Research Project Proposal:

A Non-Cooperative approach in Configurable Markov Decision Processes

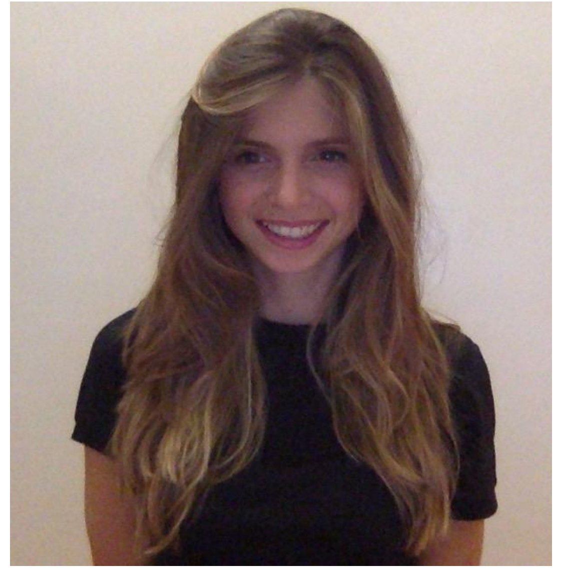
Alessandro Concetti
alessandro.concetti@mail.polimi.it
CSE Track

POLITECNICO
MILANO 1863

HONOURS PROGRAMME HP-SR
in Information Technology

A Non-Cooperative approach in Configurable Markov Decision Processes






Prof. Marcello Restelli Alberto Metelli Giorgia Ramponi Alessandro Concetti

Outline

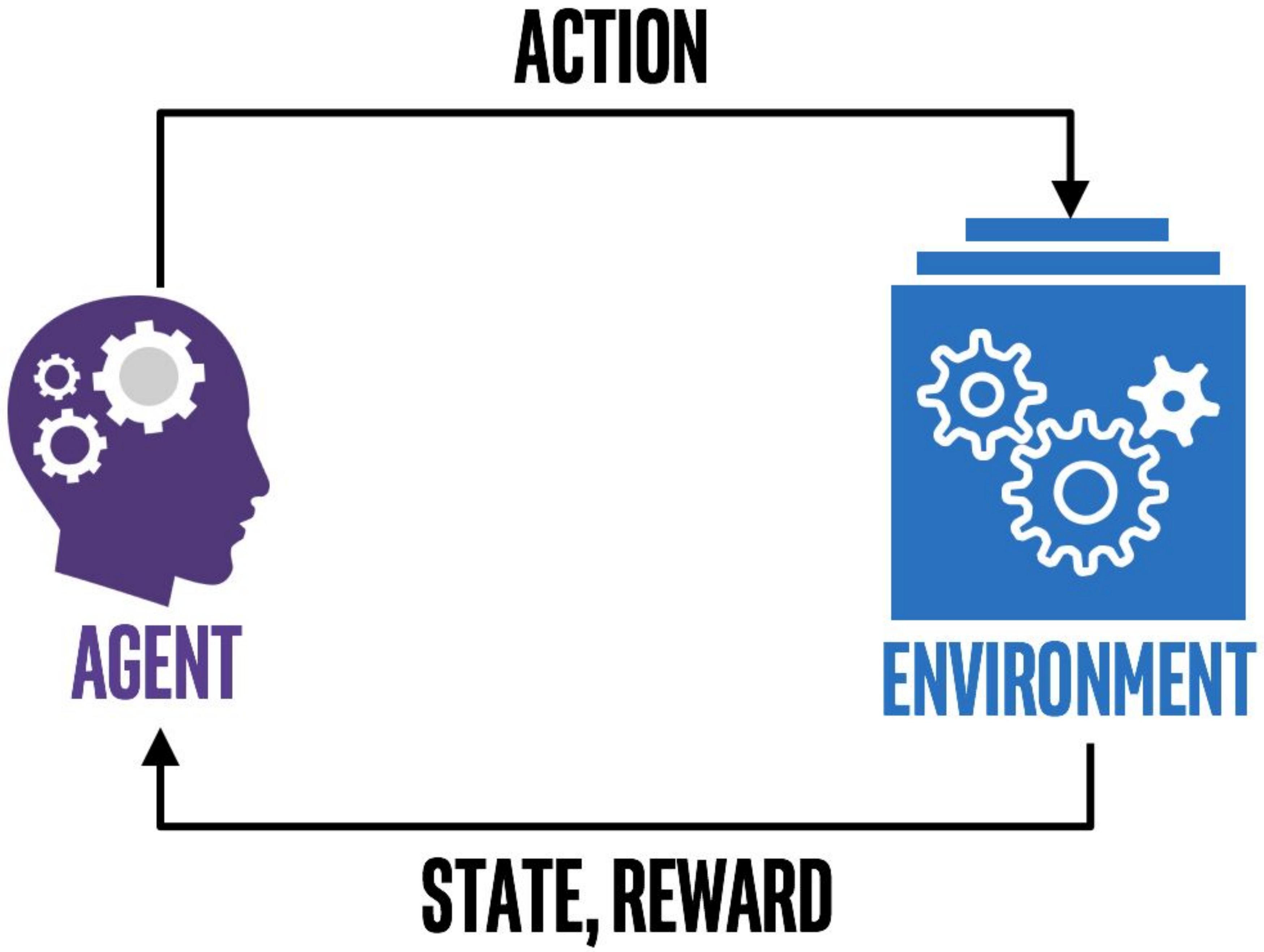
- Preliminaries
- Motivation
- State of the art
- Research plan

30-35 minutes

Preliminaries

- Reinforcement Learning
- Markov Decision Processes (MDPs)
- Configurable Markov Decision Processes (Conf-MDPs)

Reinforcement learning (RL)



Markov Decision Processes (MDPs)

Formally an MDP is a tuple $(S, A, P, R, \gamma, \mu)$, where:

- S is the set of states
- A is the set of actions
- $P(s' | s,a)$ is the transition model, i.e. the probability distribution over the next state, starting from state s and performing action a
- $R(s,a)$ is the immediate reward, given the current state s and the performed action a
- γ is the discount factor
- $\mu(s)$ is the probability distribution over the initial state

Let's define a **policy** as a probability distribution $\pi(a|s)$ over A given the current state s.

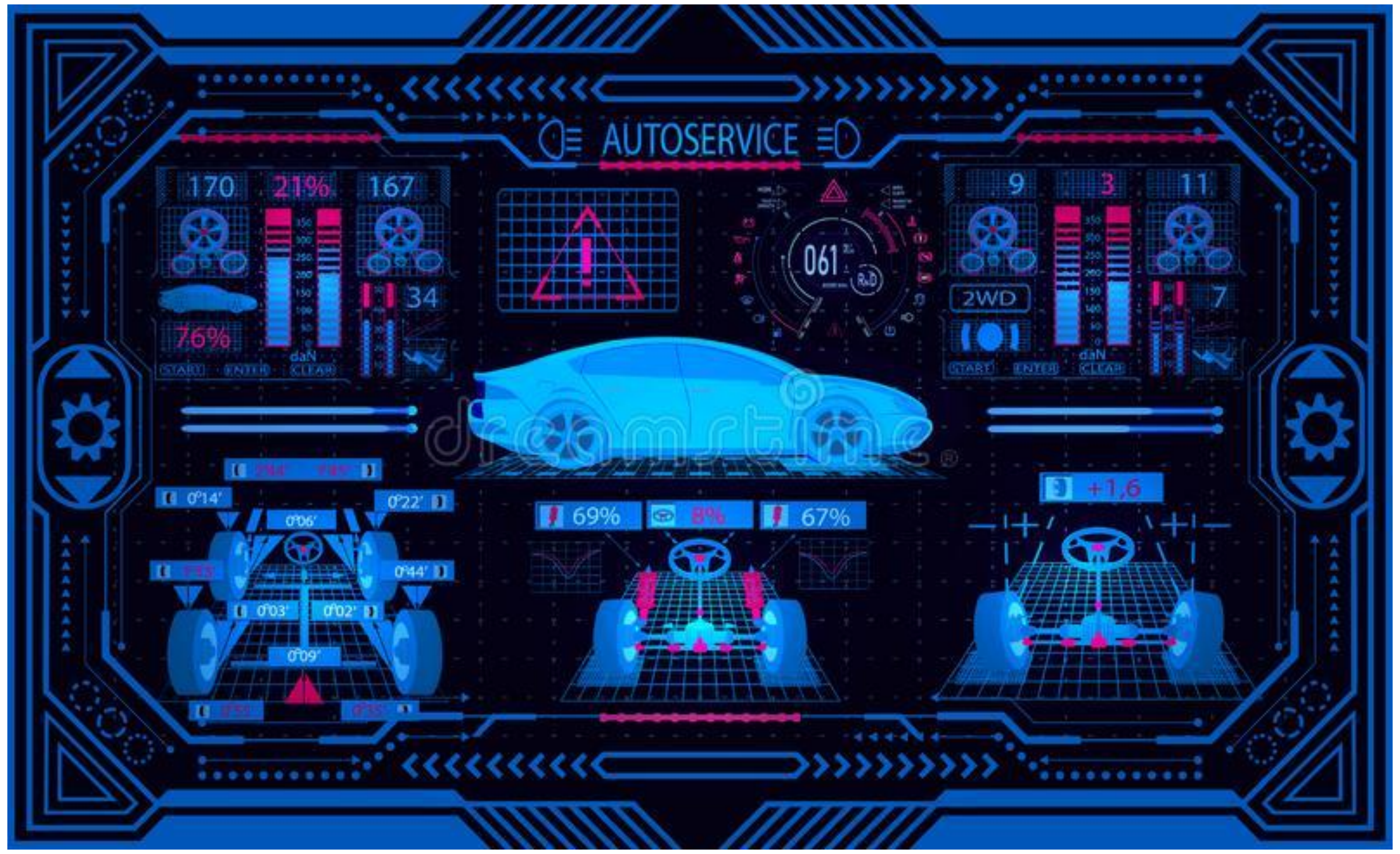
Goal

*The goal is to **find the optimal policy**, i.e the policy that maximizes the expected future reward.*

$$J^{\pi} = E[\sum_{t=0}^{\infty} \gamma^t R(s_t) | \pi]$$

Environmental parameters

In many real-world problems, there is the possibility to configure some environmental parameters.



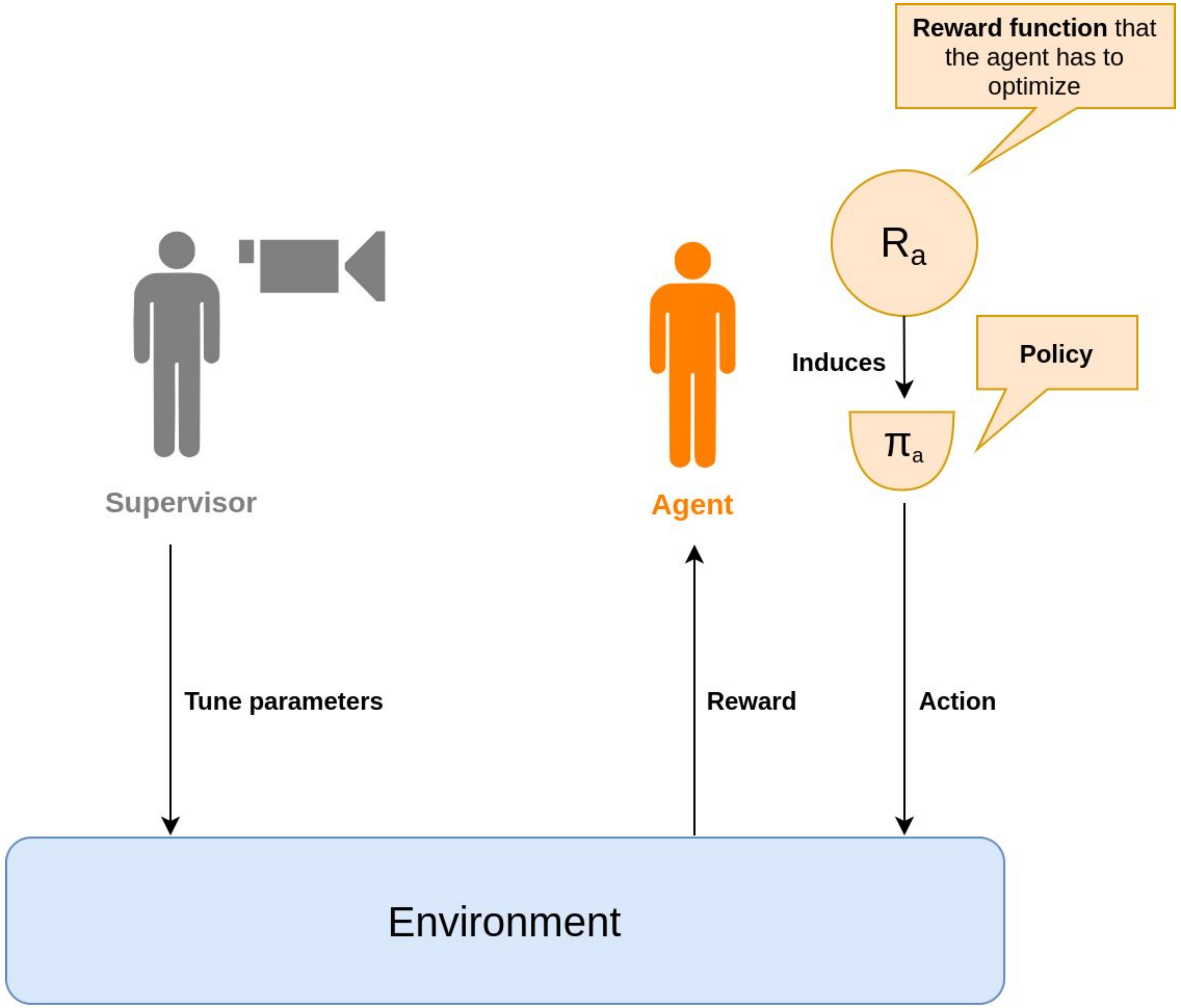
Configurable Markov Decision Processes (Conf-MDPs)

Formally a Conf-MDP is a tuple $(S, A, R, \gamma, \mu, \mathcal{P}, \Pi)$, where:

- (S, A, R, γ, μ) is the classical MDP without the transition model P
- $\mathcal{P}$ is the set of transition models
- Π the set of policies

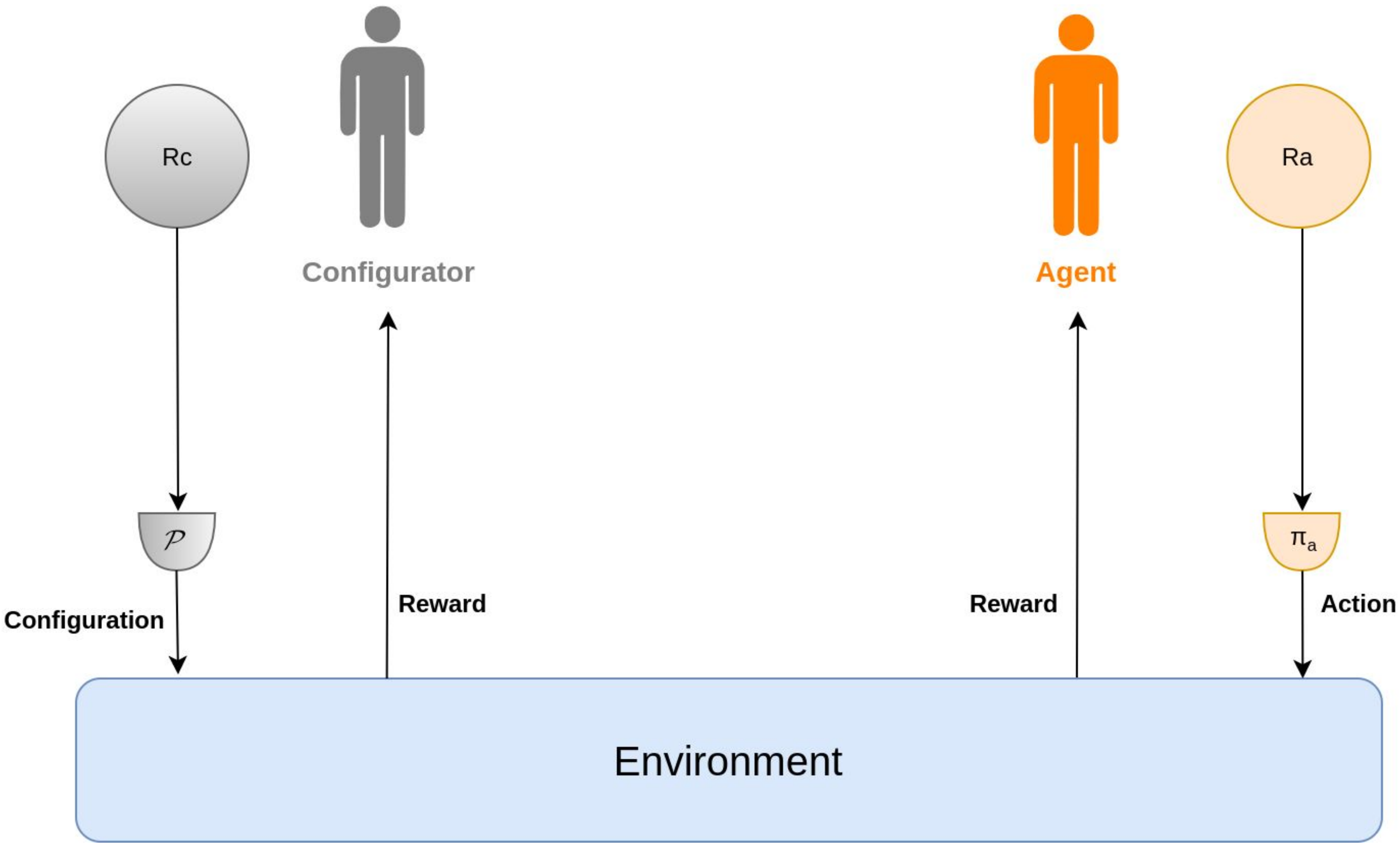
The goal is to find the optimal model-policy pair $(P, \pi) \in \mathcal{P} \times \Pi$.

Configurable Markov Decision Processes (Conf-MDPs)

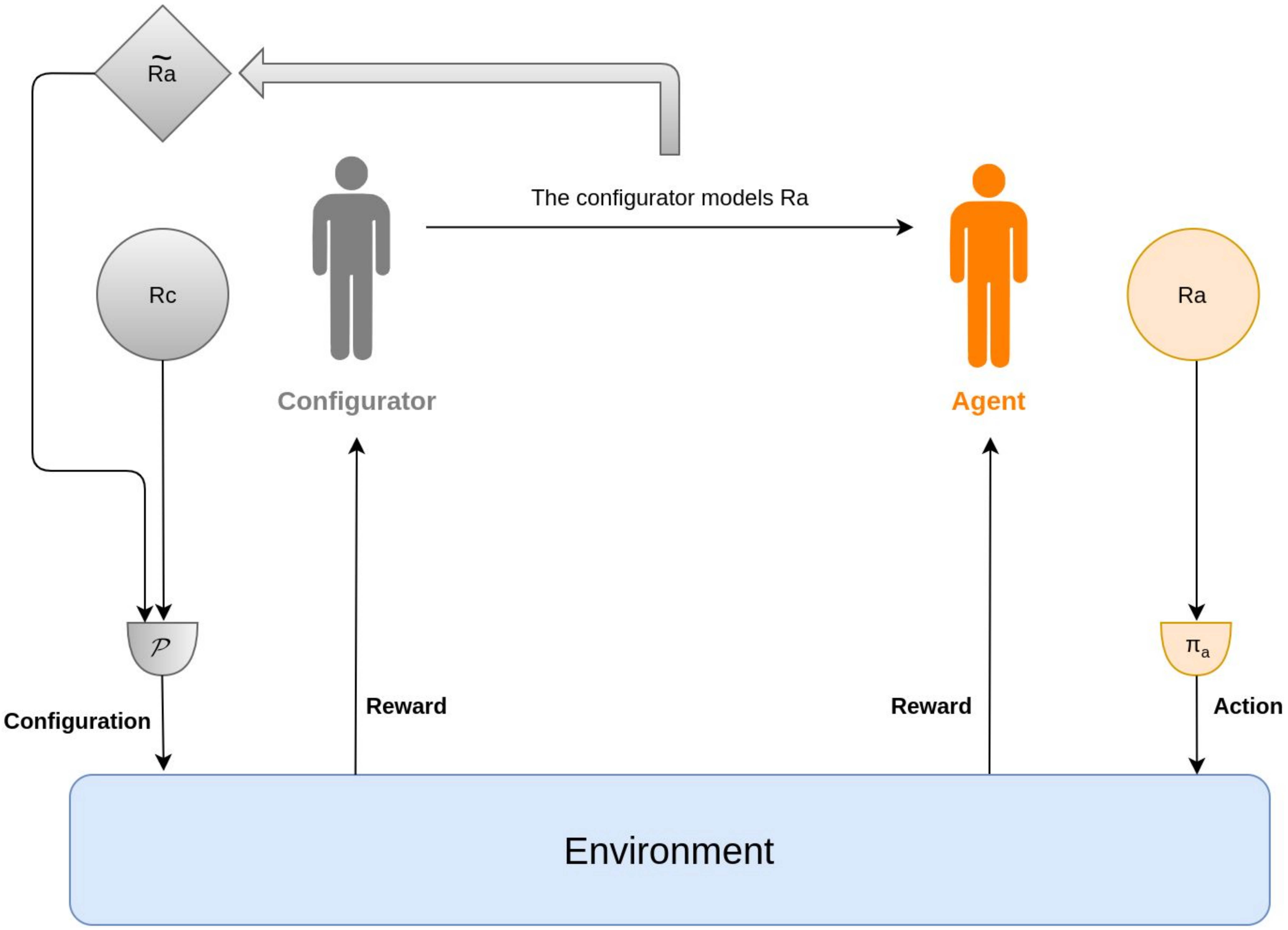


What if the supervisor and the agent were no longer cooperative?

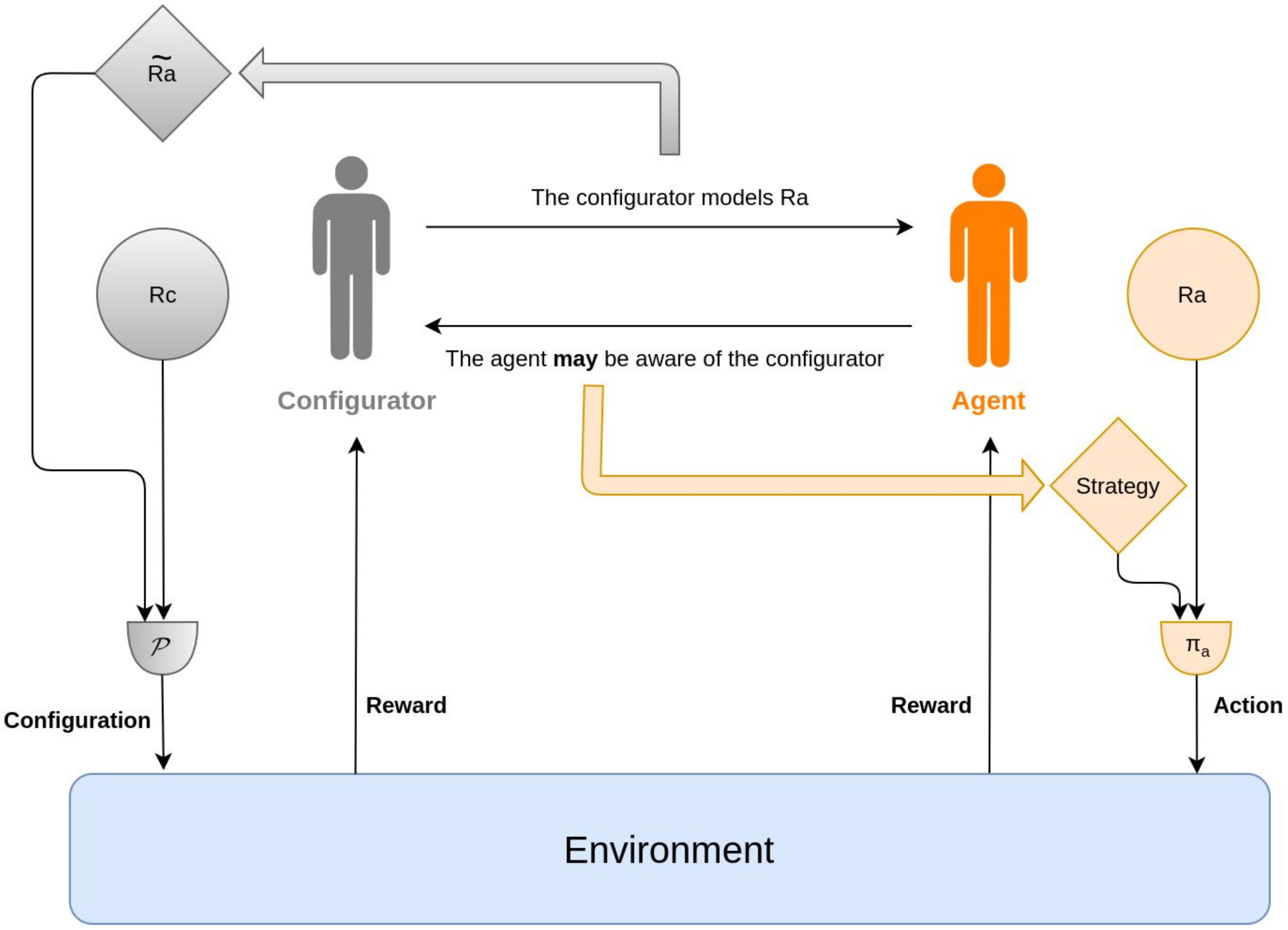
Non-cooperative scenario



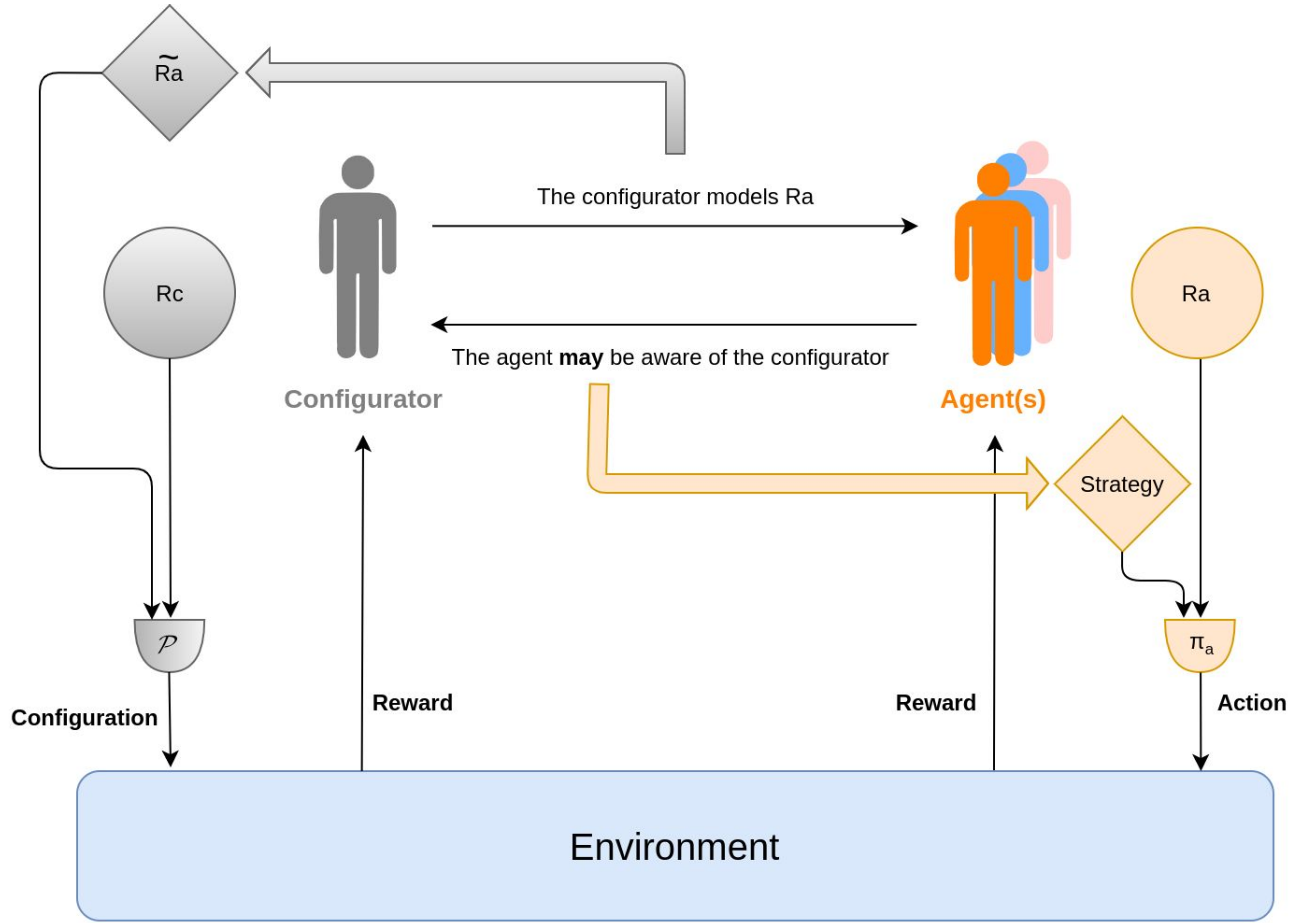
Non-cooperative scenario



Non-cooperative scenario



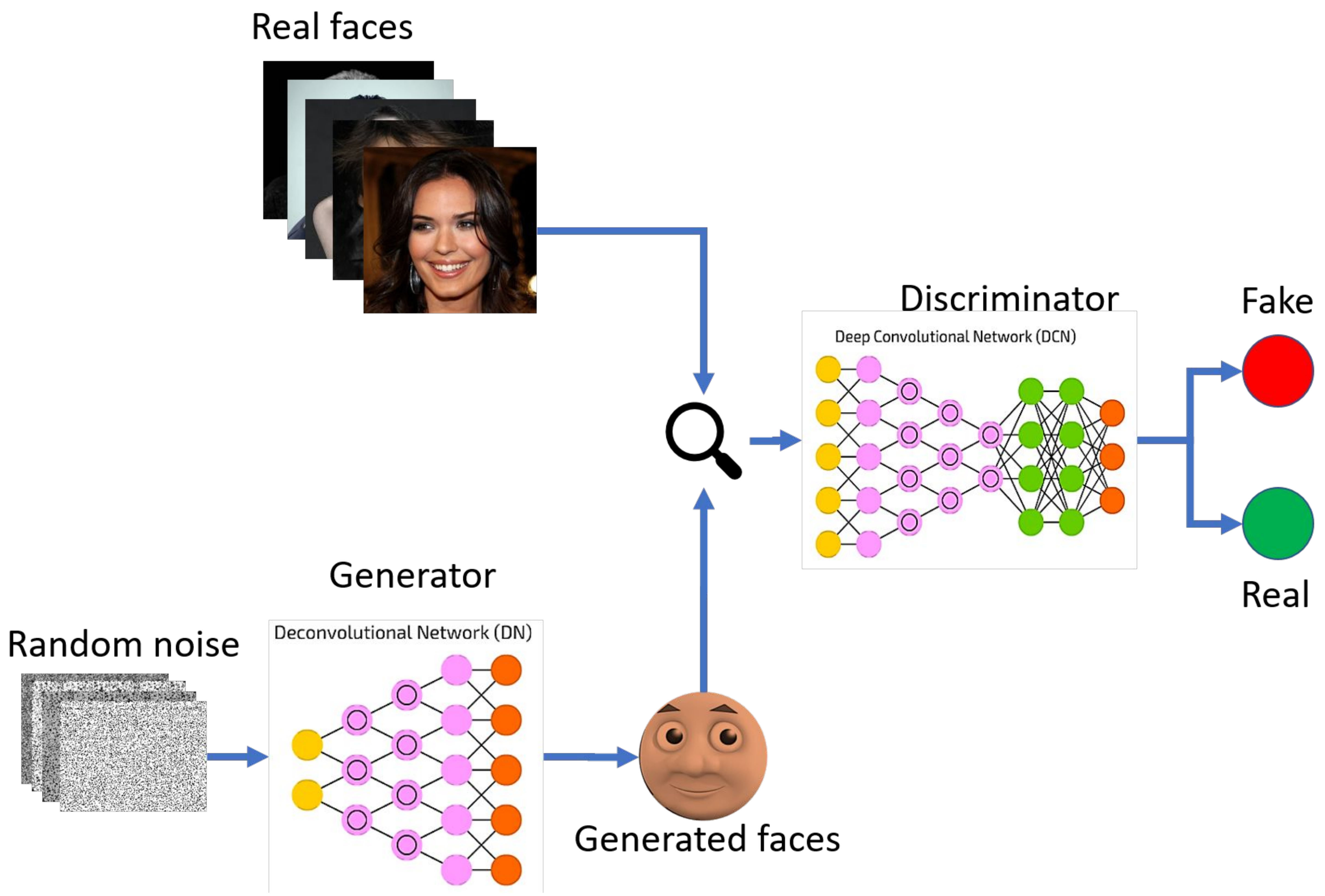
Non-cooperative scenario



Outline

- Preliminaries
- **Motivation**
- State of the art
- Research plan

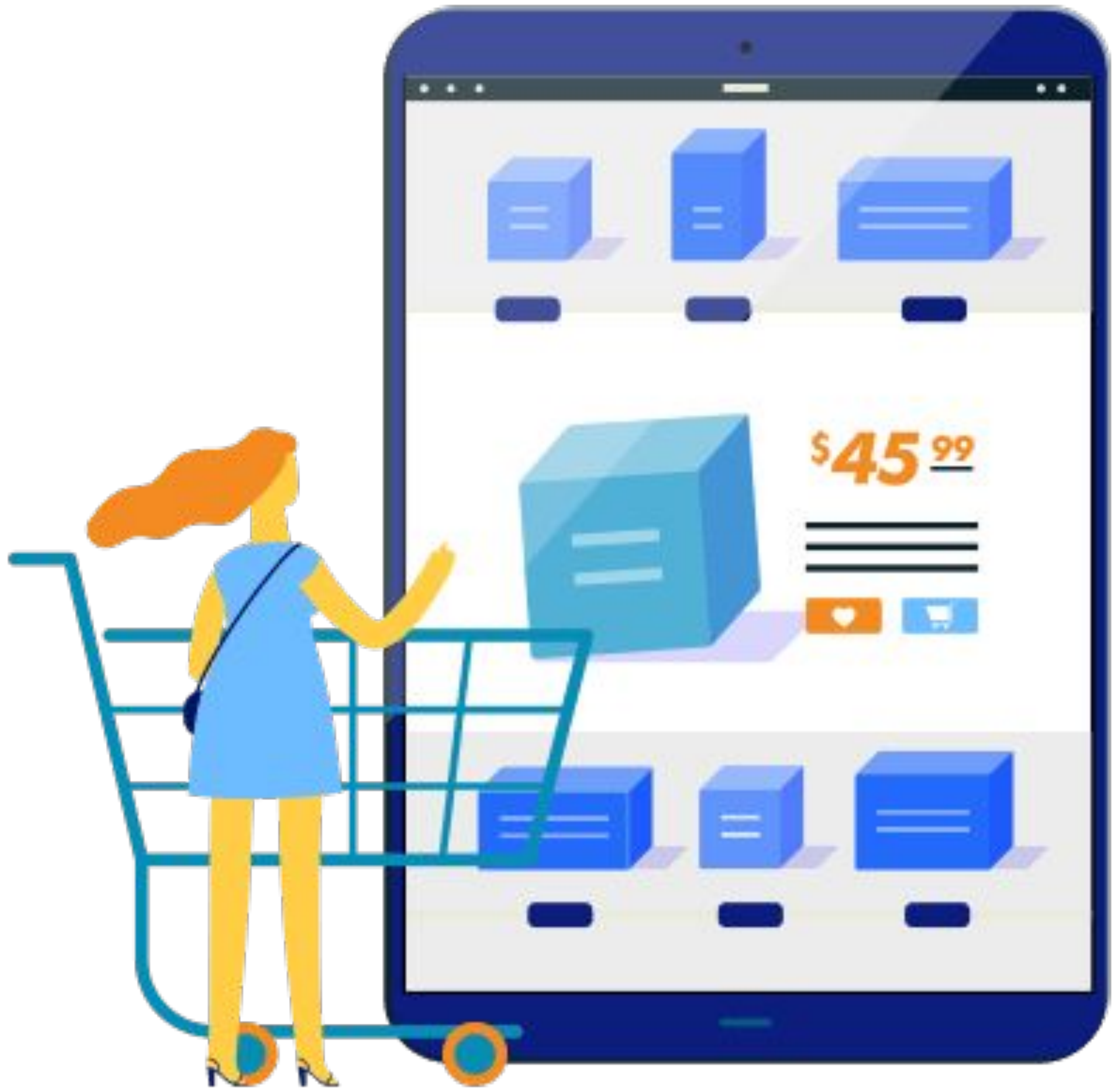
Successes of non-cooperative models in Machine Learning



Real-world applications of Non-Cooperative Conf-MDPs

Supermarket

Real-world applications of Non-Cooperative Conf-MDPs



E-commerce

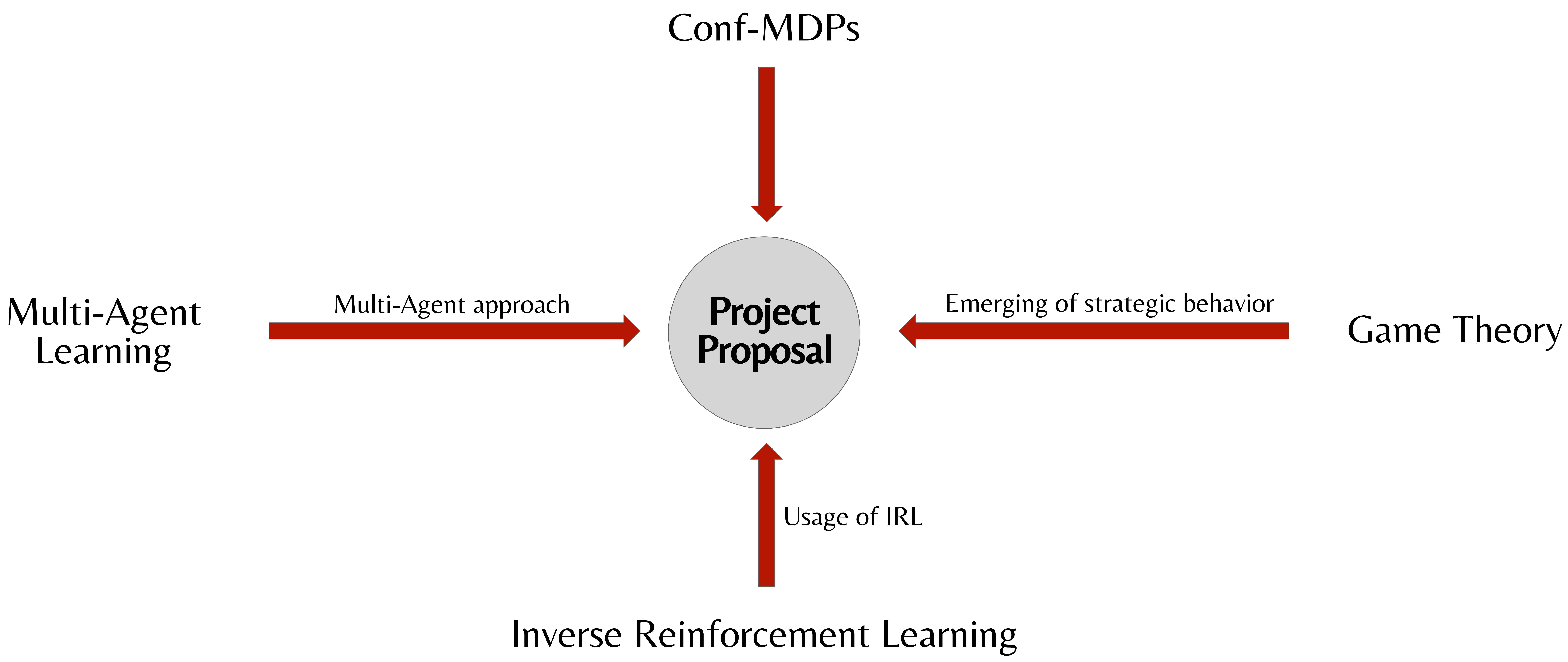
Real-world applications of Non-Cooperative Conf-MDPs

Design of road networks

Outline

- Preliminaries
- Motivation
- **State of the art**
- Research plan

State of the art



Conf-MDP

Configurable Markov Decision Processes

Alberto Maria Metelli [1*] Mirco Mutti [1*] Marcello Restelli [1]

I

Reinforcement Learning in Configurable Continuous Environments

Alberto Maria Metelli [1] Emanuele Ghelfi [1] Marcello Restelli [1]

II

Policy Space Identification in Configurable Environments

Alberto Maria Metelli, Guglielmo Manneschi, Marcello Restelli
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Piazza Leonardo da Vinci, 32, 20133, Milano, Italy
albertomaria.metelli@polimi.it, guglielmo.manneschi@mail.polimi.it, marcello.restelli@polimi.it

III

Conf-MDP (I)

Configurable Markov Decision Processes

Alberto Maria Metelli [1*] Mirco Mutti [1*] Marcello Restelli [1]

(Jun 2018)

- Theoretical formalization of the novel framework
- Safe Model-Policy Iteration (SMPI)
- Applicable in **finite** and **completely known** environments

Conf-MDP (II)

Reinforcement Learning in Configurable Continuous Environments

Alberto Maria Metelli [1] Emanuele Ghelfi [1] Marcello Restelli [1]

(Jun 2019)

- New learning algorithm: *Relative Entropy Model-Policy Search* (REMPS)
- Two phases:
  - Optimization
  - Projection
- Applicable to **unknown** and **continuous** environments

Conf-MDP (III)

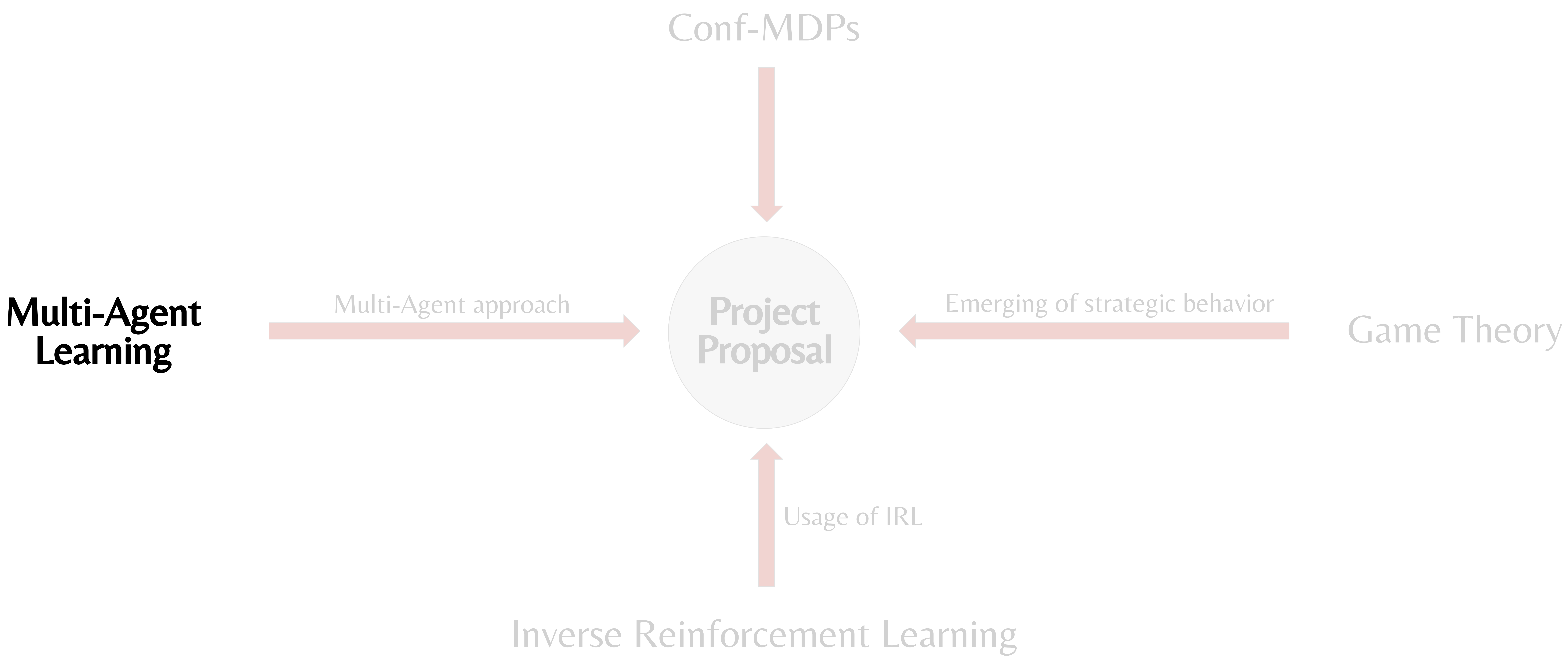
Policy Space Identification in Configurable Environments

Alberto Maria Metelli, Guglielmo Manneschi, Marcello Restelli
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Piazza Leonardo da Vinci, 32, 20133, Milano, Italy
albertomaria.metelli@polimi.it, guglielmo.manneschi@mail.polimi.it, marcello.restelli@polimi.it

(Sep 2019)

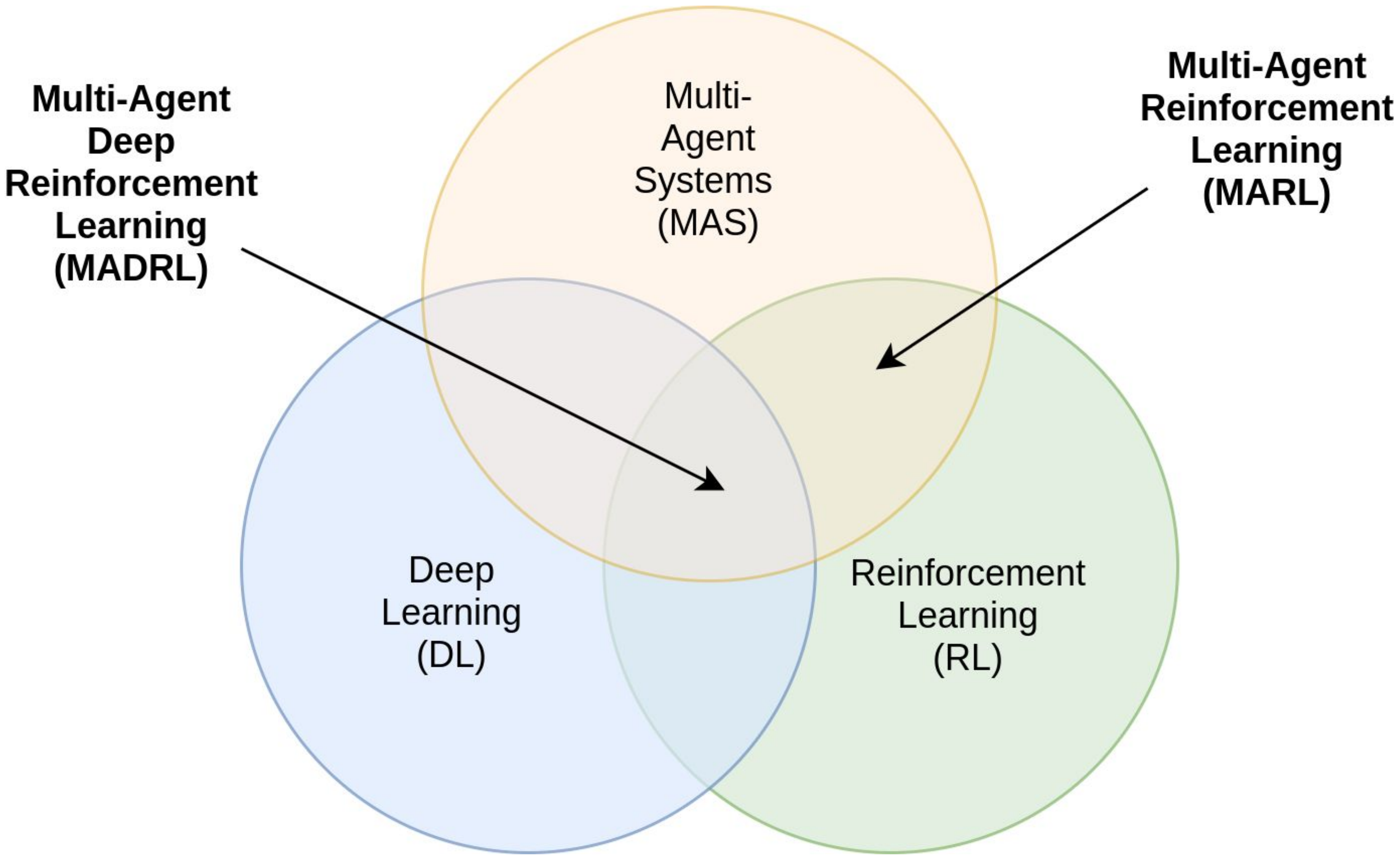
- The Conf-MDP is used to simplify the identification of the policy of an agent.
- Configuring the environment is useful to distinguish useless parameters from non-controllable ones.

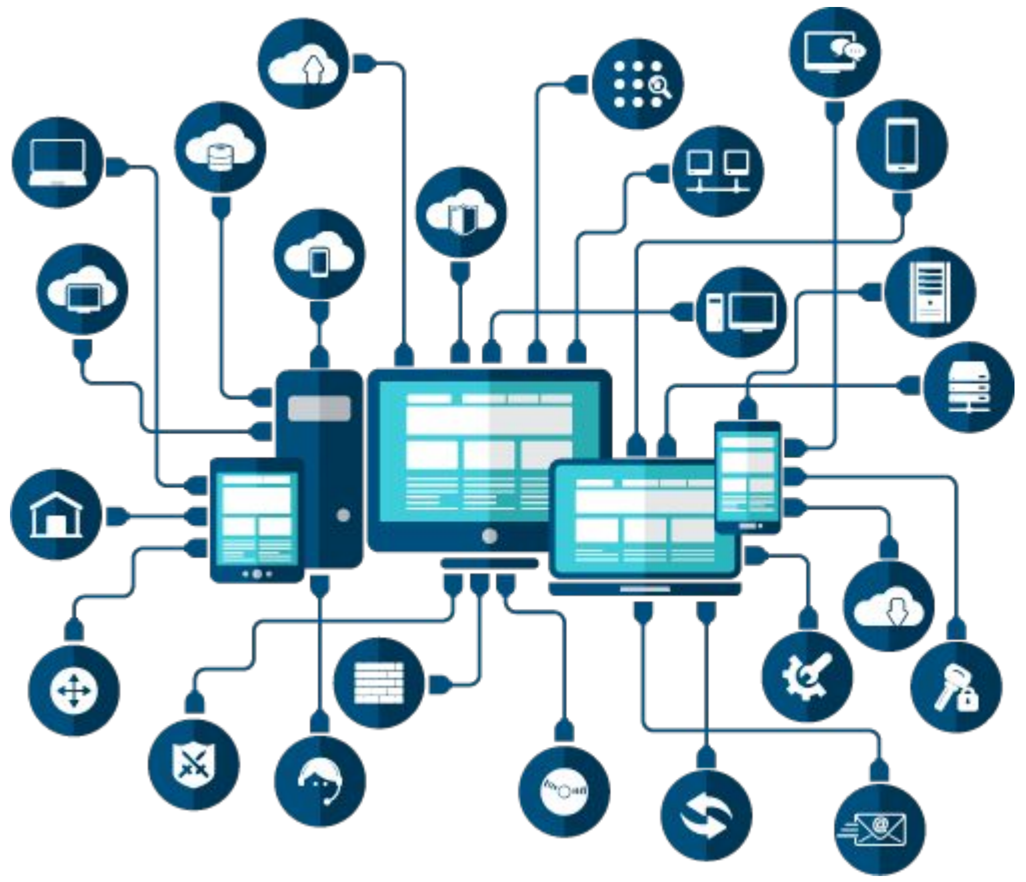
State of the art



Multi-Agent Learning (MAL)

~~MDP~~ → Stochastic Games

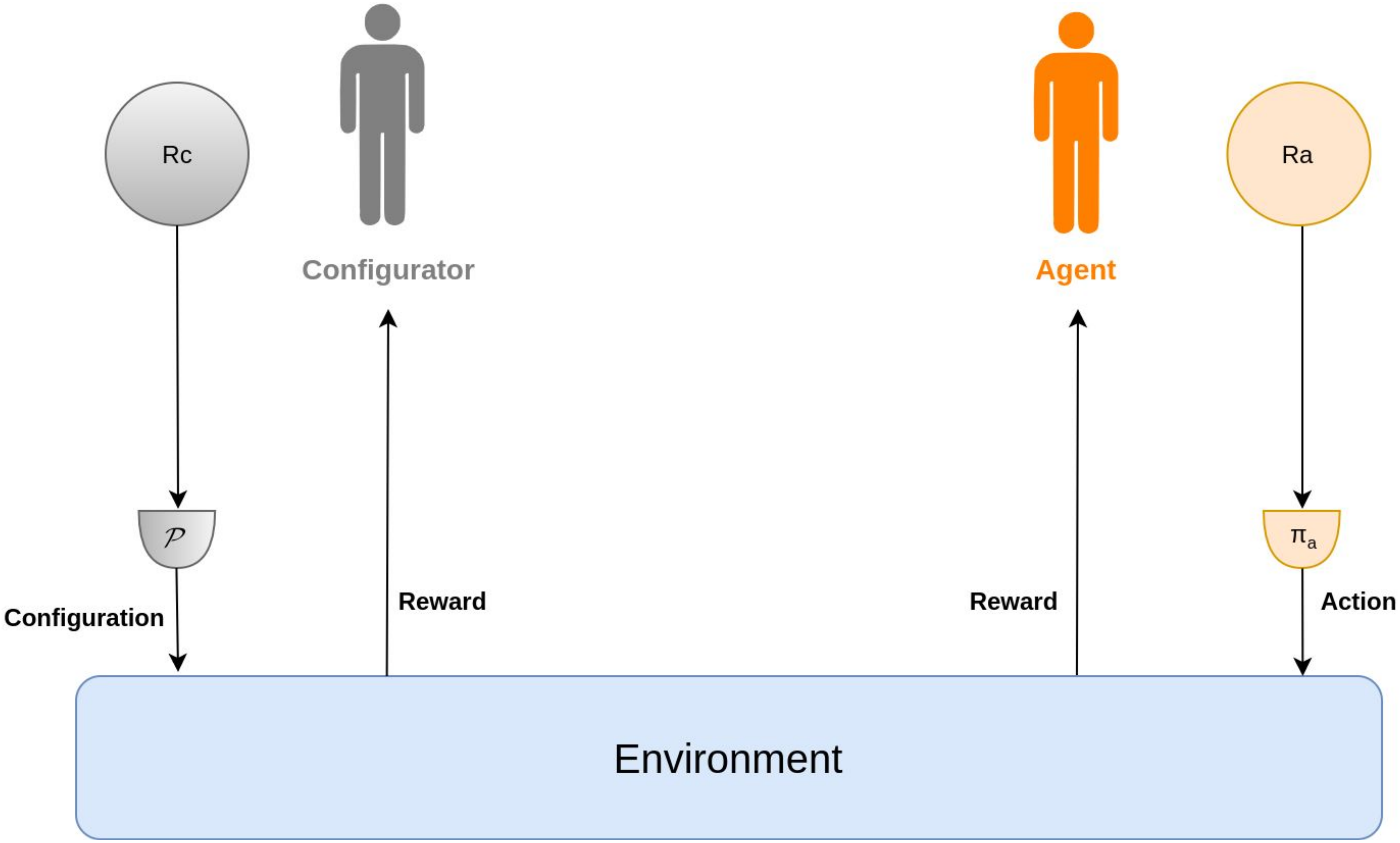




Learning in Multiagent environments

- Finding the optimal policy is not as obvious as the single agent case
- Coalition formation
- Partially observable environments
- Non-stationary environments

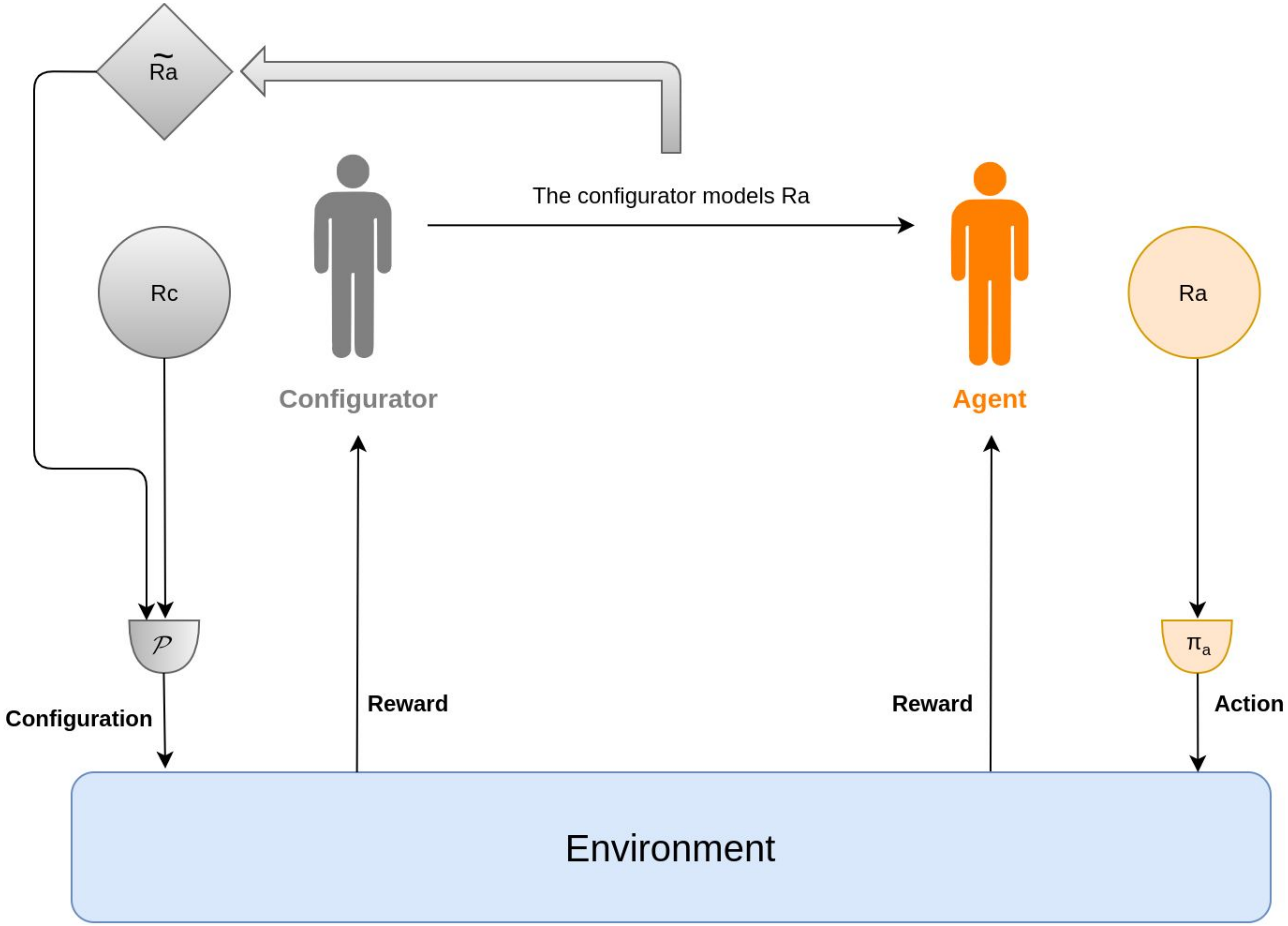
MAL in Conf-MDP



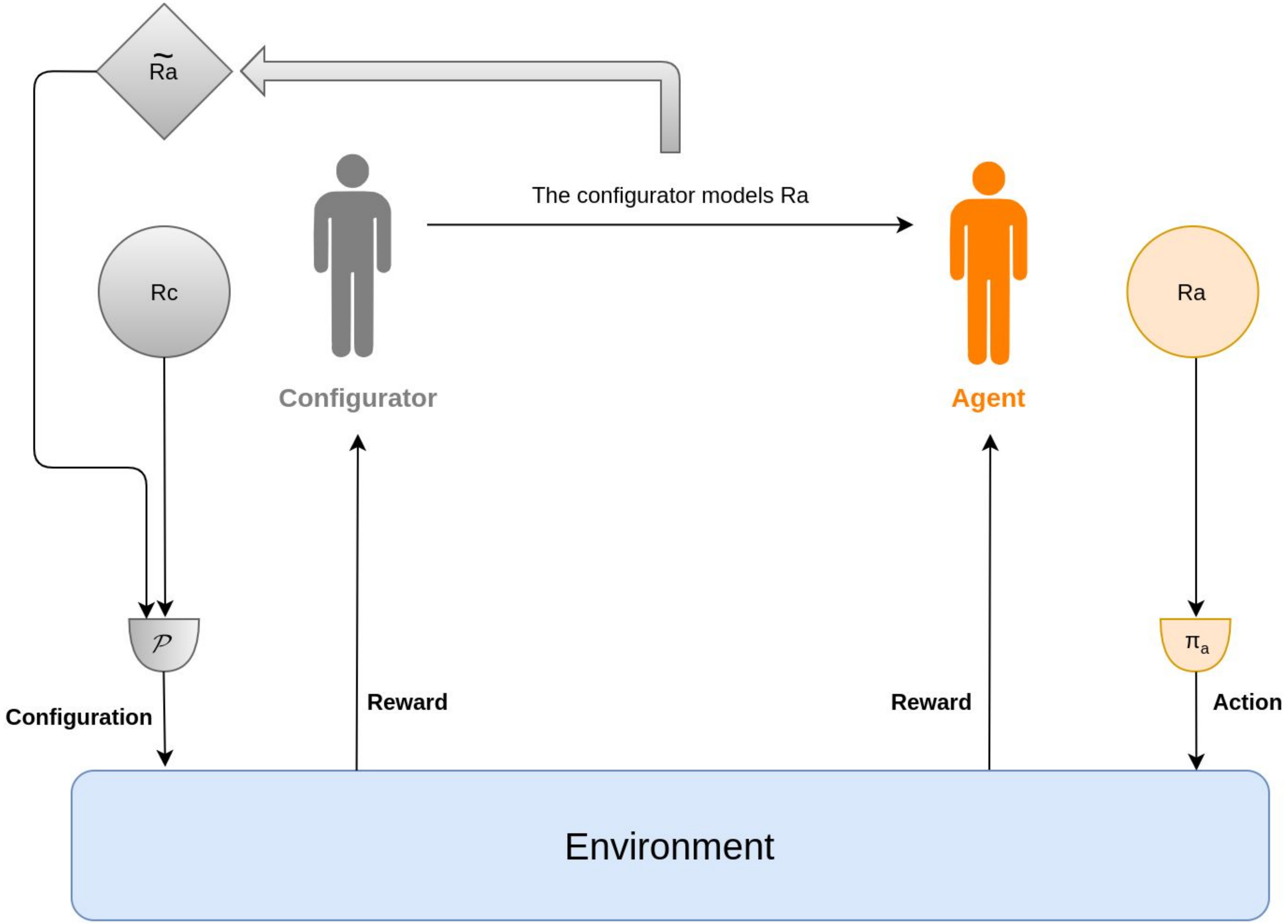
MAL in Conf-MDP

The **configurator** models the agent's behavior recovering its reward function

- More difficult if it has partial information

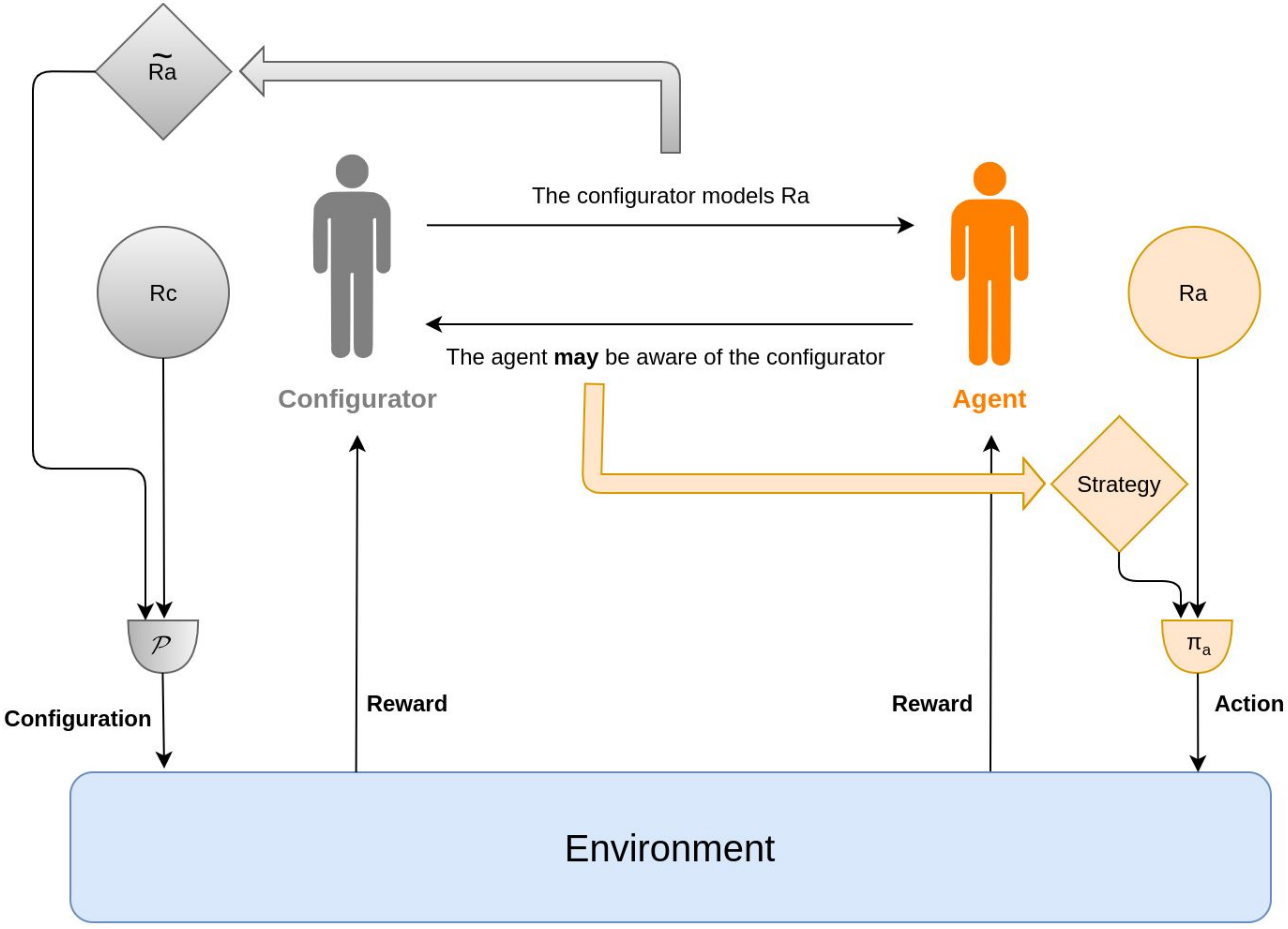


MAL in Conf-MDP



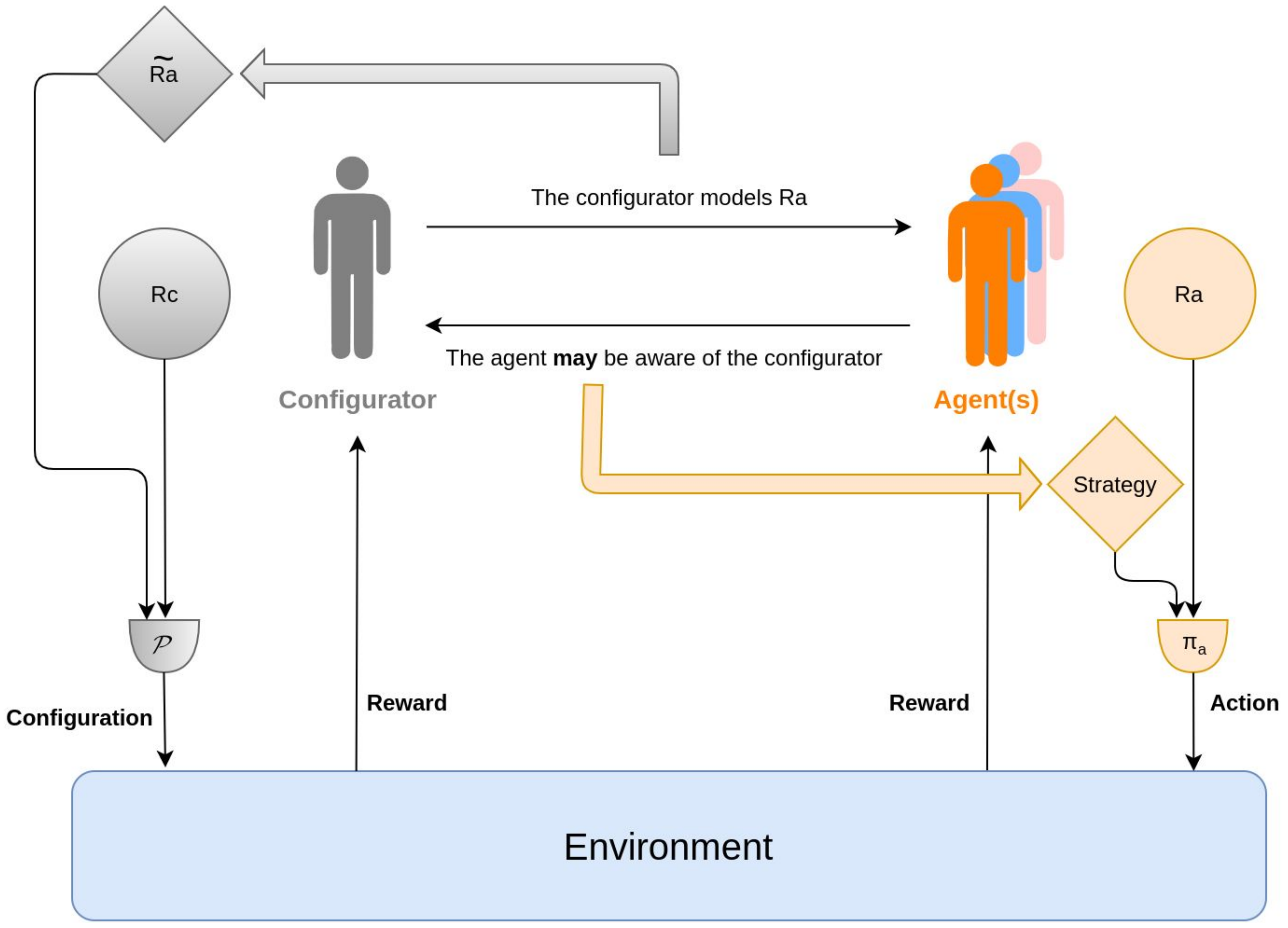
- The **agent** could follow possible strategies:
  1. Ignore environmental changes
  2. Forget previous configurations

MAL in Conf-MDP



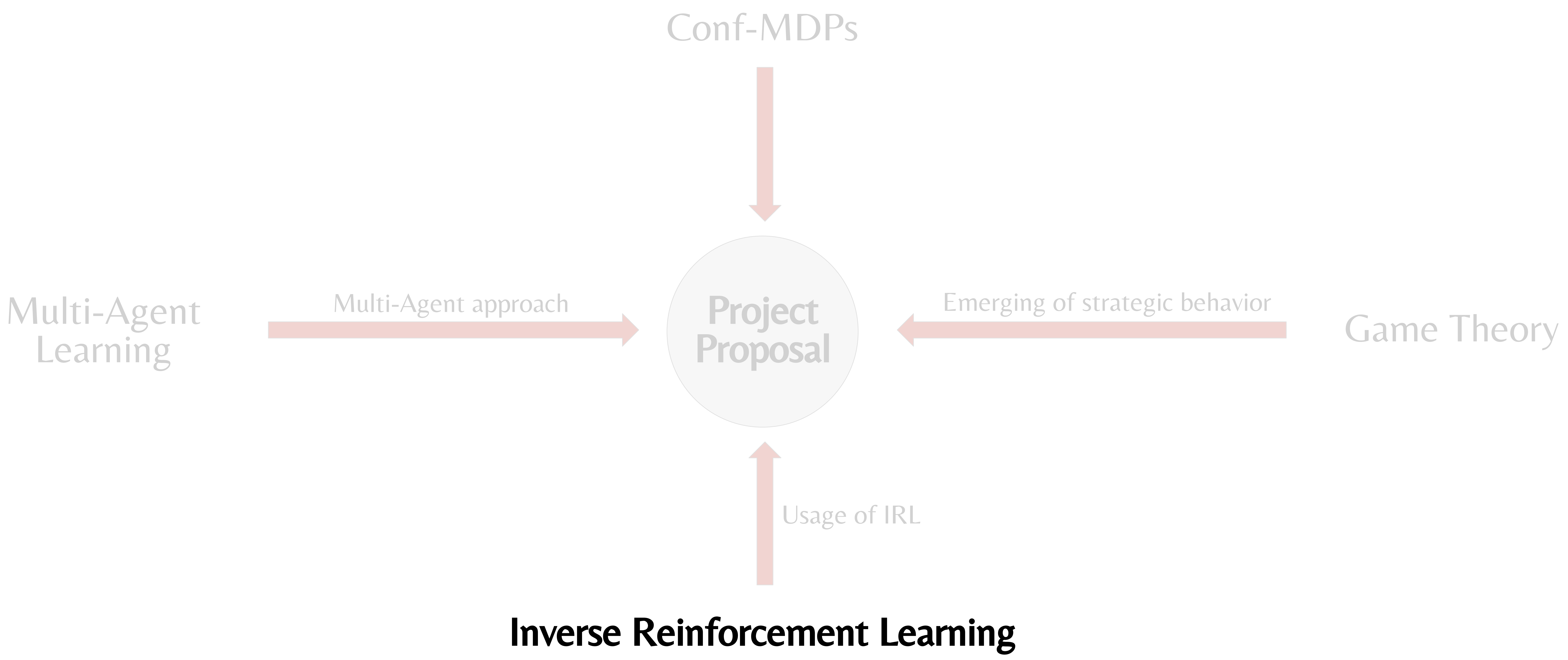
- The **agent** could follow possible strategies:
  1. Ignore environmental changes
  2. Forget previous configurations
  3. Awareness of the configurator

MAL in Conf-MDP

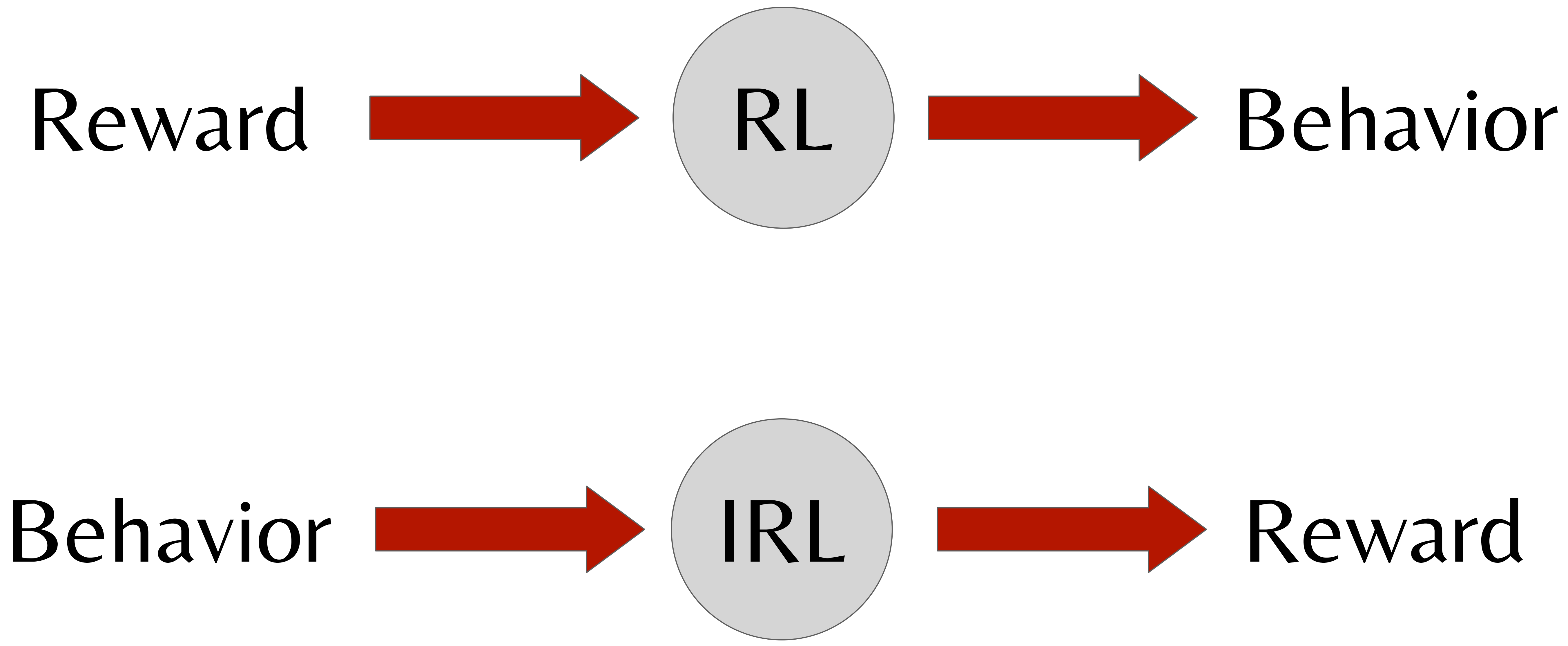


- The **agent** could follow possible strategies:
    1. Ignore environmental changes
    2. Forget previous configurations
    3. Awareness of the configurator
    4. Possible coalition formation

State of the art



Inverse Reinforcement Learning (IRL)



The goal of IRL is to recover the unknown reward function from the expert's demonstrations.

Why should we use IRL?

- When we want to know what are the reasons that induce the agent to choose some behaviors
- When the reward function is hard to design

Exemple of IRL

- A set of expert demonstrations *D* is given.
- **Goal:** find R(s,a) that is equivalent, in term of performance, to the *unknown* reward function R_E(s,a) of the expert
    - This means that we want similar state-action visitation frequency: $\mu_E \simeq \mu$

➔ Evaluate μ_E from D
➔ Initialize randomly the reward R
➔ Repeat until convergence
  ◆ Find the current policy π induced by R with RL techniques
  ◆ Evaluate μ of the current policy π
  ◆ Update R based on the comparison between μ and μ_E

Inverse Reinforcement Learning (IRL)

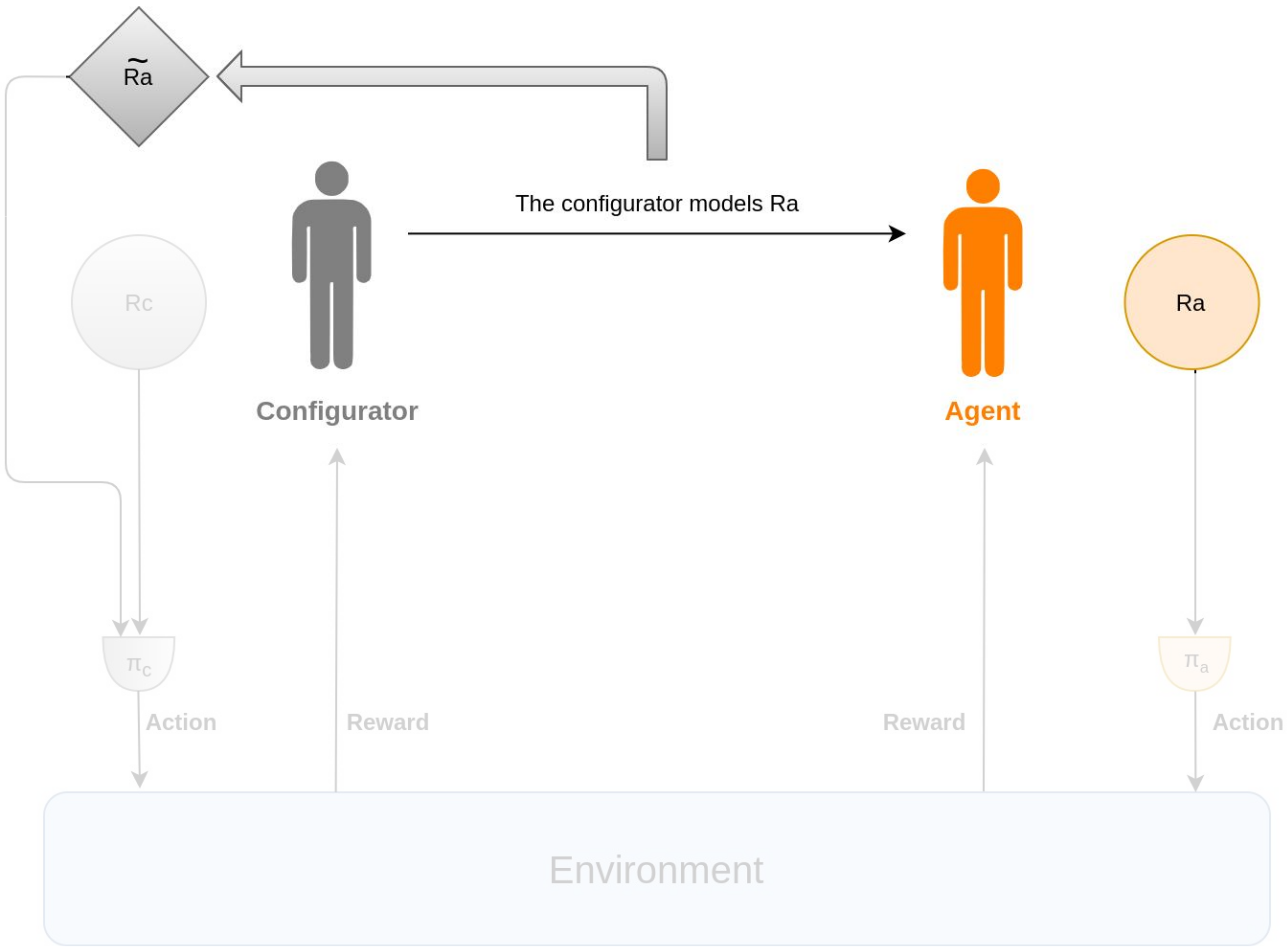
IRL is an **ill-posed** problem

- maximize the entropy
- maximize the margin between the optimal policy and the others

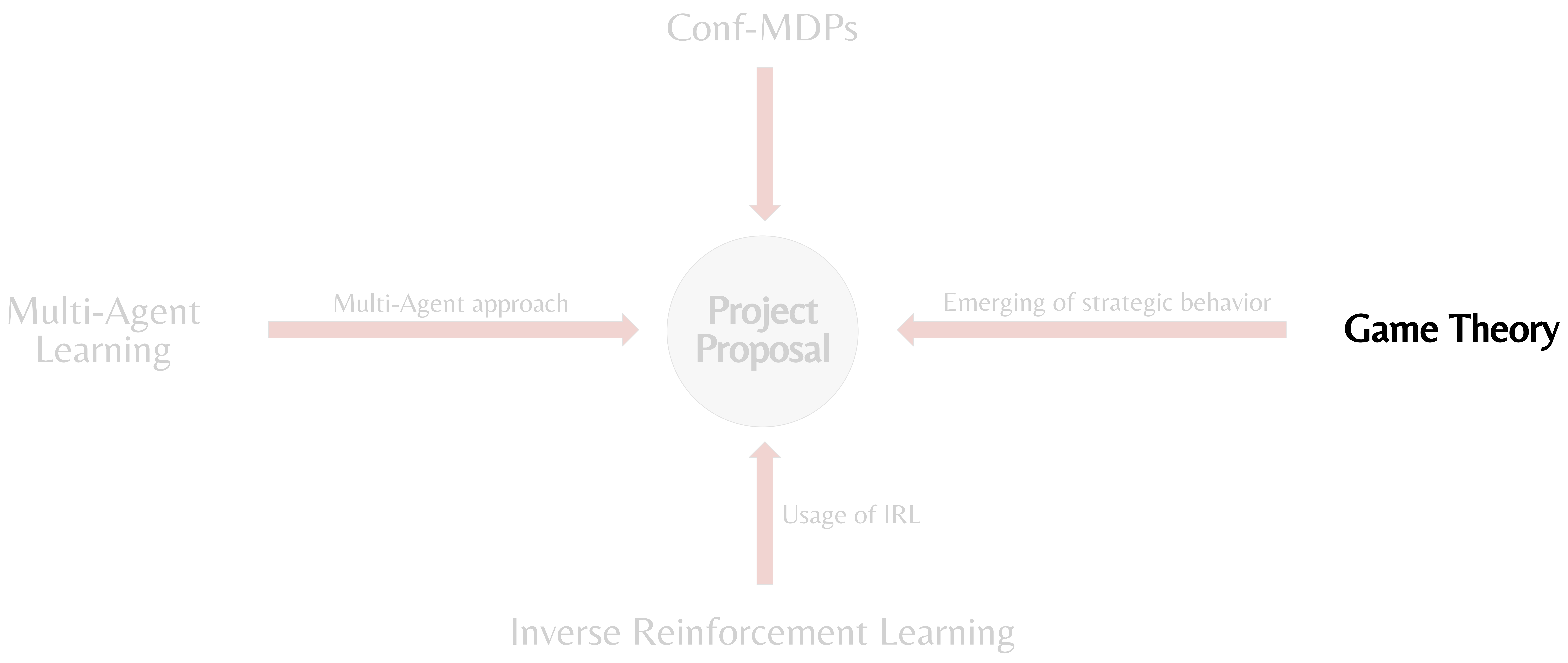
Two categories:

- Model-based
- Model-free
    - Interactive model-free
    - Batch model-free

IRL in Conf-MDP



State of the art



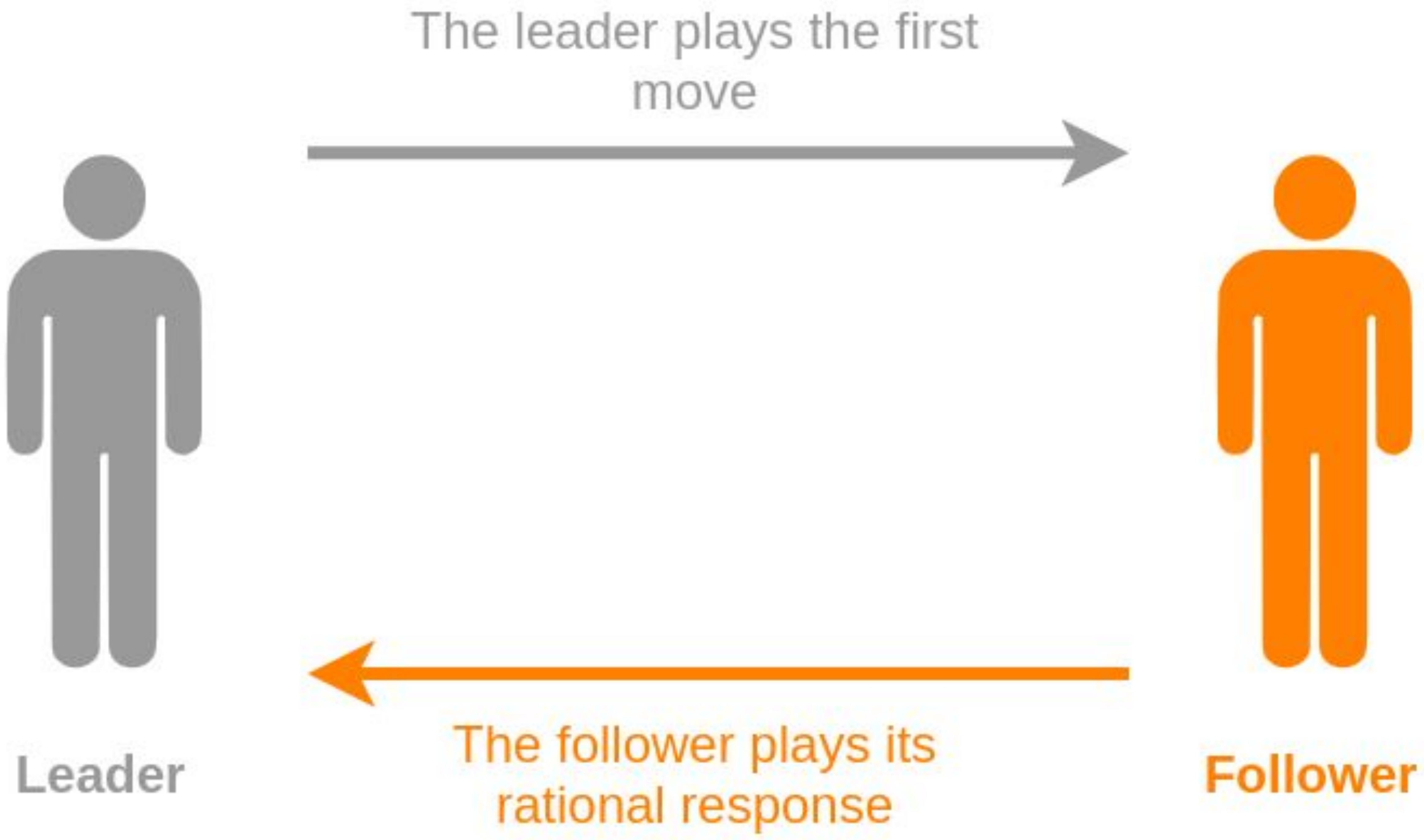
Game Theory (GT)

Game theory *is the study of mathematical models of strategic interaction among rational decision-makers.*

⬇

Stackelberg Games

Stackelberg Games



Stackelberg equilibrium

The leader (player 1) and the follower (player 2) aim to solve these optimization problems:

$$\min_{x_1 \in X_1} \left\{ f_1(x_1, x_2) \middle| \, x_2 \in \arg\min_{y \in X_2} f_2(x_1, y) \right\}$$

$$\min_{x_2 \in X_2} f_2(x_1, x_2)$$

A strategy x1* is called a **Stackelberg equilibrium strategy** for the leader if

$$\sup_{x_2 \in \mathcal{R}(x_1^*)} f_1(x_1^*, x_2) \leq \sup_{x_2 \in \mathcal{R}(x_1)} f_1(x_1, x_2), \quad \forall x_1 \in X_1,$$

where $\mathcal{R}(x_1) = \{y \in X_2 | \, f_2(x_1, y) \leq f_2(x_1, x_2), \forall x_2 \in X_2\}$ is the rational reaction set of x2.

Stackelberg Games

Convergence of Learning Dynamics in Stackelberg Games

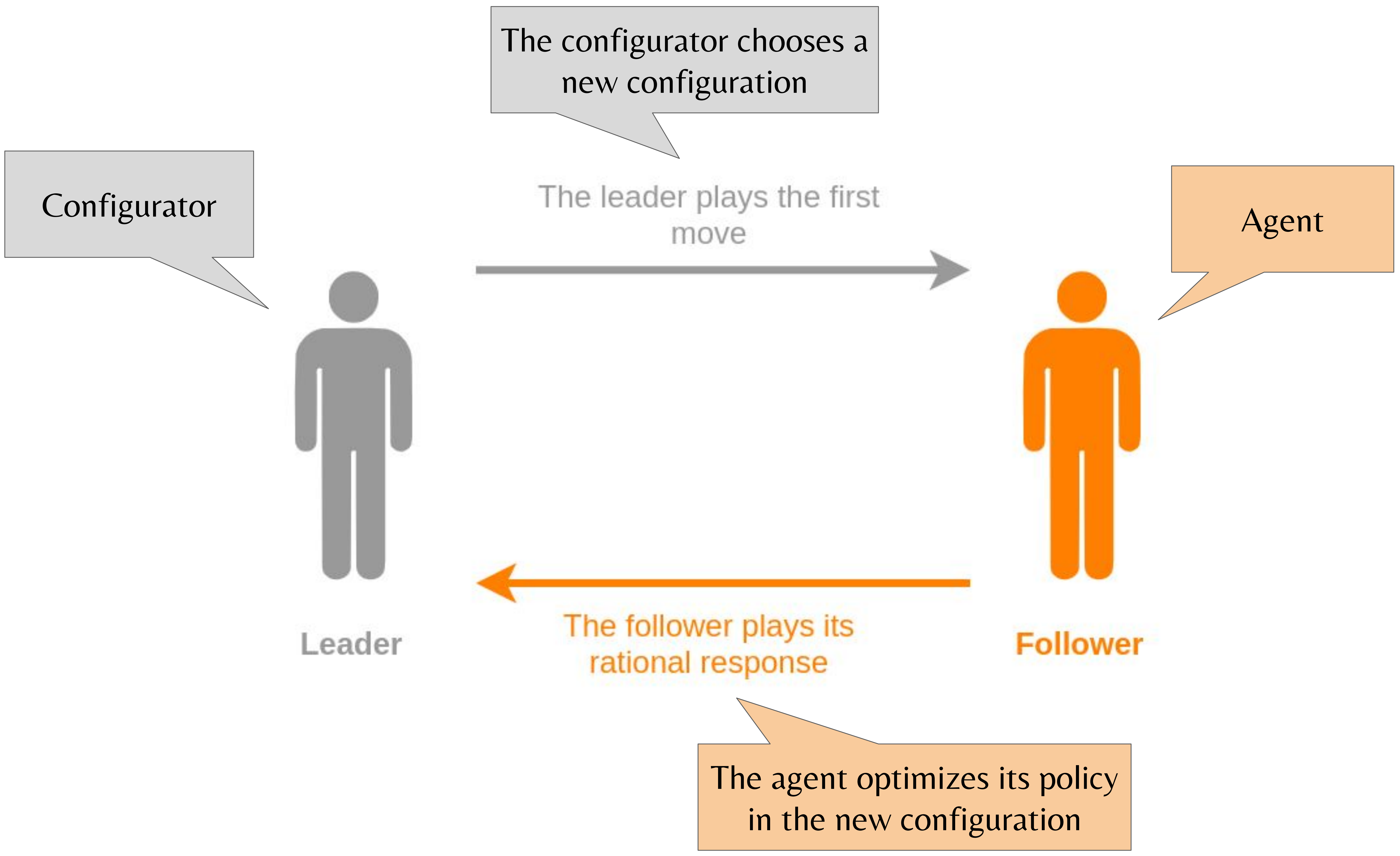
Tanner Fiez FIEZT@UW.EDU
Department of Electrical and Computer Engineering
University of Washington

Benjamin Chasnov BCHASNOV@UW.EDU
Department of Electrical and Computer Engineering
University of Washington

Lillian J. Ratliff RATLIFFL@UW.EDU
Department of Electrical and Computer Engineering
University of Washington

- Investigate the relationship between Nash and Stackelberg equilibria
- Provide a learning rule for the leader that provably converges to a Stackelberg equilibrium

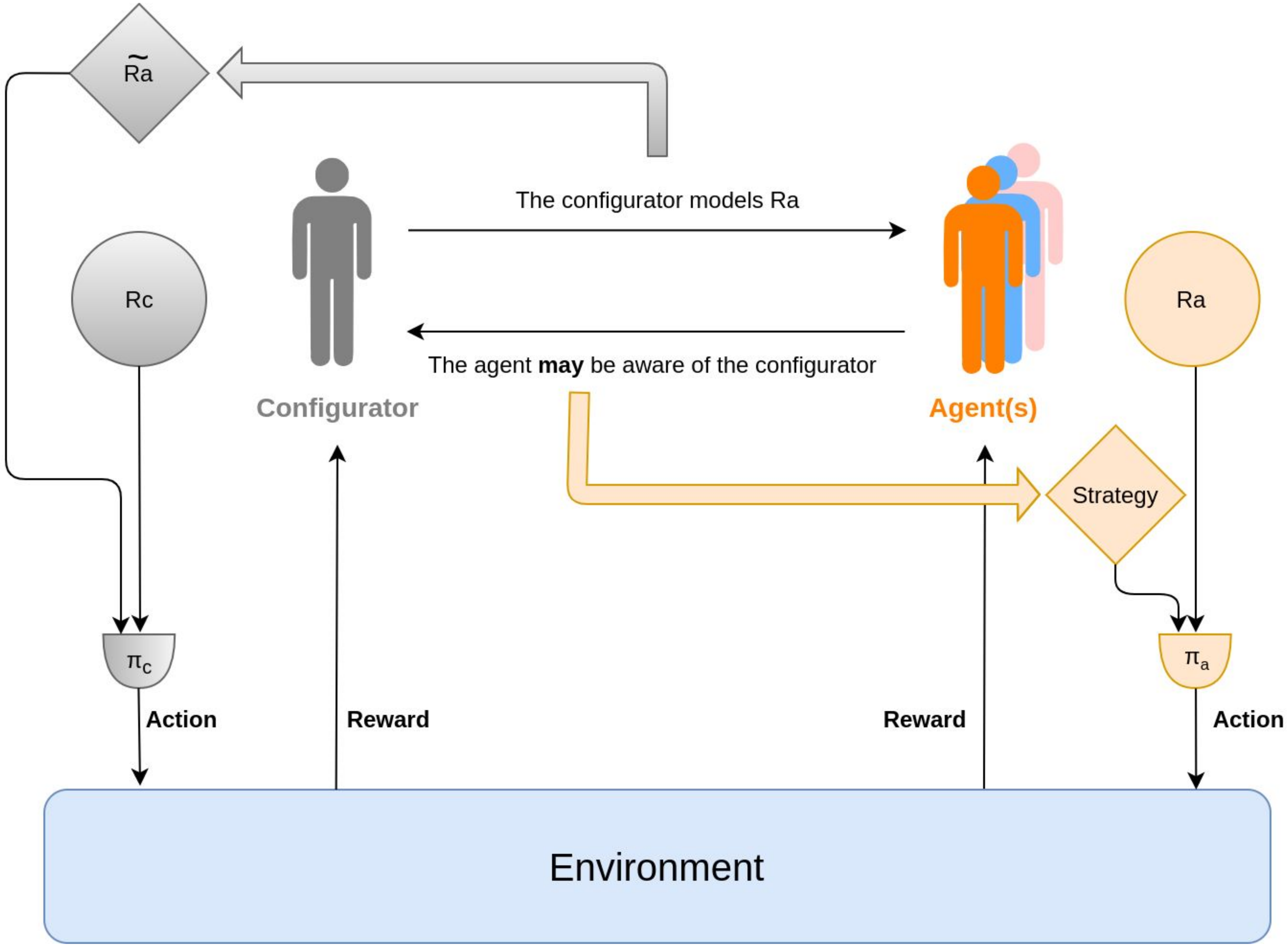
Stackelberg Games in Conf-MDP



Outline

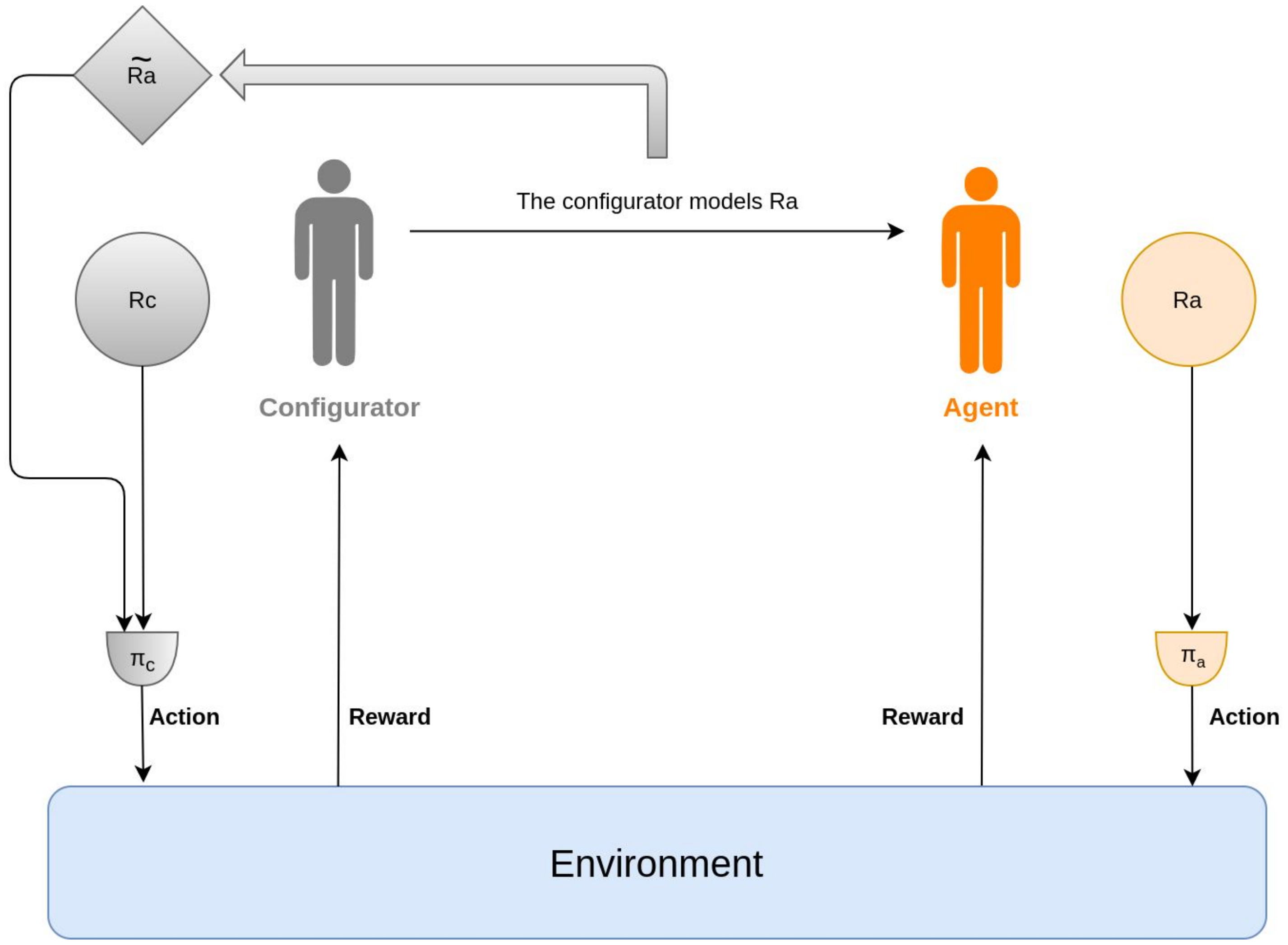
- Preliminaries
- Motivation
- State of the art
- **Research plan**

Non-cooperative Conf-MDP



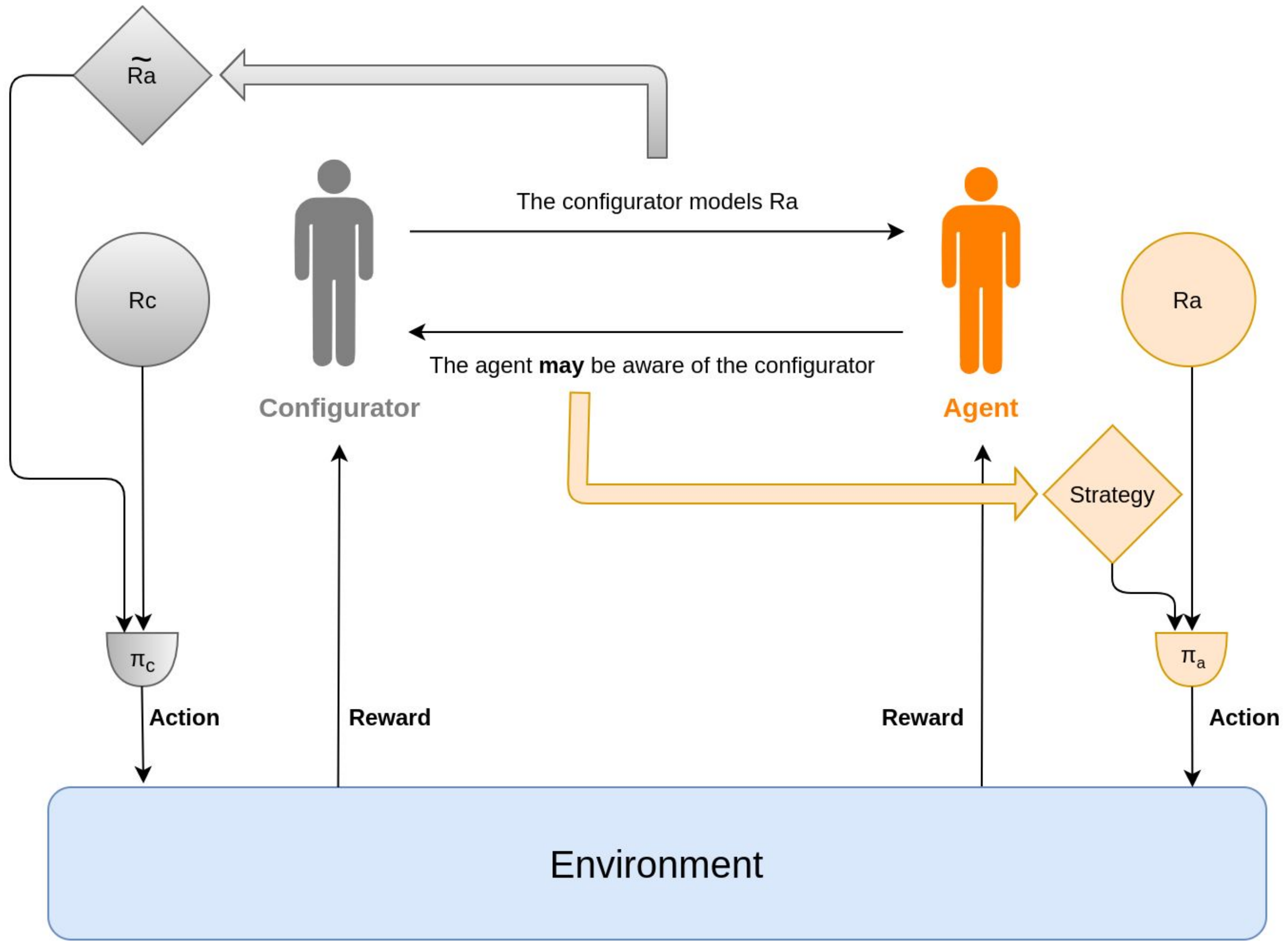
Possible assumptions

IRL process



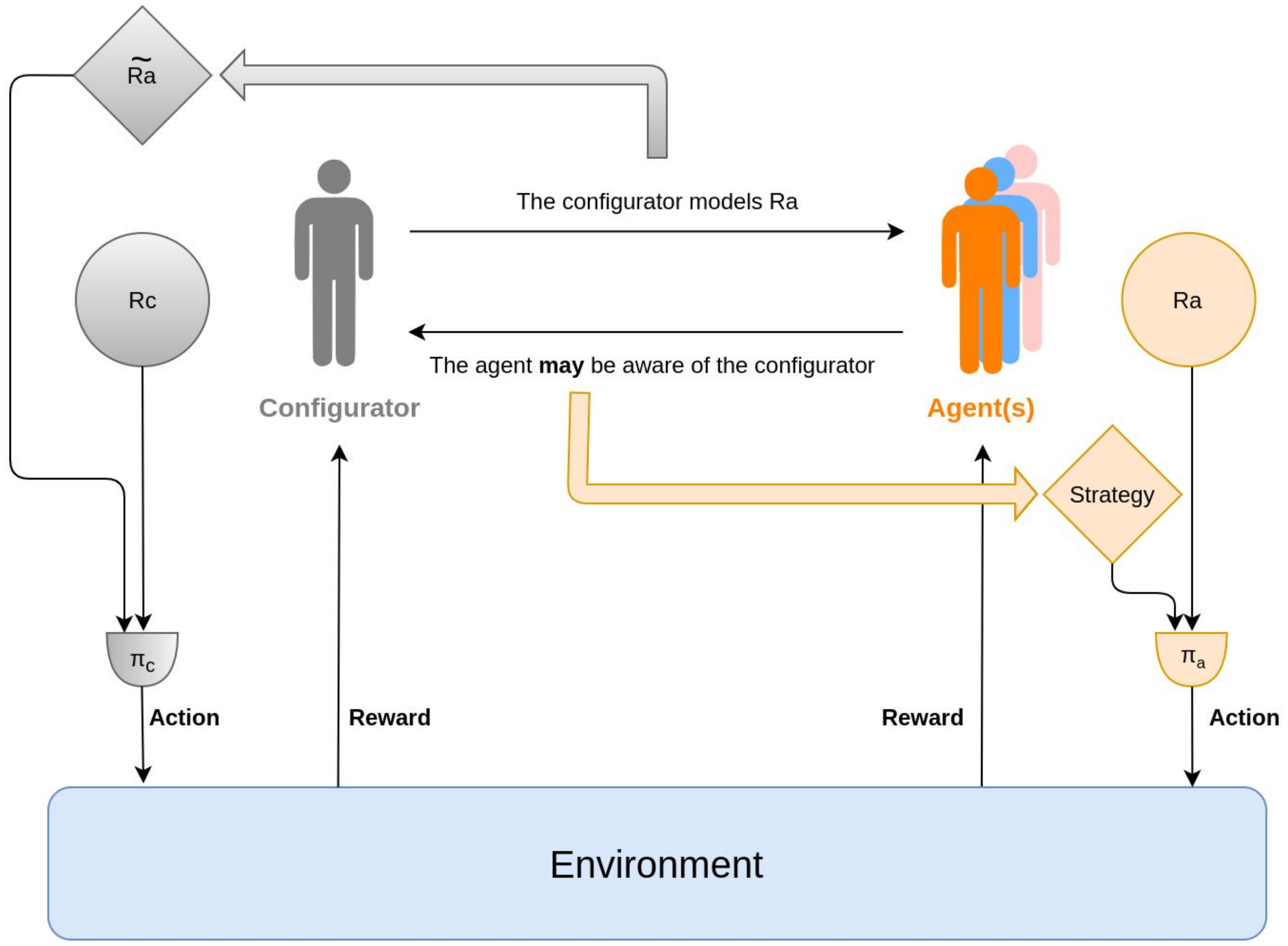
- The configurator is omniscient
- The configurator has partial information

Agent's awareness



- The agent is unaware
- The agent is aware

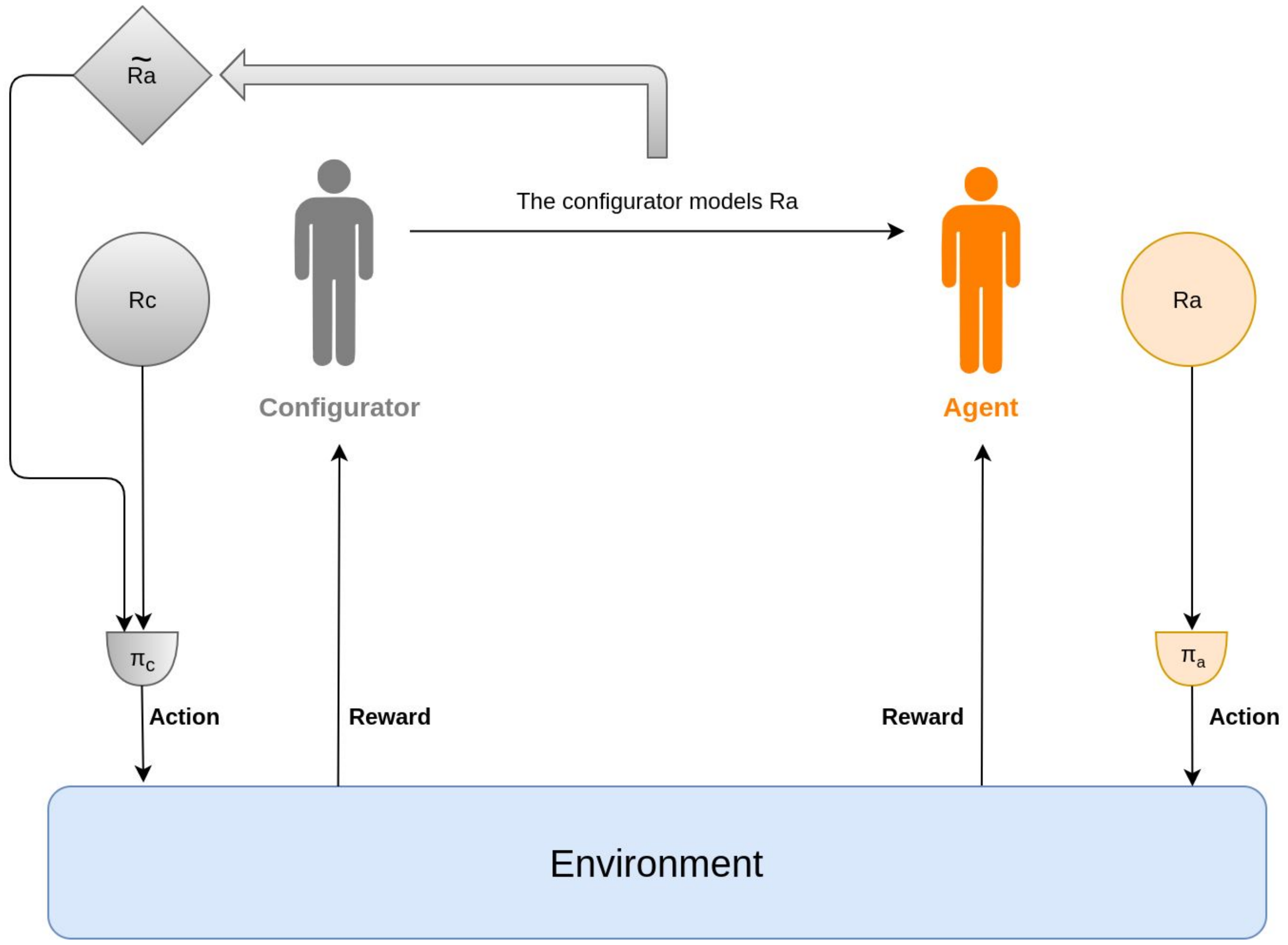
Possible multiple agents



- Single agent
- Multiple agents

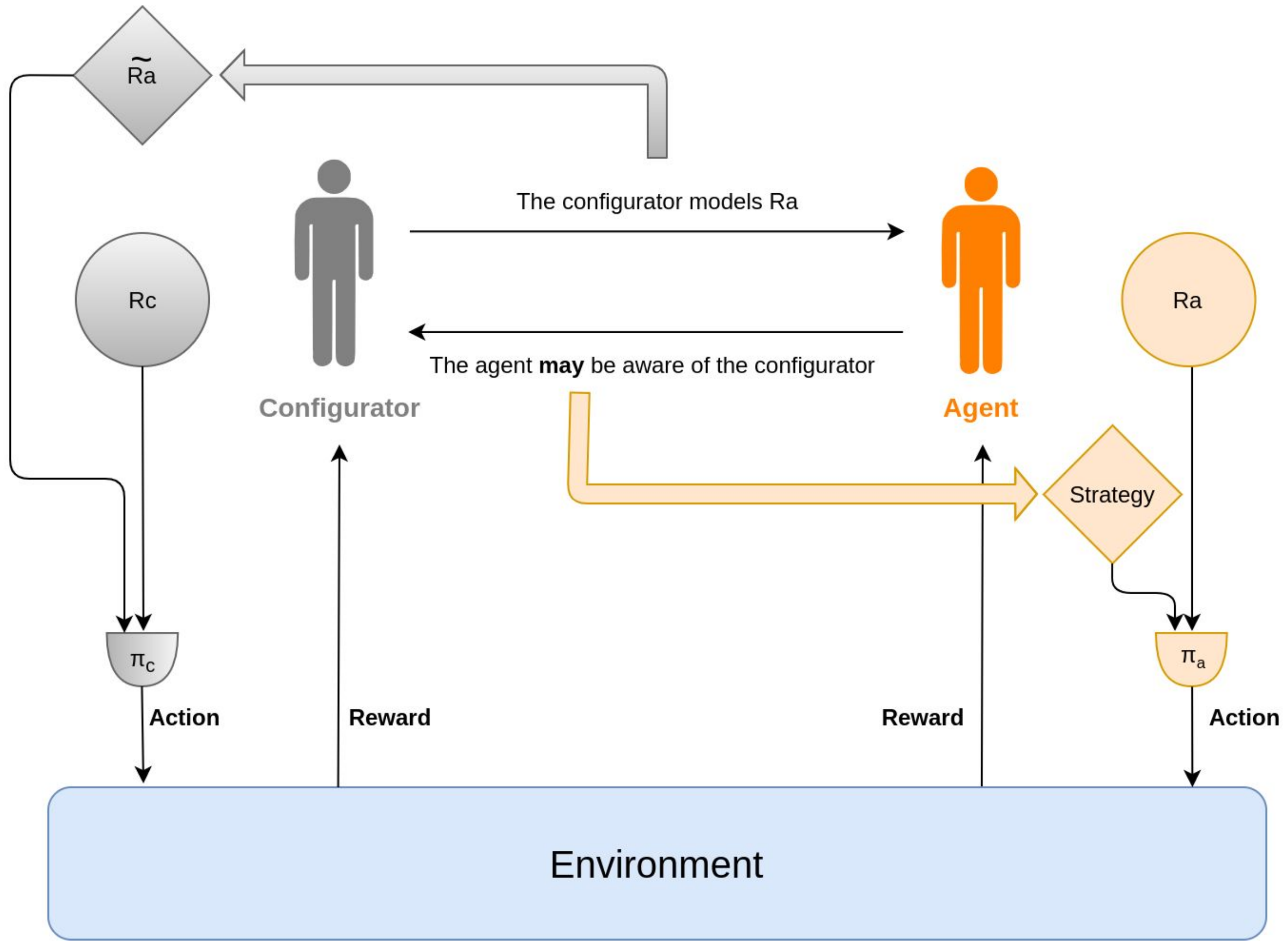
Possible assumptions

IRL process



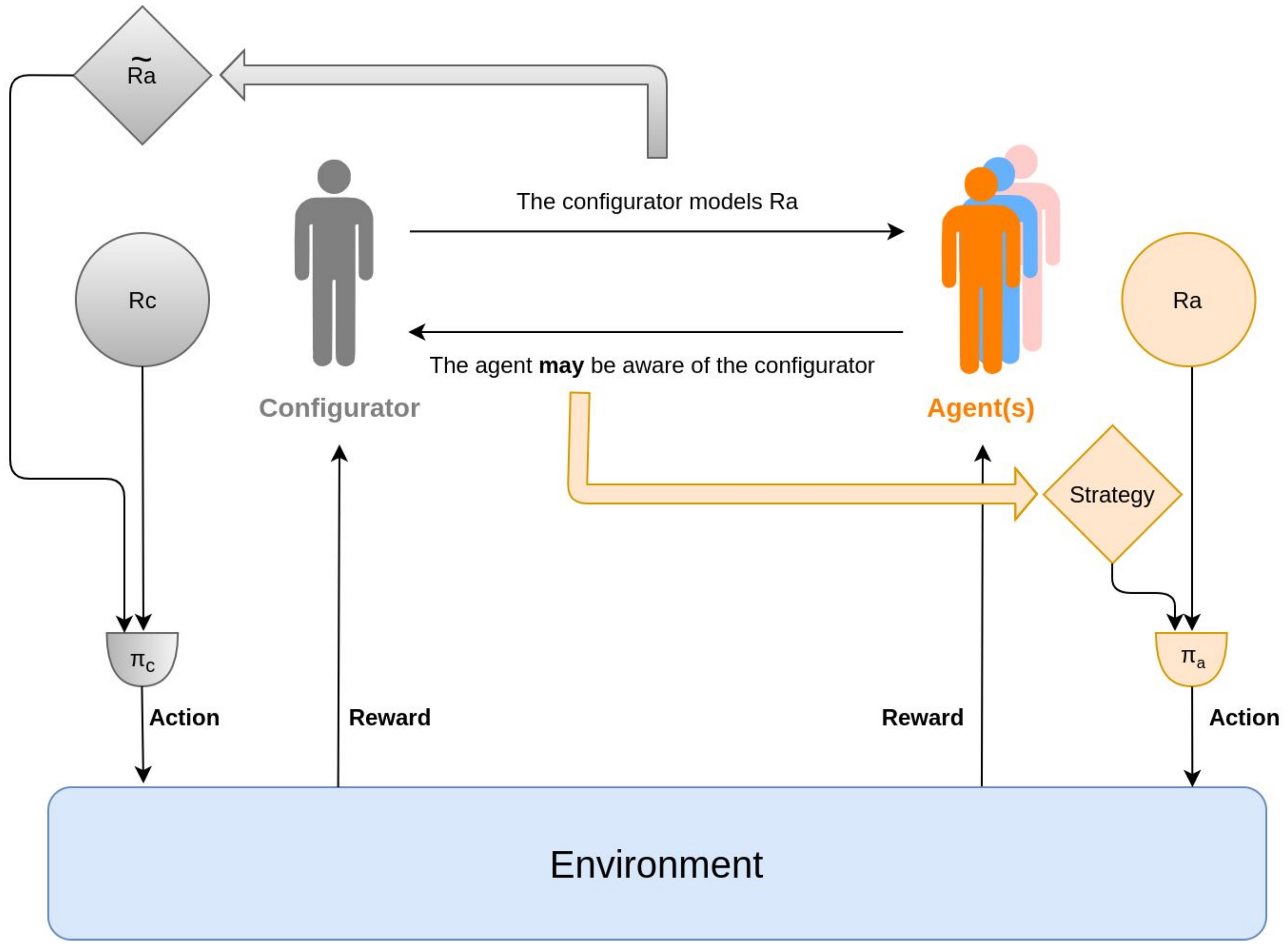
- **The configurator is omniscient**
- The configurator has partial information

Agent's awareness



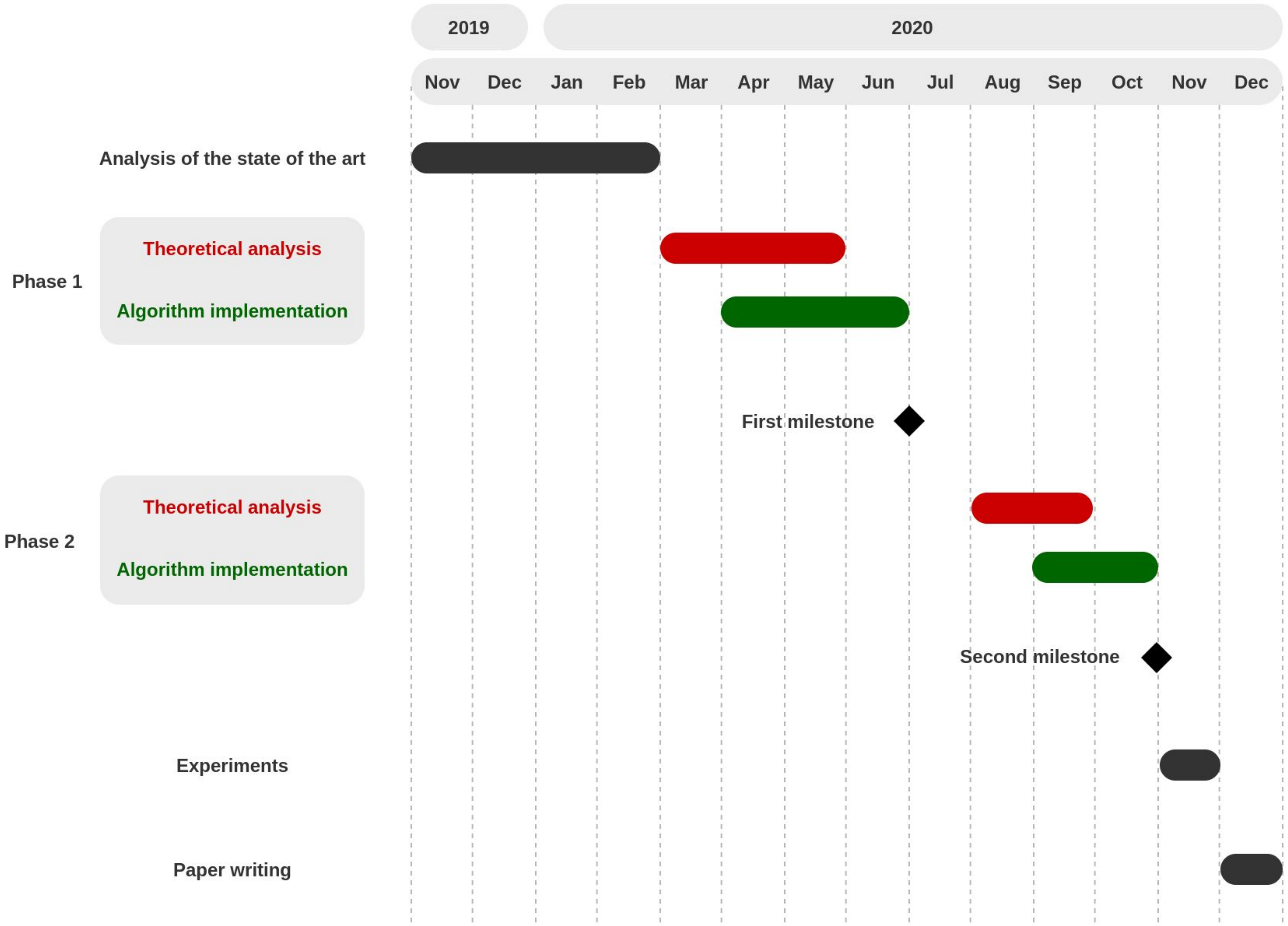
- **The agent is unaware**
- **The agent is aware**

Multiple agents



- **Single agent**
- Multiple agents

Project plan



Thank you for your attention!

Alessandro Concetti