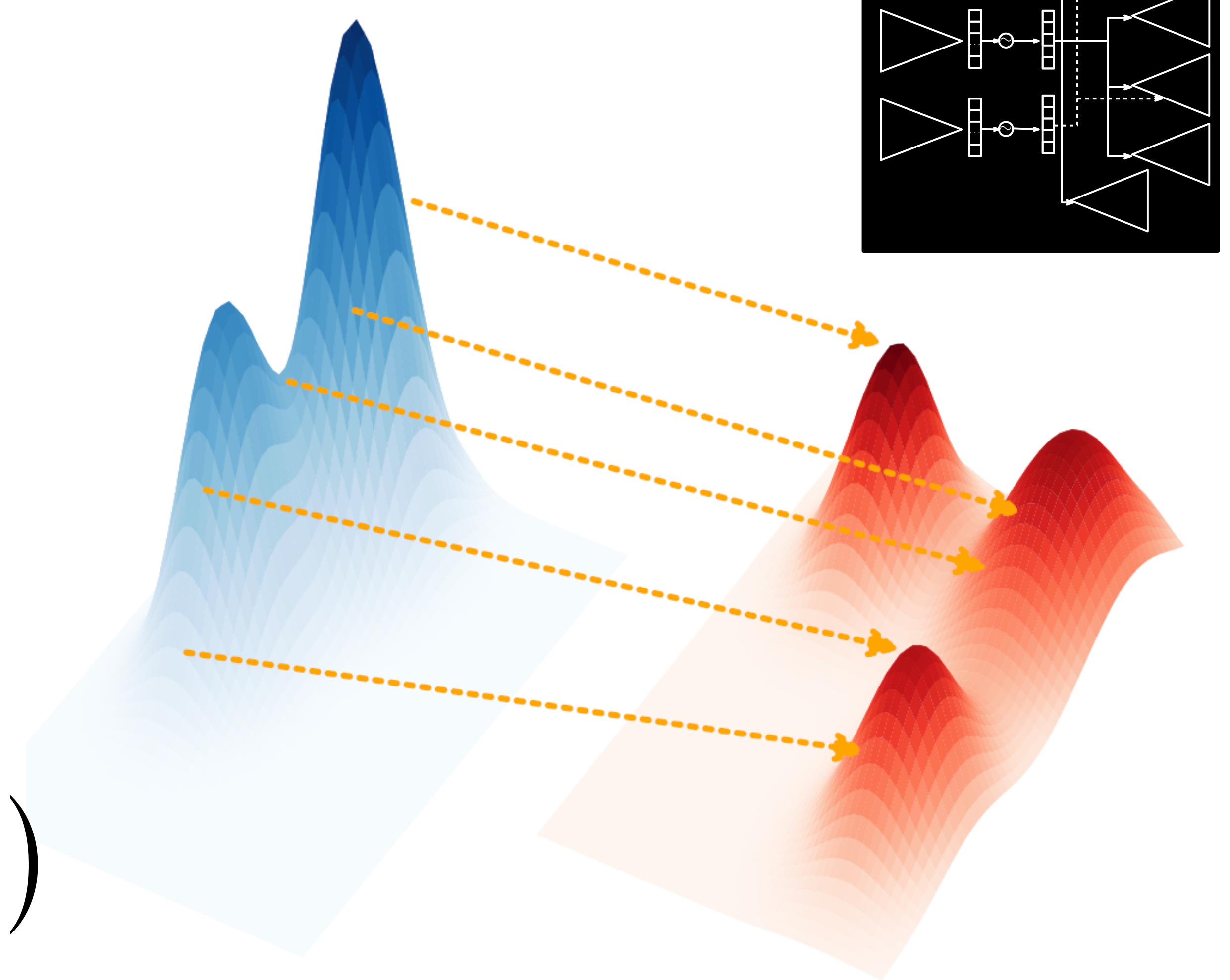


Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ



$$\min_\theta W\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

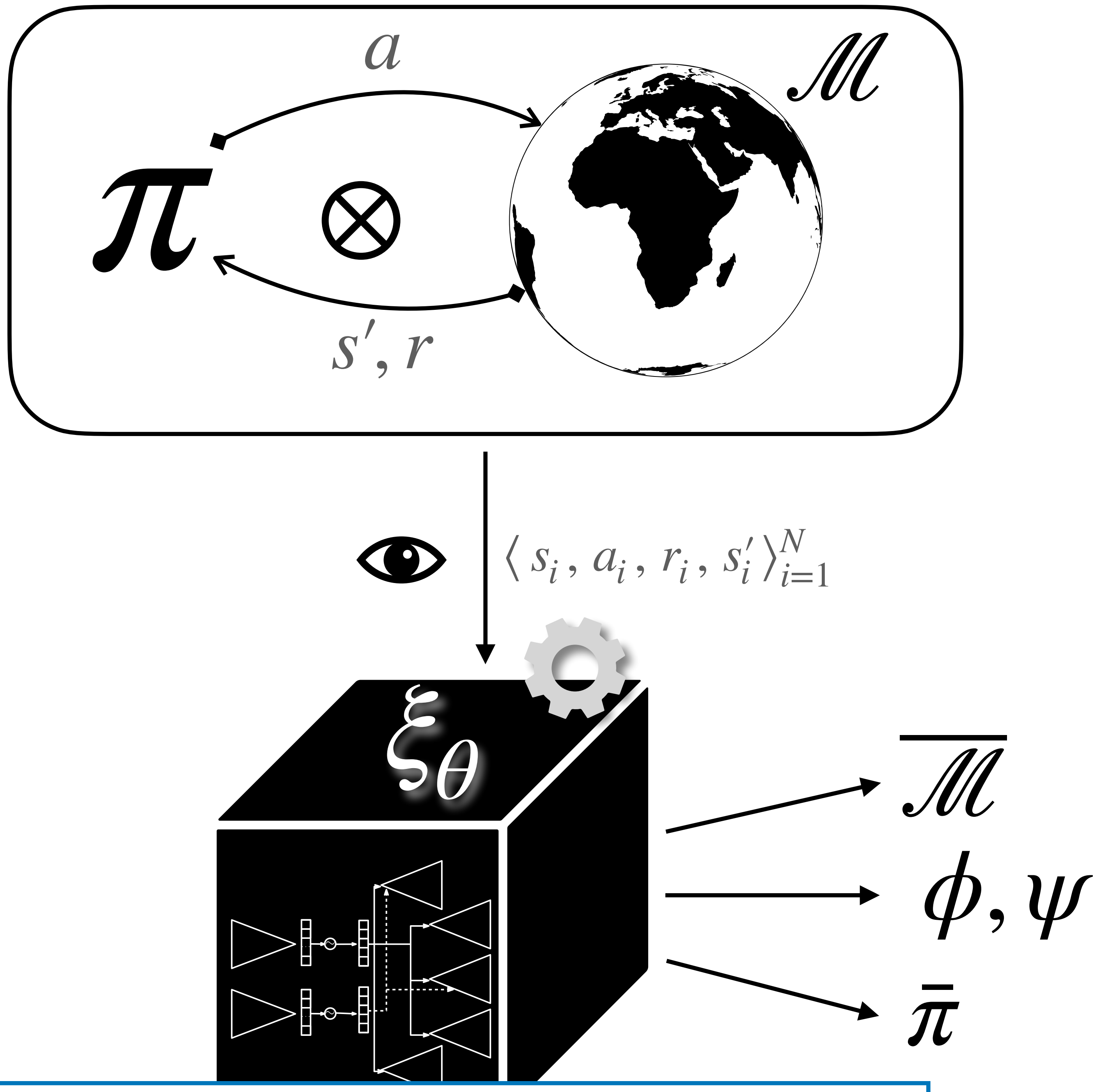


- Choose the *Wasserstein Distance*

$$W\left(P, Q\right) = \inf_{\lambda \in \Lambda(P,Q)} \mathbb{E}_{x,y \sim \lambda}\, d\left(x, y\right) = \sup_{\|f\| \leq 1} \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{y \sim Q} f\left(y\right)$$

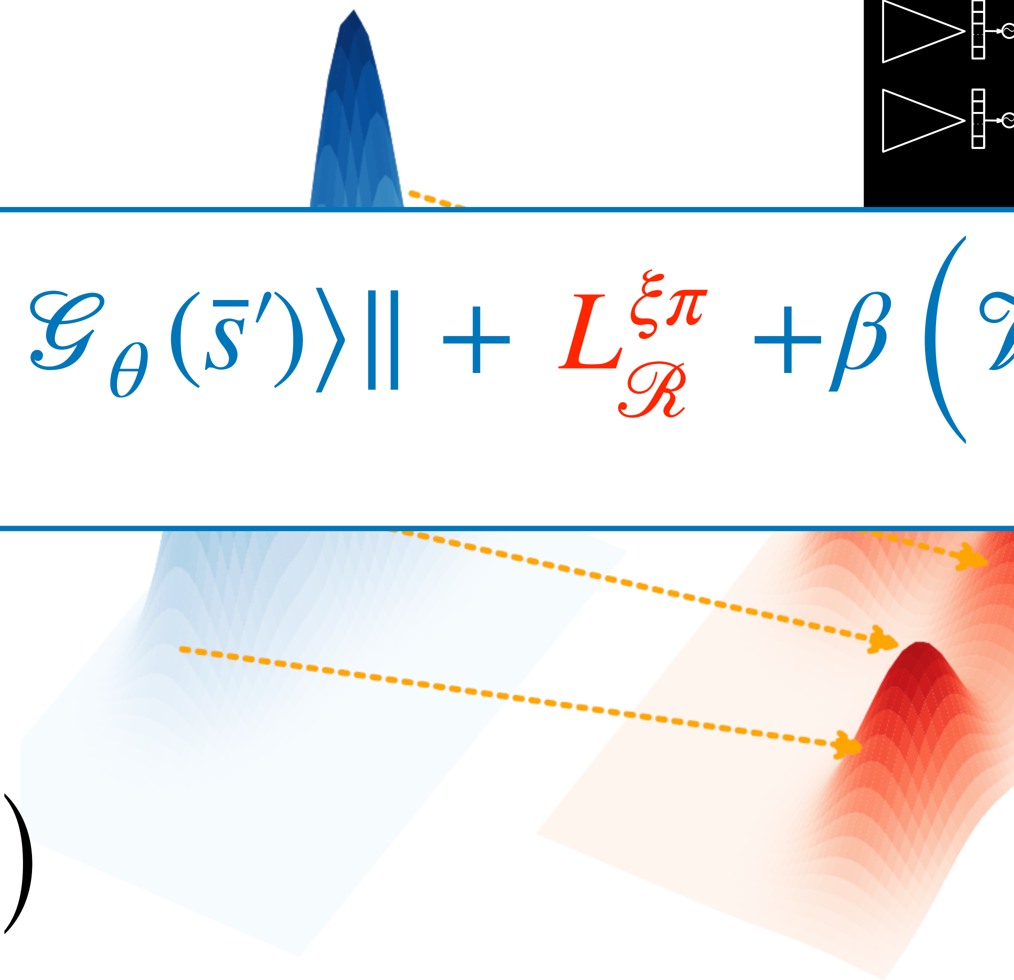
Learning the Latent Space Model

- Train a *behavioral model* ξ_θ by learning from traces produced by executing the RL policy π in the original model $\mathcal{M}$
- Goal: learn ξ_θ so that we can retrieve:
  - The latent MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$
  - The embedding functions ϕ, ψ
  - A latent policy $\bar{\pi}$ distilled from π
- Minimize a *discrepancy* D between $\mathcal{M} \otimes \pi$ and ξ_θ



$$\min_{\theta} W\left(\mathcal{M} \otimes \pi, \xi_\theta\right)$$

$$\leq \min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\,\cdot\,\mid s,a,s')} \left\|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar{s}), \psi_\theta(\bar{s},\bar{a}), \mathcal{G}_\theta(\bar{s}')\rangle\right\| + L_{\mathcal{R}}^{\xi\pi} + \beta\left(\mathcal{W}_{\xi_\pi} + L_{\mathbf{P}}^{\xi_\pi}\right)$$



- Choose the *Wasserstein Distance*

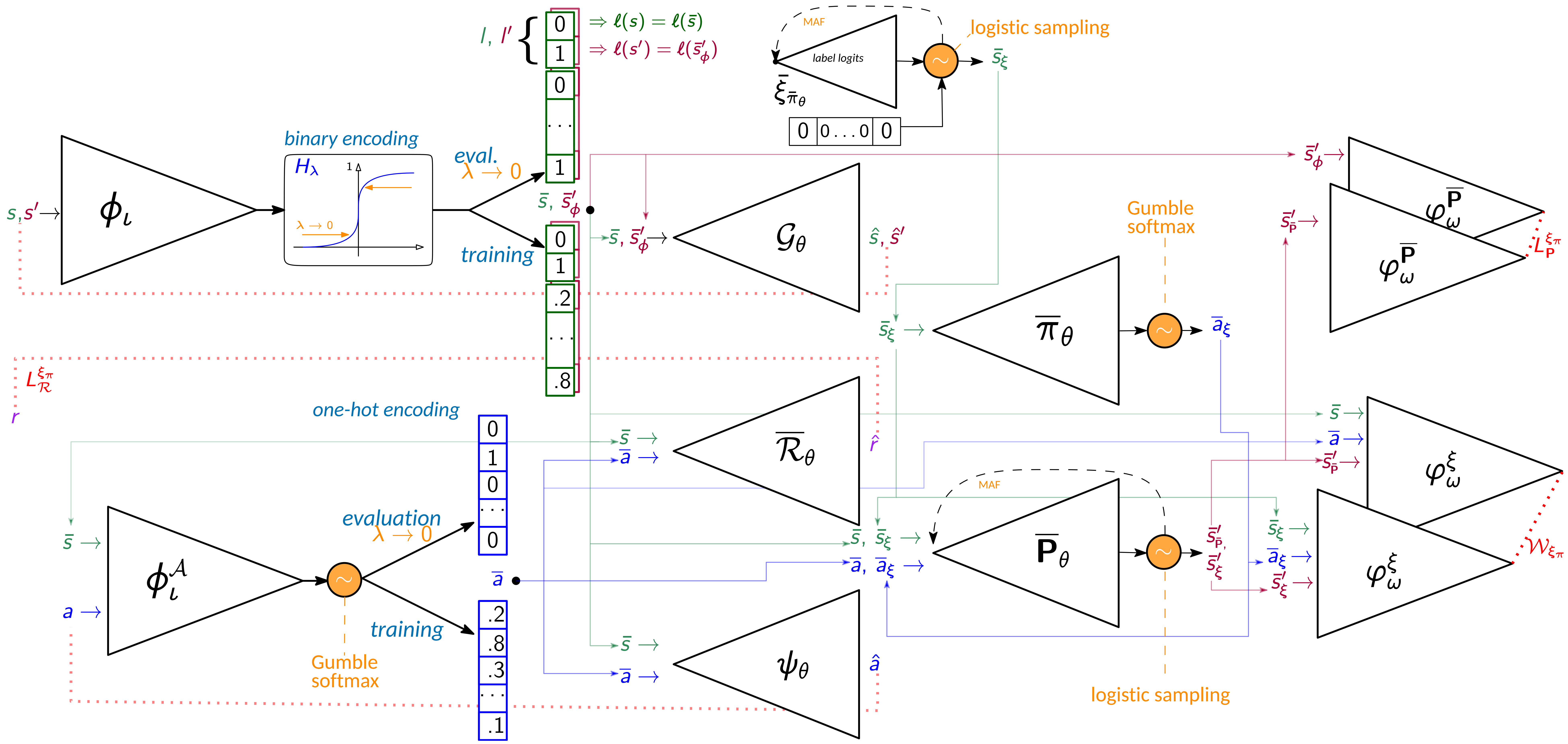
$$W\left(P,Q\right) = \inf_{\lambda\in\Lambda(P,Q)} \mathbb{E}_{x,y\sim\lambda}\, d\left(x,y\right) = \sup_{\|f\|\leq 1} \mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{y\sim Q} f\left(y\right)$$

$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\,\cdot\mid s,a,s')} \left\|\langle s, a, s'\rangle - \langle \mathscr{G}_\theta(\bar{s}), \psi_\theta(\bar{s},\bar{a}), \mathscr{G}_\theta(\bar{s}')\rangle\right\| + L_{\mathscr{R}}^{\xi\pi} + \beta\left(\mathscr{W}_{\xi_\pi} + L_{\mathbf{P}}^{\xi_\pi}\right)$$

$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\cdot\mid s,a,s')} \|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar{s}), \psi_\theta(\bar{s},\bar{a}), \mathcal{G}_\theta(\bar{s}')\rangle\| + L_{\mathscr{R}}^{\xi\pi} + \beta\left(\mathscr{W}_{\xi_\pi} + L_{\mathbf{P}}^{\xi_\pi}\right)$$

Wasserstein Auto-encoded Markov Decision Process

$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\,\cdot\mid s,a,s')} \left\|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar{s}), \psi_\theta(\bar{s},\bar{a}), \mathcal{G}_\theta(\bar{s}')\rangle\right\| + L^{\xi_\pi}_{\mathcal{R}} + \beta\left(\mathcal{W}_{\xi_\pi} + L^{\xi_\pi}_{\mathbf{P}}\right)$$



- $\mathcal{W}_{\xi_\pi} = \max_{\omega:\,\|\varphi^\xi_\omega\|\leq 1} \mathbb{E}_{s,a\sim\xi_\pi}\mathbb{E}_{\bar{a}\sim\phi^{\mathcal{A}}_\iota(\,\cdot\mid\phi_\iota(s),a)}\mathbb{E}_{\bar{s}'\sim\overline{\mathbf{P}}_\theta(\,\cdot\mid\bar{s},\bar{a})}\,\varphi^\xi_\omega\left(\phi_\iota(s),\bar{a},\bar{s}'\right) - \mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\xi_{\bar{\pi}}}\,\varphi^\xi_\omega(\bar{s},\bar{a},\bar{s}')$
- $L^{\xi_\pi}_{\mathbf{P}} = \max_{\omega:\,\|\varphi^{\mathbf{P}}_\omega\|\leq 1} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\,\cdot\mid s,a,s')}\left[\varphi^{\mathbf{P}}_\omega(s,a,\bar{s},\bar{a},\bar{s}') - \mathbb{E}_{\bar{s}'_{\mathbf{P}}\sim\overline{\mathbf{P}}_\theta(\,\cdot\mid\bar{s},\bar{a})}\,\varphi^{\mathbf{P}}_\omega\left(s,a,\bar{s},\bar{a},\bar{s}'_{\mathbf{P}}\right)\right]$

Evaluation

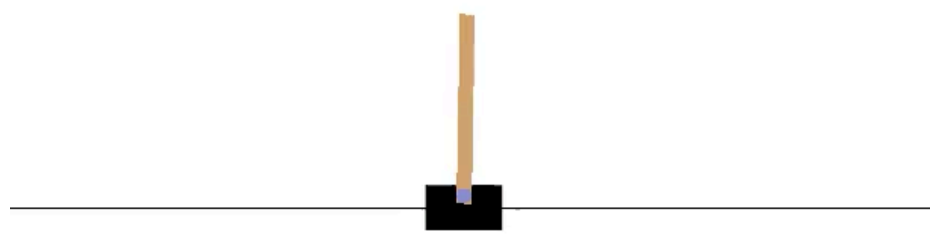
Distillation: performance of $\bar{\pi}$

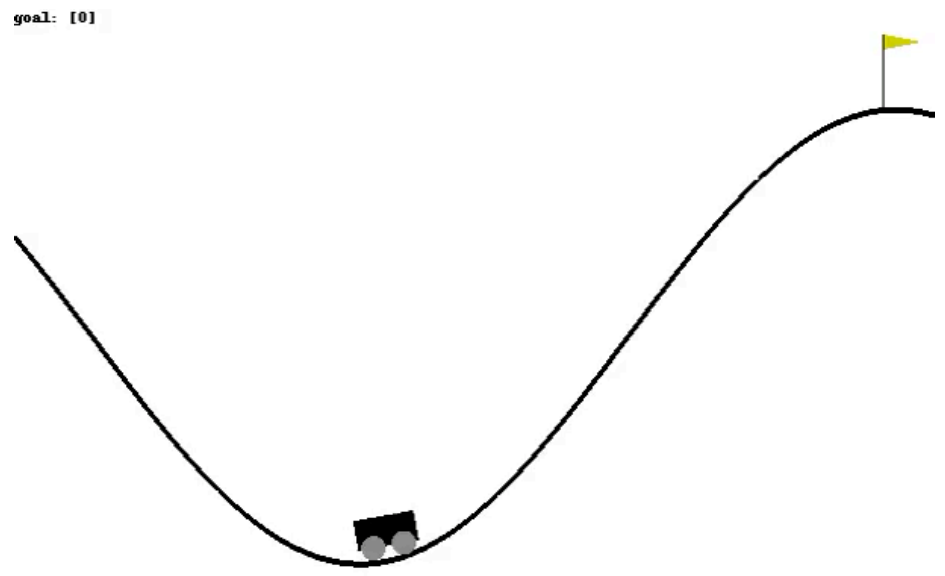
WAE-MDPs distill policies up to 10 times faster than VAE-MDPs

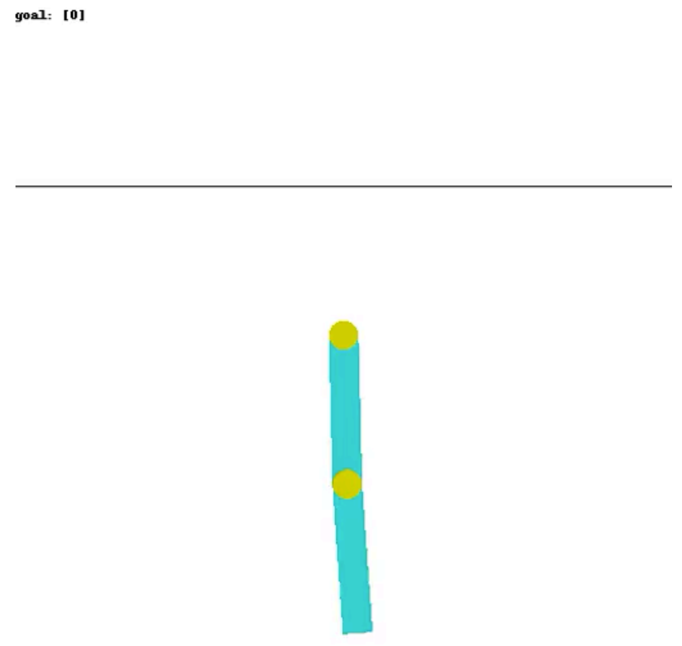
- *Faster*
- *Better performance*
- *Learning guarantees*
- *Similar or even better model quality*

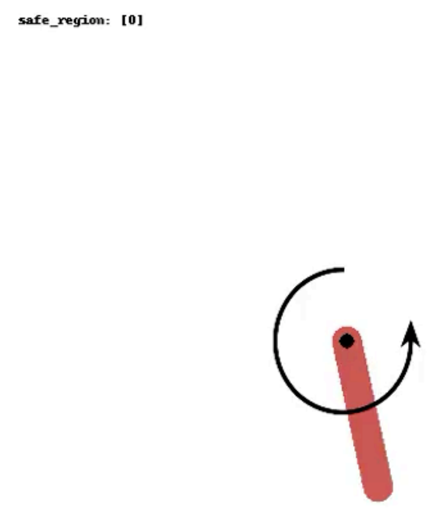
Evaluation

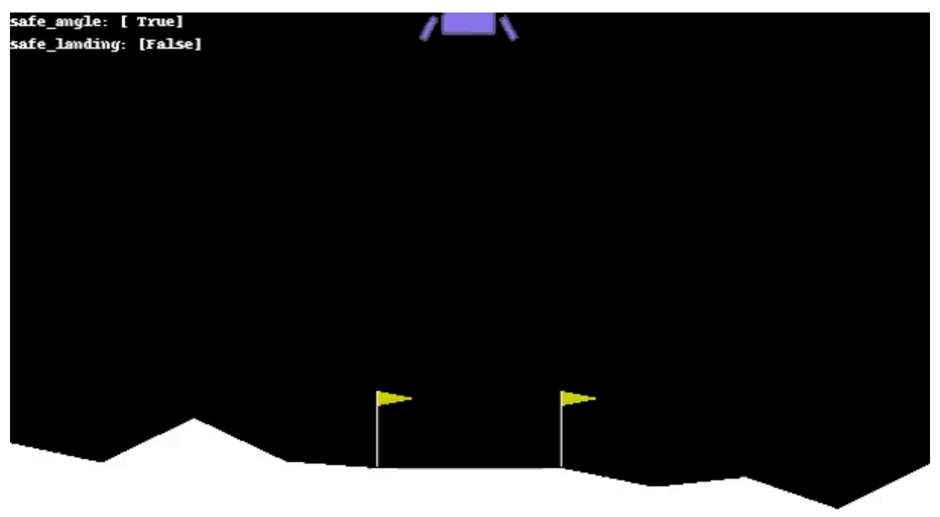
CartPole	*MountainCar*	*Acrobot*	*Pendulum*	*LunarLander*







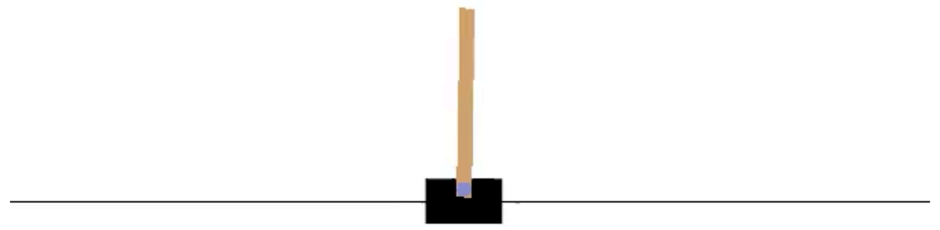
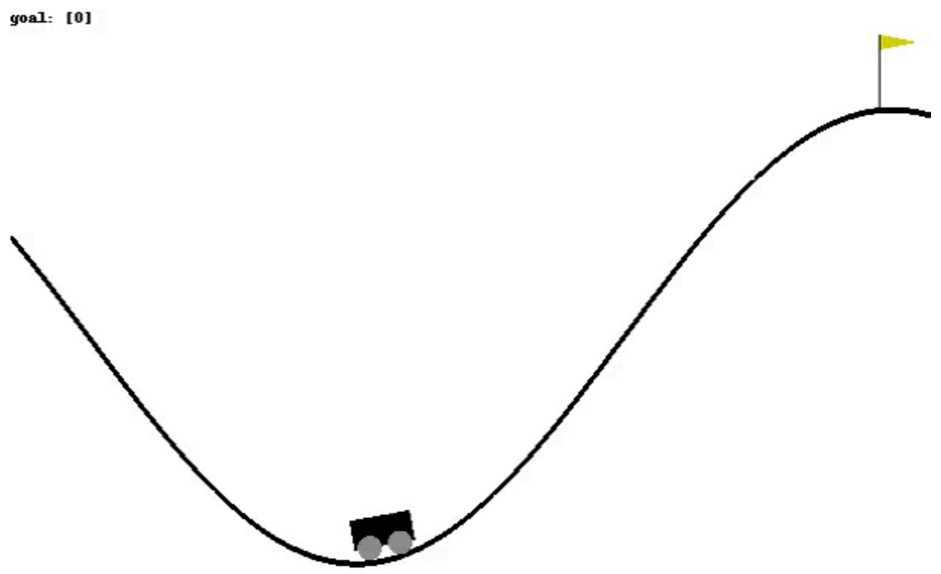
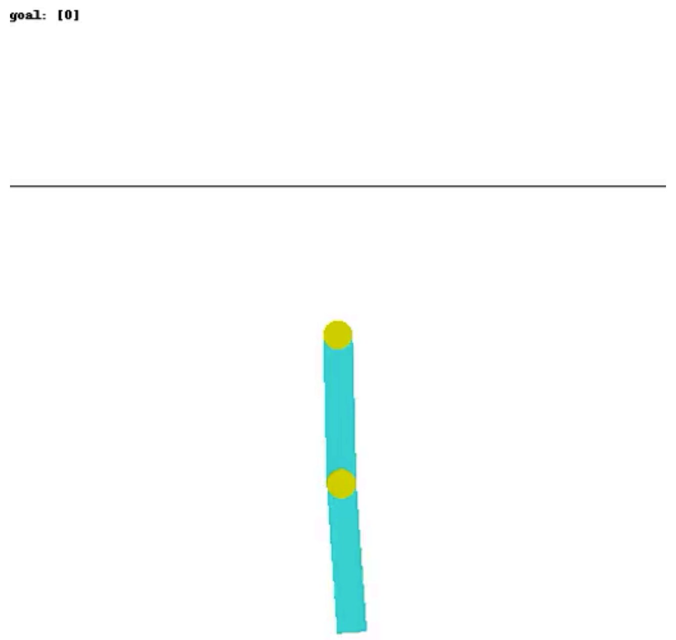
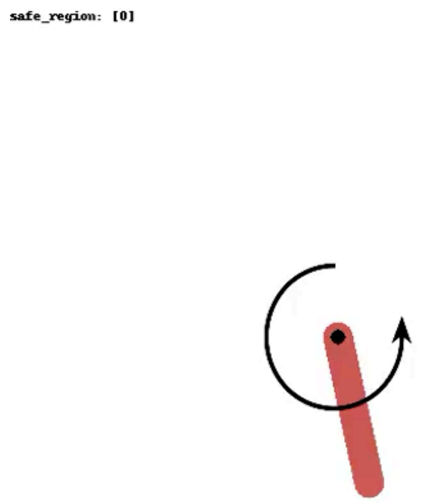
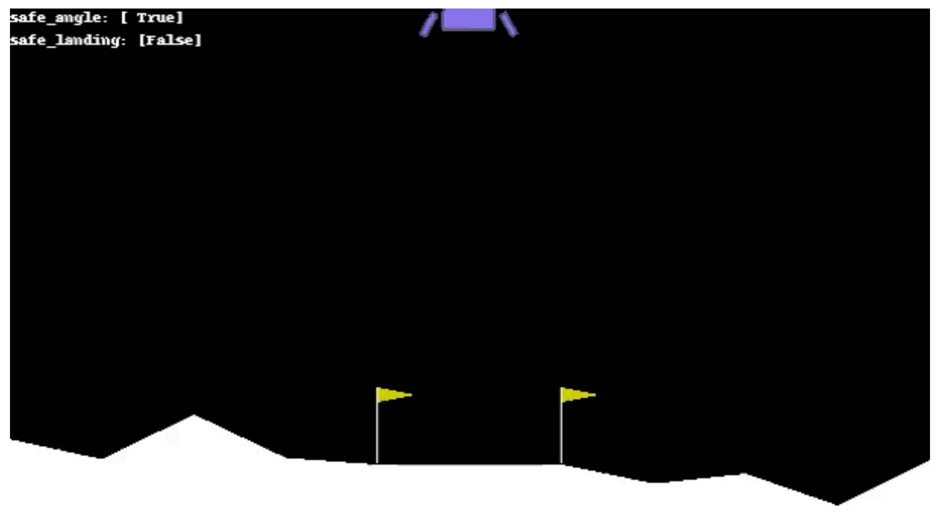




Time-to-failure properties (lower is better)

CartPole	MountainCar	Acrobot	Pendulum	LunarLander
$\varphi = \neg\text{Reset}\,\mathcal{U}\,\neg\text{Safe}$	$\varphi = \neg\text{Goal}\,\mathcal{U}\,\text{Reset}$	$\varphi = \neg\text{Goal}\,\mathcal{U}\,\text{Reset}$	$\varphi = \Diamond(\neg\text{Safe} \wedge \bigcirc\text{Reset})$	$\varphi = \neg\text{SafeLanding}\,\mathcal{U}\,\text{Reset}$
$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.032$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.0022$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.037$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.0702$

Evaluation

CartPole	*MountainCar*	*Acrobot*	*Pendulum*	*LunarLander*
				

Time-to-failure properties (lower is better)

CartPole	MountainCar	Acrobot	Pendulum	LunarLander
$\varphi = \neg \text{Reset}\, \mathcal{U}\, \neg \text{Safe}$	$\varphi = \neg \text{Goal}\, \mathcal{U}\, \text{Reset}$	$\varphi = \neg \text{Goal}\, \mathcal{U}\, \text{Reset}$	$\varphi = \Diamond(\neg \text{Safe} \wedge \bigcirc \text{Reset})$	$\varphi = \neg \text{SafeLanding}\, \mathcal{U}\, \text{Reset}$
$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.032$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.0022$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.037$	$\overline{V}^{\varphi}_{\bar{\pi}_\theta}(\bar{s}_I) = 0.0702$