Learning behaviours aligned with moral values in a multi-agent system:
guiding reinforcement learning with symbolic judgments

PhD Defence of **Rémy CHAPUT**

LIRIS CNRS UMR5205 / Université Claude Bernard Lyon 1

Supervised by: Salima Hassas, Olivier Boissier, Mathieu Guillermin

27 October 2022

Apprentissage de comportements alignés sur des valeurs morales dans un système multi-agent :

guider l'apprentissage par renforcement avec des jugements symboliques

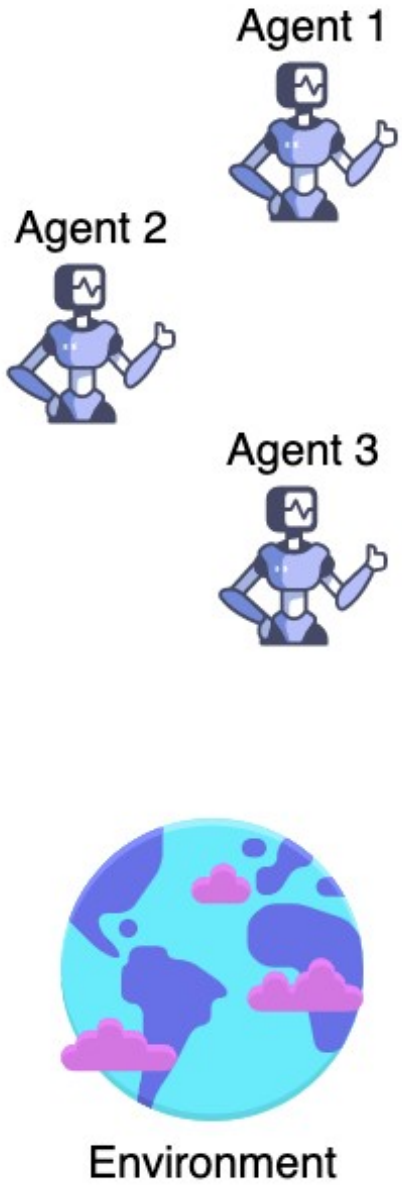
Soutenance de thèse de **Rémy CHAPUT**

LIRIS CNRS UMR5205 / Université Claude Bernard Lyon 1

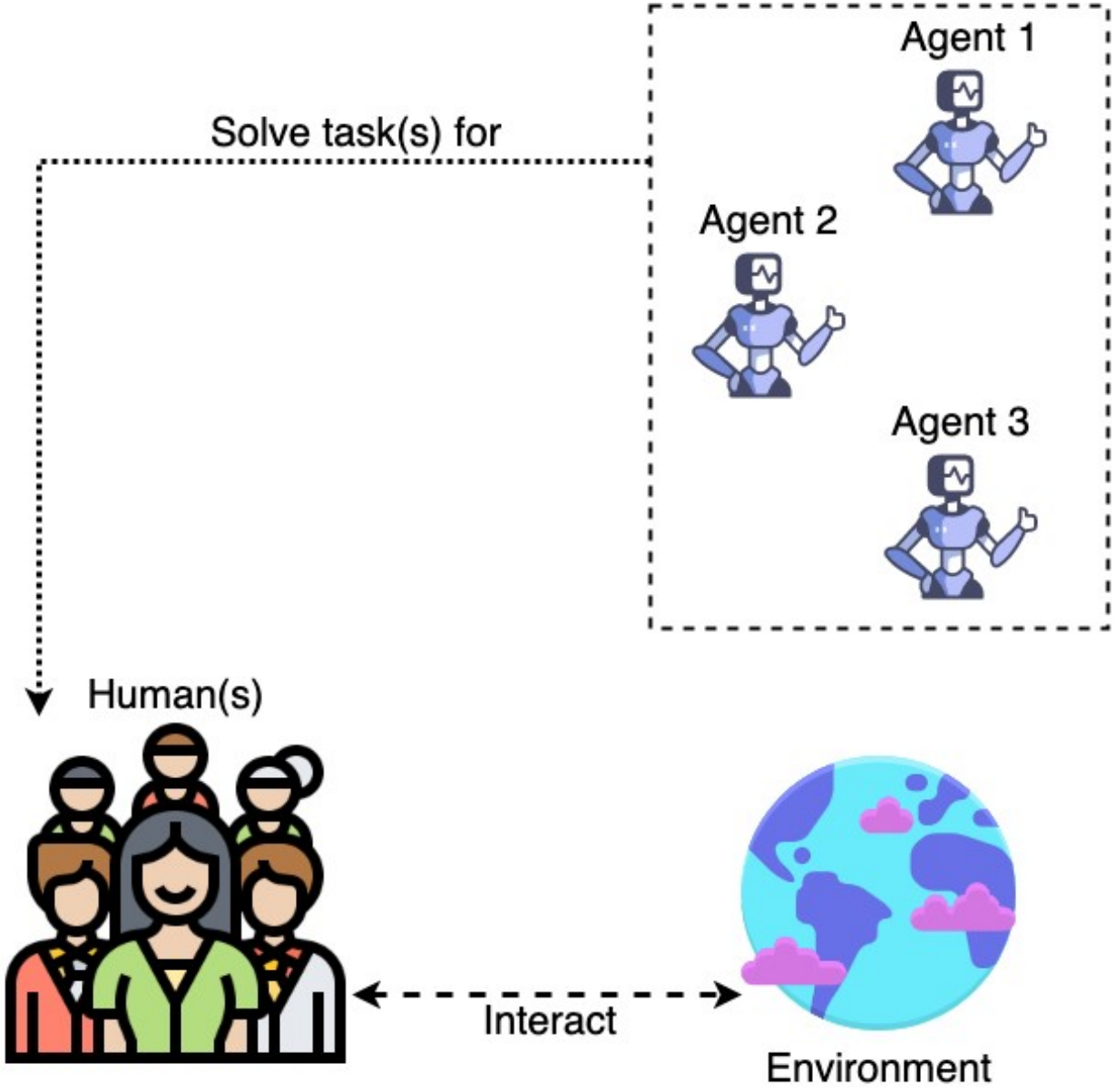
Supervisé par : Salima Hassas, Olivier Boissier, Mathieu Guillermin

27 Octobre 2022

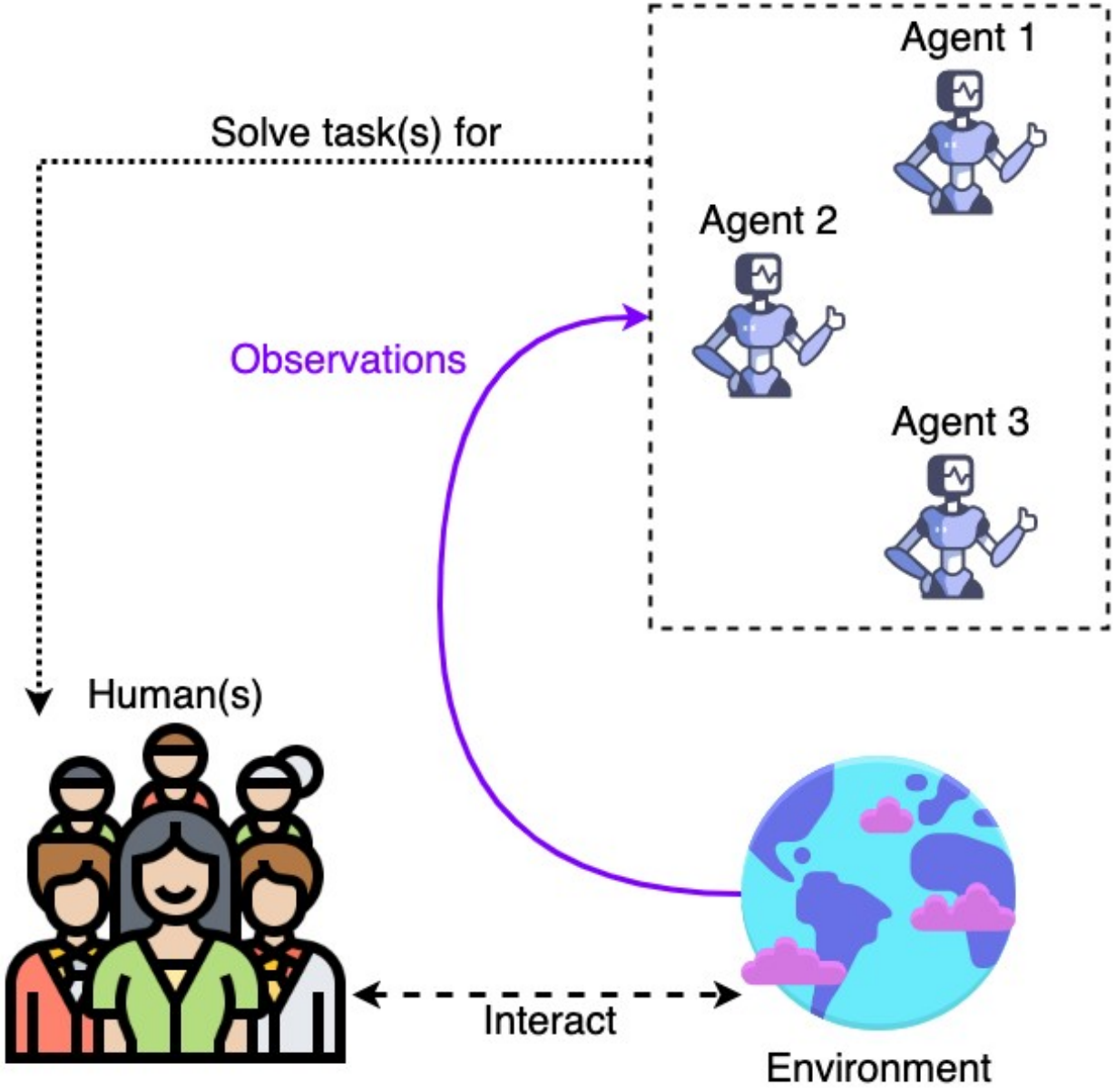
AI systems have an impact over humans



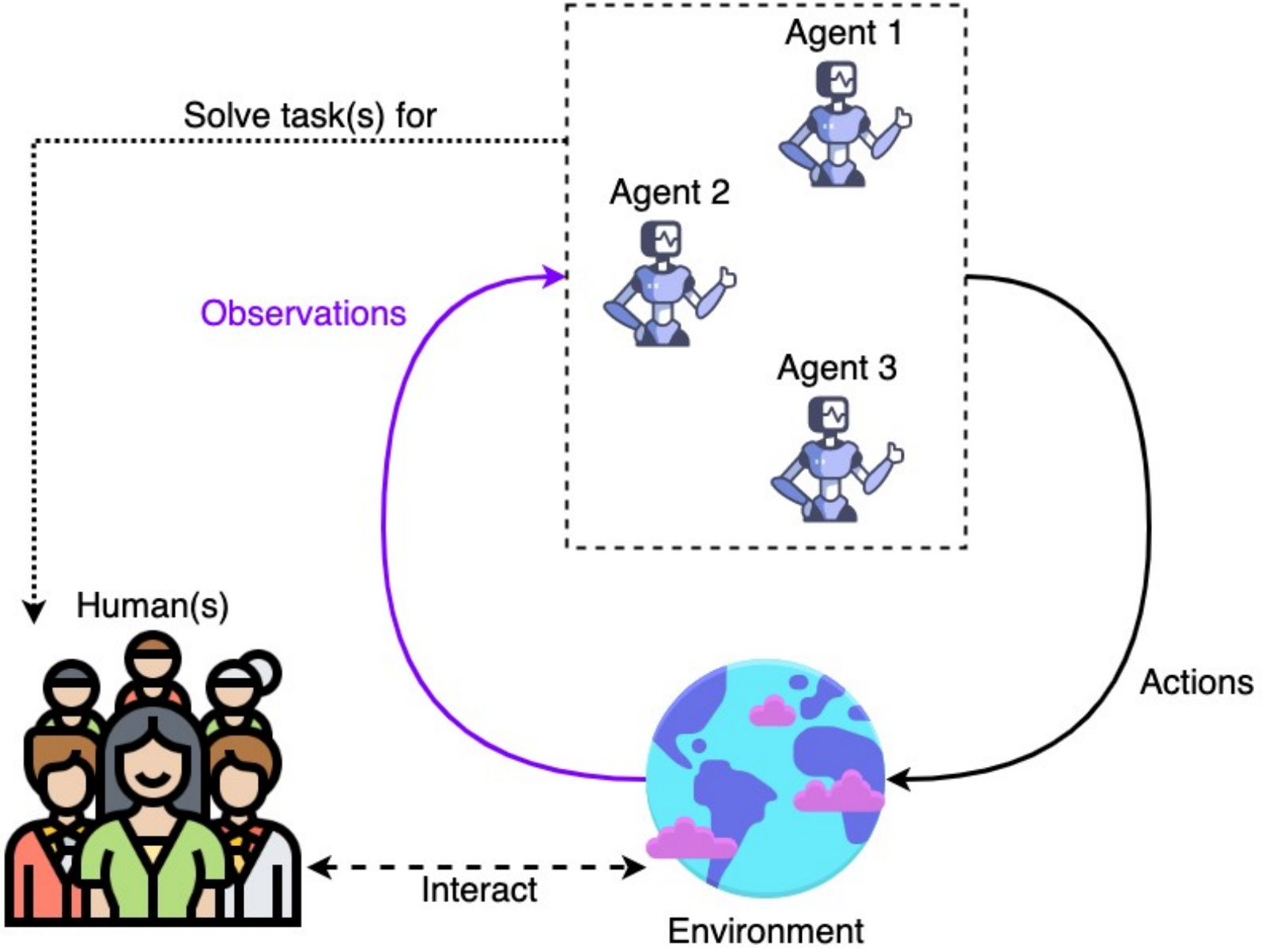
AI systems have an impact over humans



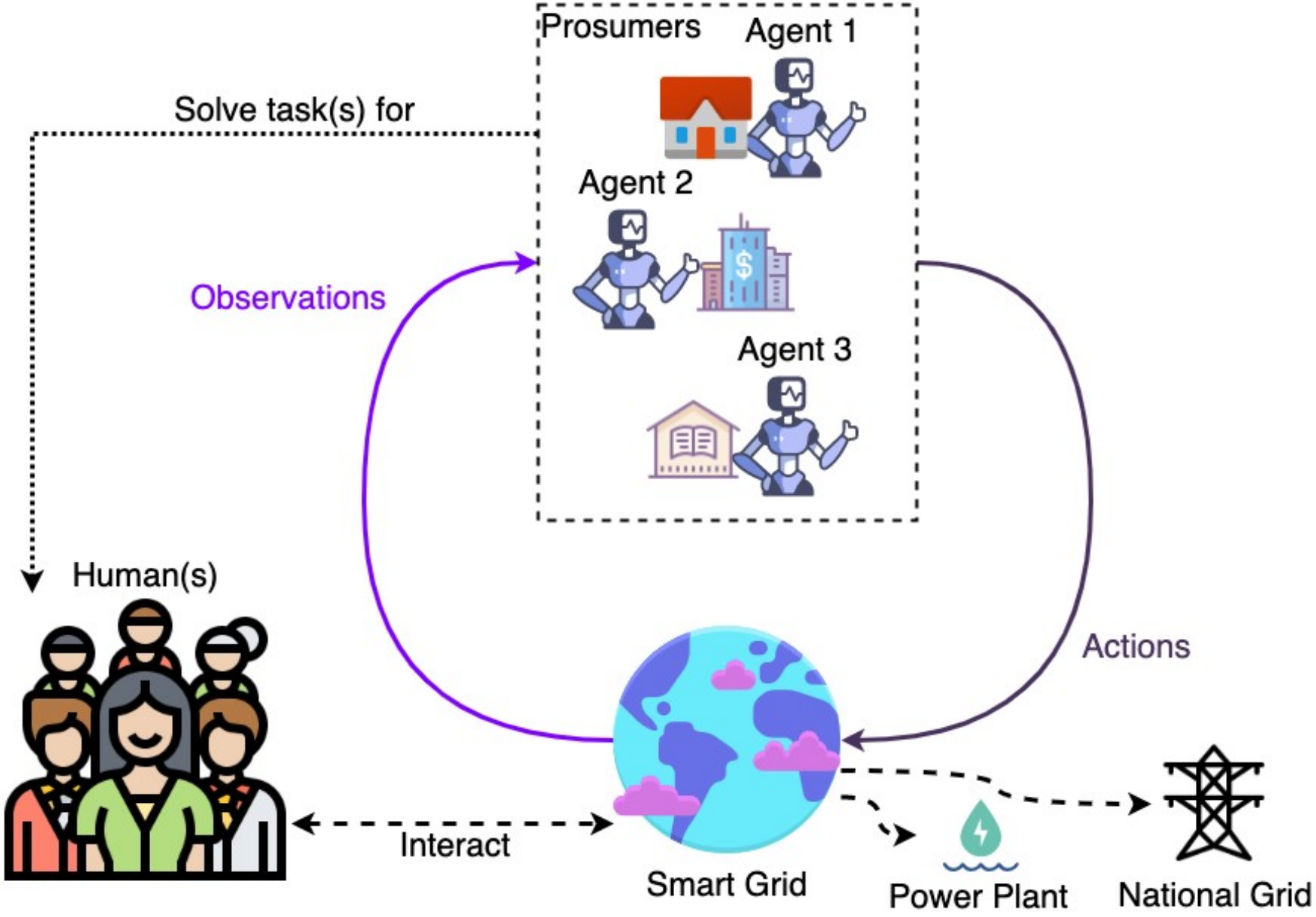
AI systems have an impact over humans



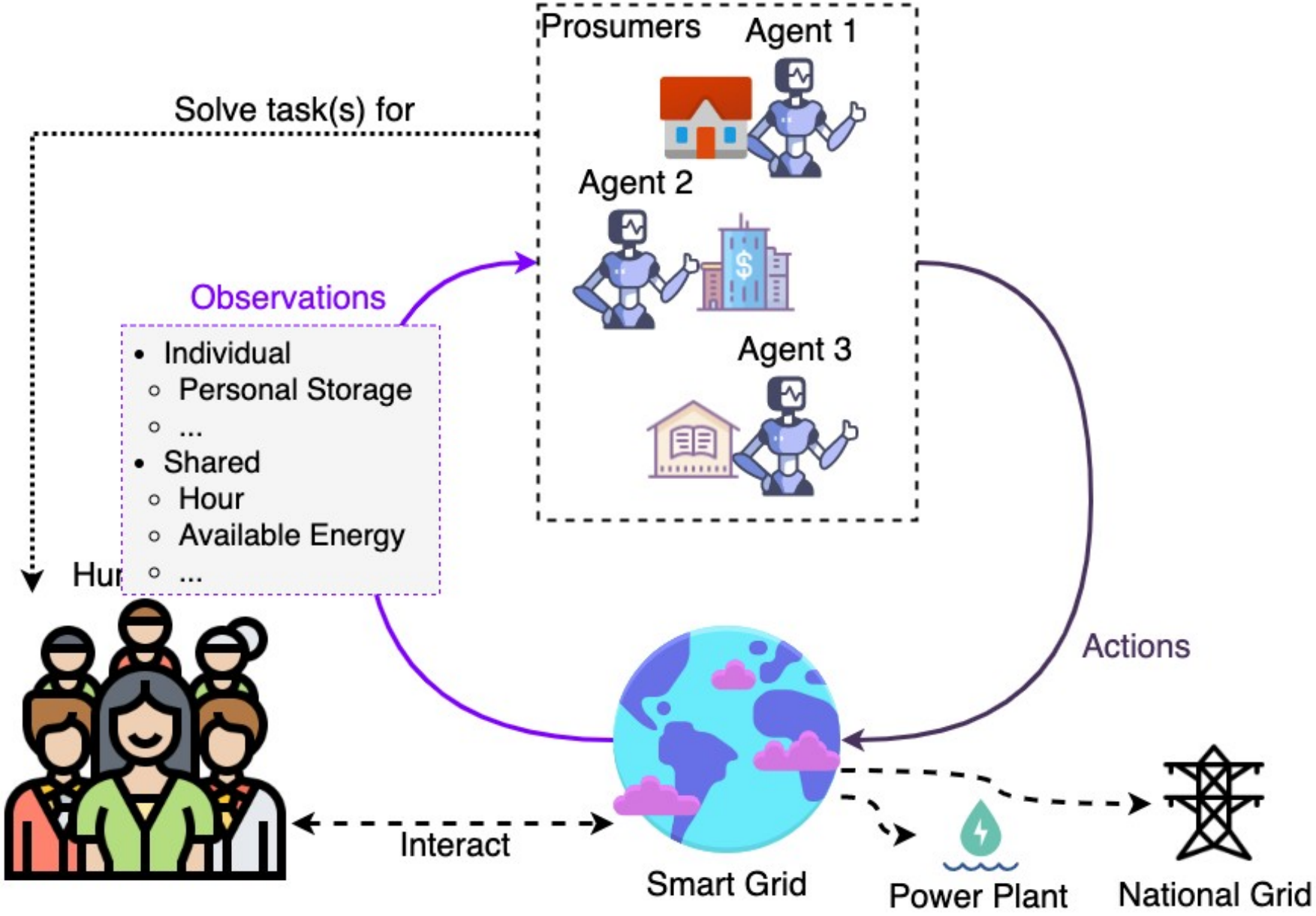
AI systems have an impact over humans



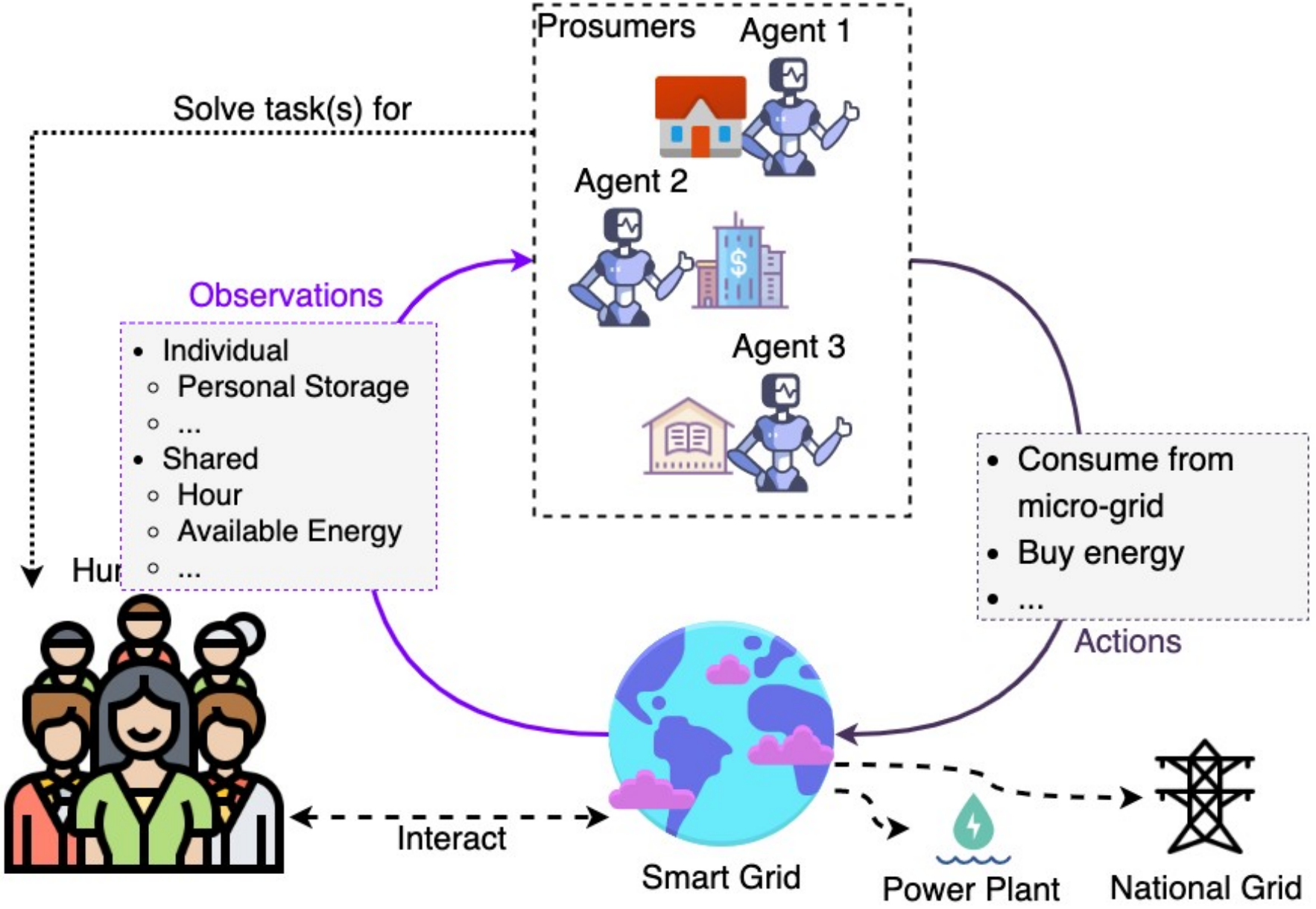
Application to Smart Grids



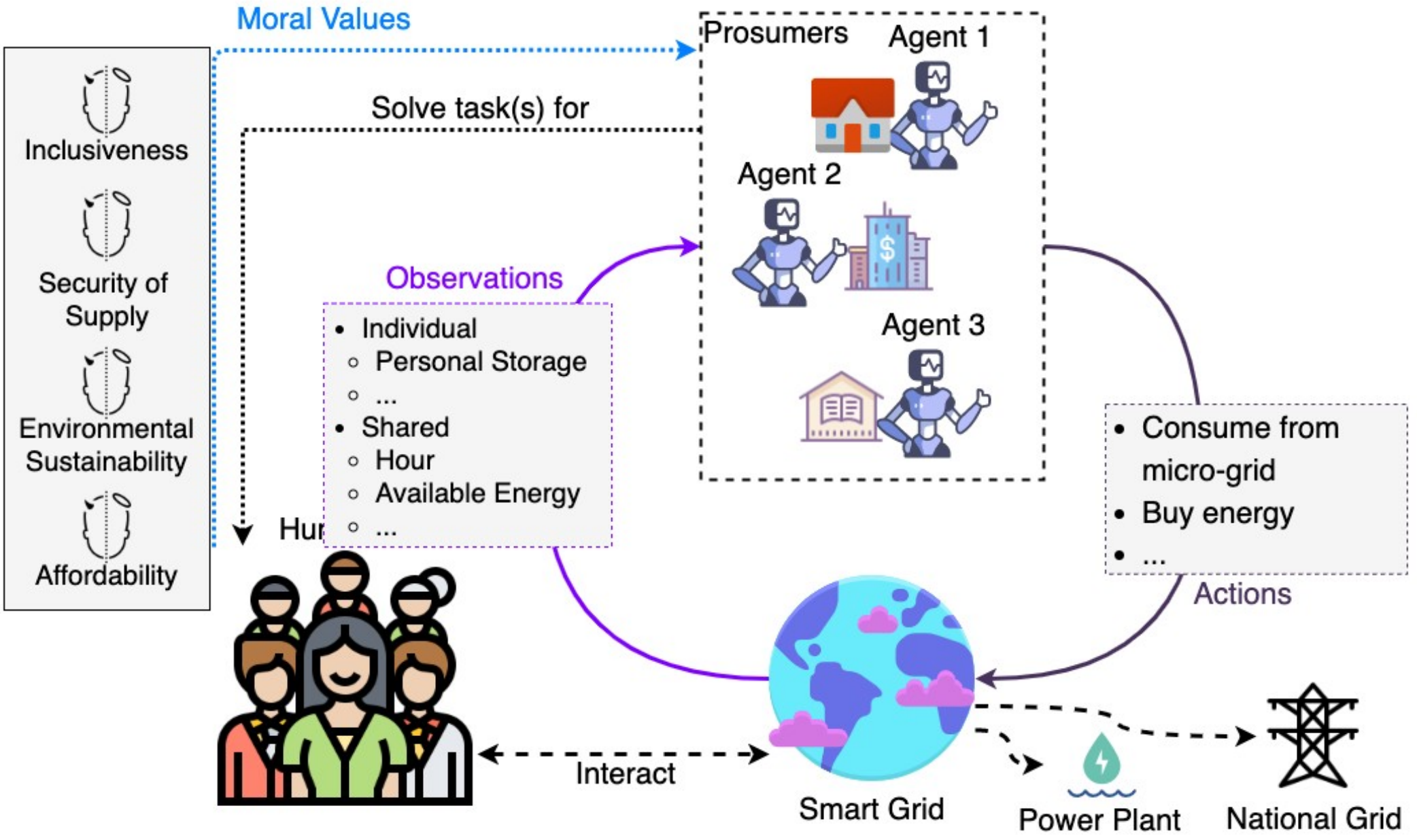
Application to Smart Grids



Application to Smart Grids



Application to Smart Grids



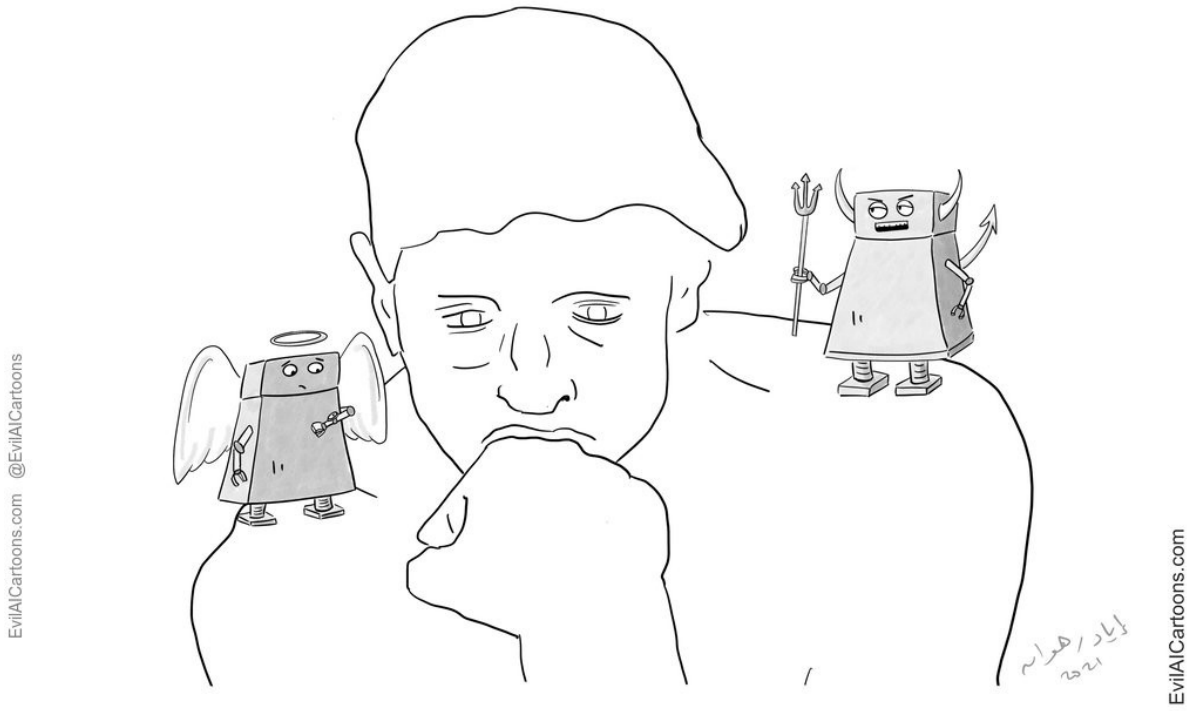
General objective: Learn behaviours morally-aligned

Learning to exhibit behaviours **aligned with our moral values**

General objective: Learn behaviours morally-aligned

Learning to exhibit behaviours **aligned with our moral values**

Ensuring that AI systems' impact is beneficial to humans



<< Look, man, we haven't got all day. How about Jim and I just figure it out between ourselves? >>

Machine Ethics

General objective: Learn behaviours morally-aligned

Learning to exhibit behaviours **aligned with our moral values**

Ensuring that AI systems' impact is beneficial to humans

Ethics by Design (Dignum, 2019): implementing algorithmic capabilities

Machine Ethics

State of the Art – Machine Ethics

- Top-Down
  - Formalization of existing ethical principle(s)
  - Examples:
    - (Cointe *et al.*, 2016) **Ethicaa** : Logic rules, multiple agents, priority order over principles, judgment mechanism
    - (Bremner *et al.*, 2019) **Ethical Layer** : Planning module, ensures adequacy of plans with rules
  - (+) Integration of expert knowledge ; Verifiability
  - (-) Difficult to adapt to unexpected or new situations, to changes

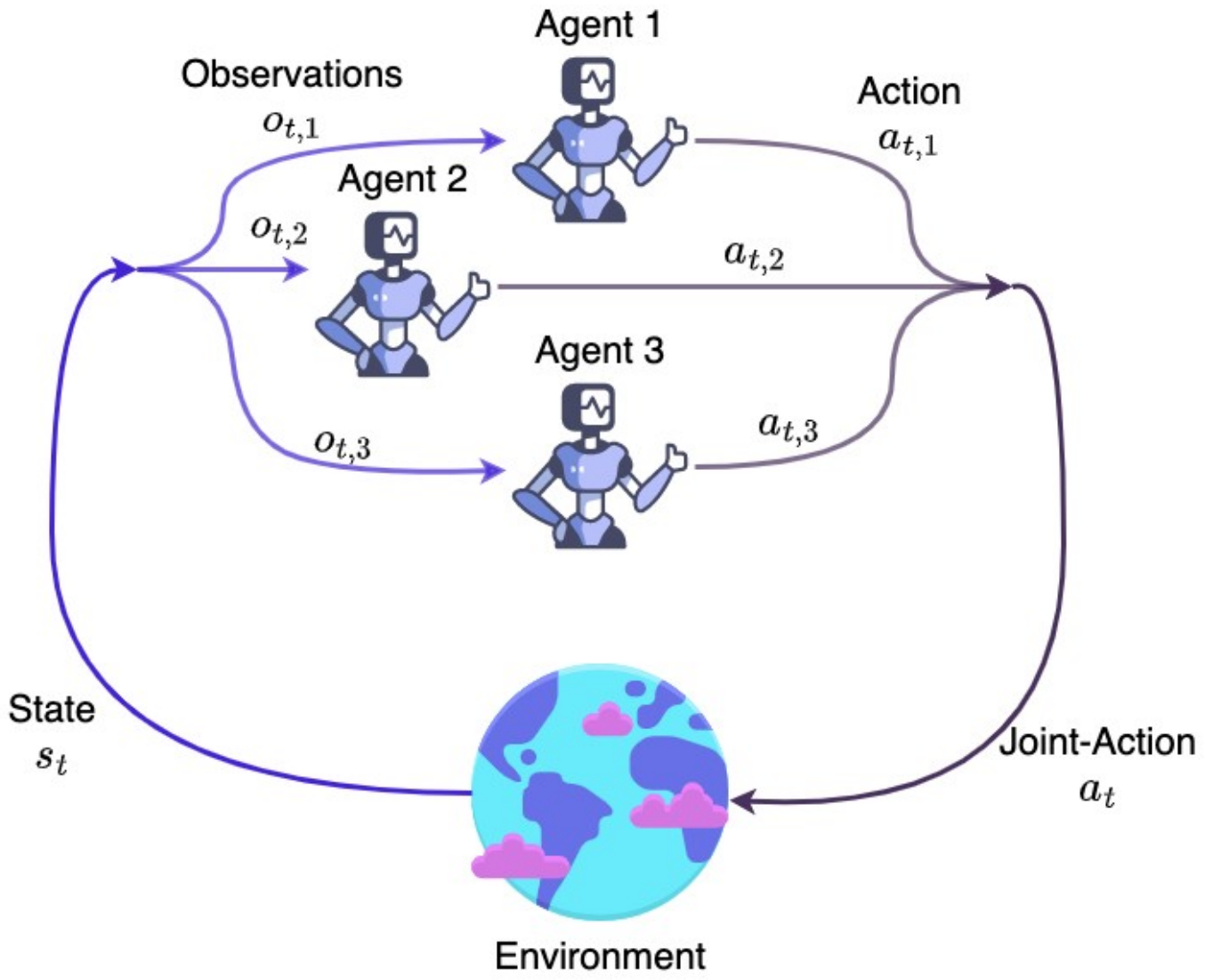
State of the Art – Machine Ethics

- Top-Down
  - Formalization of existing ethical principle(s)
  - Examples:
    - (Cointe *et al.*, 2016) **Ethicaa** : Logic rules, multiple agents, priority order over principles, judgment mechanism
    - (Bremner *et al.*, 2019) **Ethical Layer** : Planning module, ensures adequacy of plans with rules
  - (+) Integration of expert knowledge ; Verifiability
  - (-) Difficult to adapt to unexpected or new situations, to changes
- Bottom-Up
  - Learning a new principle from experiences
  - Examples:
    - (Anderson *et al.*, 2019) **GenEth** : Inductive Logic Programming, ethicists' decisions
    - (Wu *et al.*, 2017) **RL Shaping** : task reward + ethical reward, human "average" behaviour
  - (+) Adaptation
  - (-) Difficult to understand the expected behaviour

State of the Art – Machine Ethics

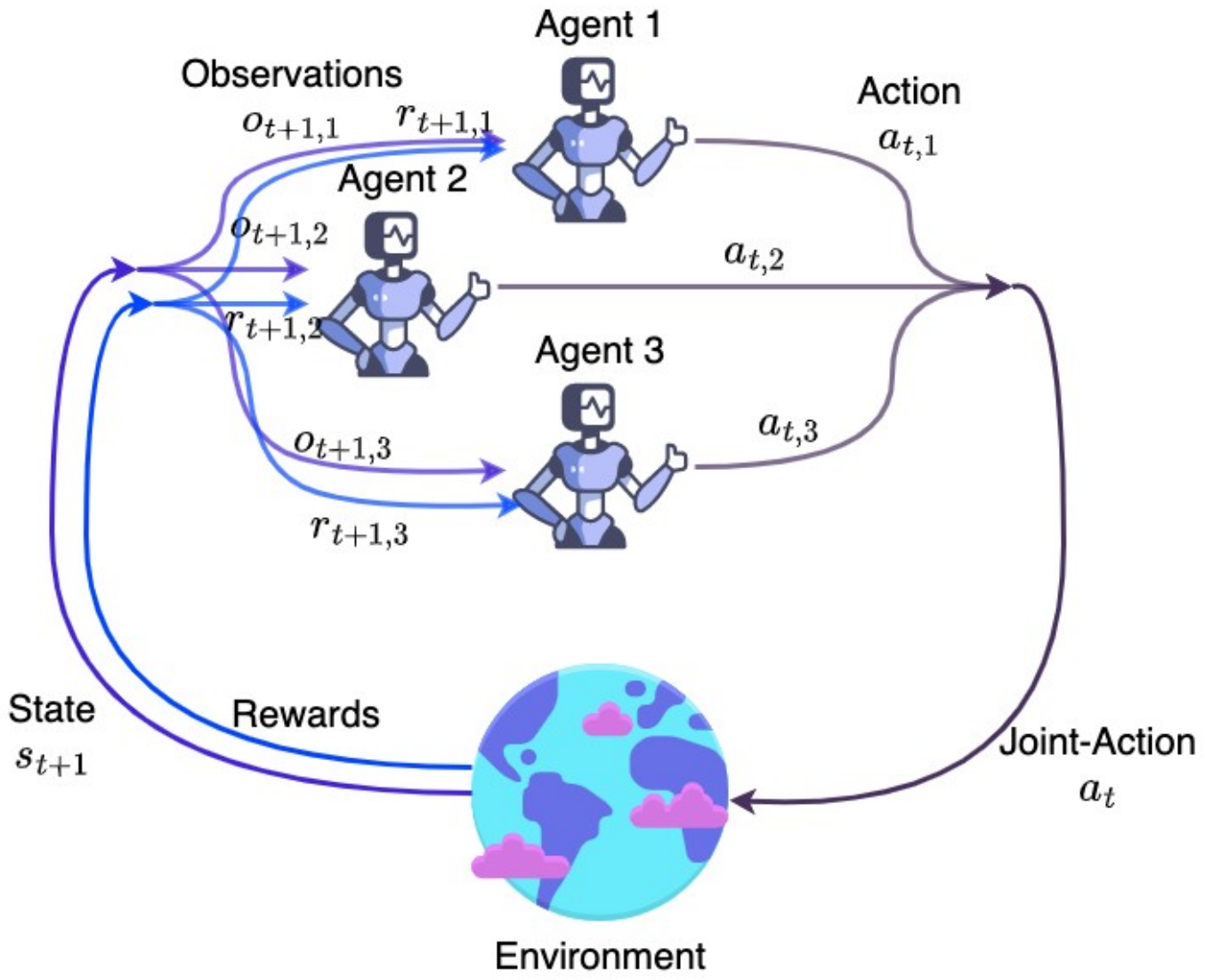
- Top-Down
  - Formalization of existing ethical principle(s)
  - Examples:
    - (Cointe *et al.*, 2016) **Ethicaa** : Logic rules, multiple agents, priority order over principles, judgment mechanism
    - (Bremner *et al.*, 2019) **Ethical Layer** : Planning module, ensures adequacy of plans with rules
  - (+) Integration of expert knowledge ; Verifiability
  - (-) Difficult to adapt to unexpected or new situations, to changes
- Bottom-Up
  - Learning a new principle from experiences
  - Examples:
    - (Anderson *et al.*, 2019) **GenEth** : Inductive Logic Programming, ethicists' decisions
    - (Wu *et al.*, 2017) **RL Shaping** : task reward + ethical reward, human "average" behaviour
  - (+) Adaptation
  - (-) Difficult to understand the expected behaviour
- **Hybrid**
  - Combines advantages from both
  - e.g., learning while constraining

Reinforcement Learning (DecPOMDP) to learn behaviours



(Sutton *et al.*, 2018 ; Bernstein *et al.*, 2002)

Reinforcement Learning (DecPOMDP) to learn behaviours



(Sutton *et al.*, 2018 ; Bernstein *et al.*, 2002)

Objectives and research questions

Objectives
Handle complex environments – Multiple persons
Handle complex environments – Multiple values
Handle complex environments – Multiple situations
Adapt to shifting ethical consensus

Objectives and research questions

Objectives
Handle complex environments – Multiple persons
Handle complex environments – Multiple values
Handle complex environments – Multiple situations
Adapt to shifting ethical consensus
Learn behaviours with non-dilemma situations

RQ1) Learning

- How to learn **behaviours aligned with moral values**, in a **complex environment**, and to **adapt** to **changes**?

Objectives and research questions

Objectives
Handle complex environments – Multiple persons
Handle complex environments – Multiple values
Handle complex environments – Multiple situations
Adapt to shifting ethical consensus
Learn behaviours with non-dilemma situations
Specify desired behaviour

RQ1) Learning

- How to learn **behaviours aligned with moral values**, in a **complex environment**, and to **adapt** to **changes**?

RQ2) Judging

- How to **guide the learning** of agents based on **several moral values**?

Objectives and research questions

Objectives
Handle complex environments – Multiple persons
Handle complex environments – Multiple values
Handle complex environments – Multiple situations
Adapt to shifting ethical consensus
Learn behaviours with non-dilemma situations
Specify desired behaviour
Learn behaviours with dilemma situations

RQ1) Learning

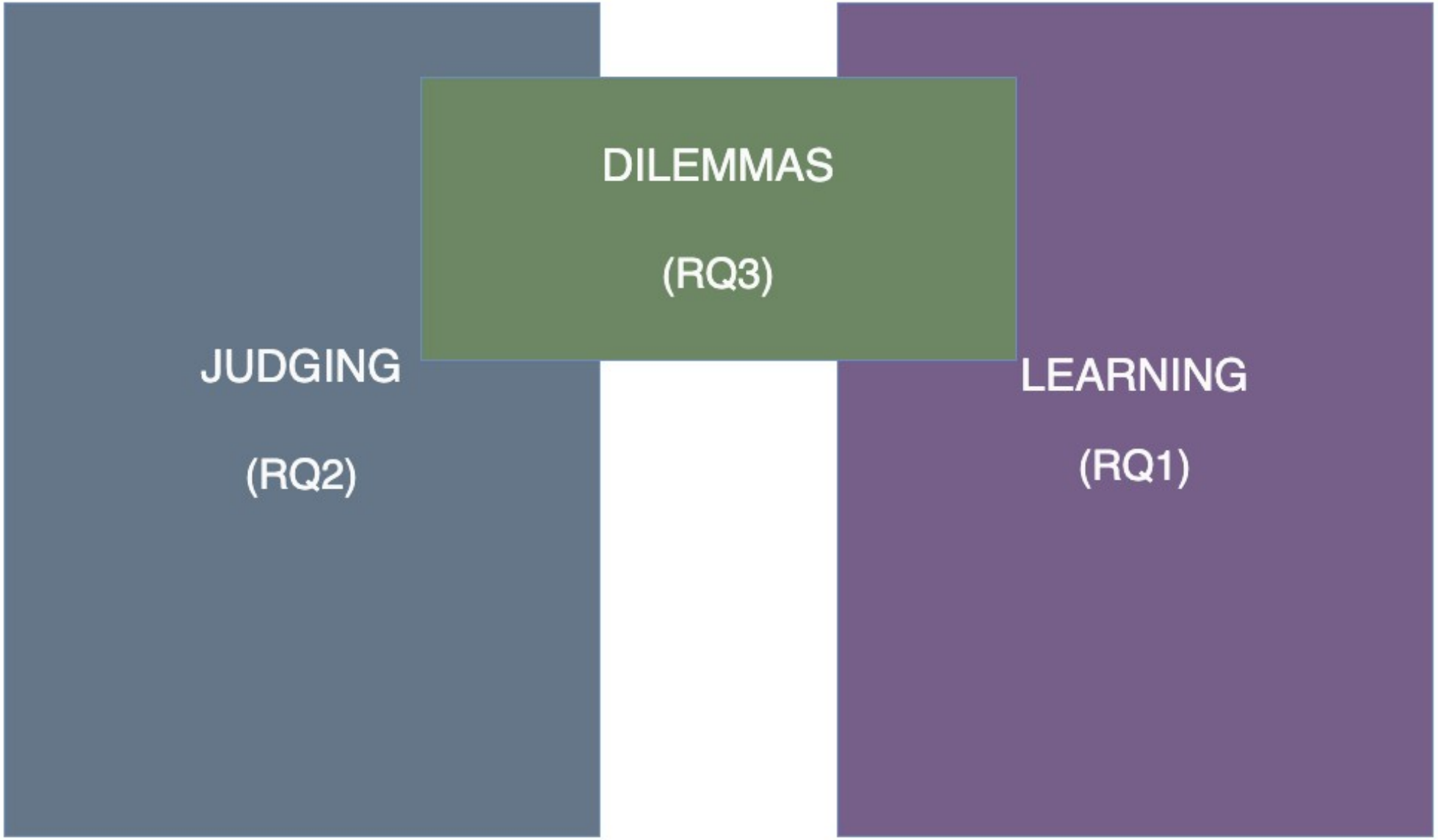
- How to learn **behaviours aligned with moral values**, in a **complex environment**, and to **adapt** to **changes**?

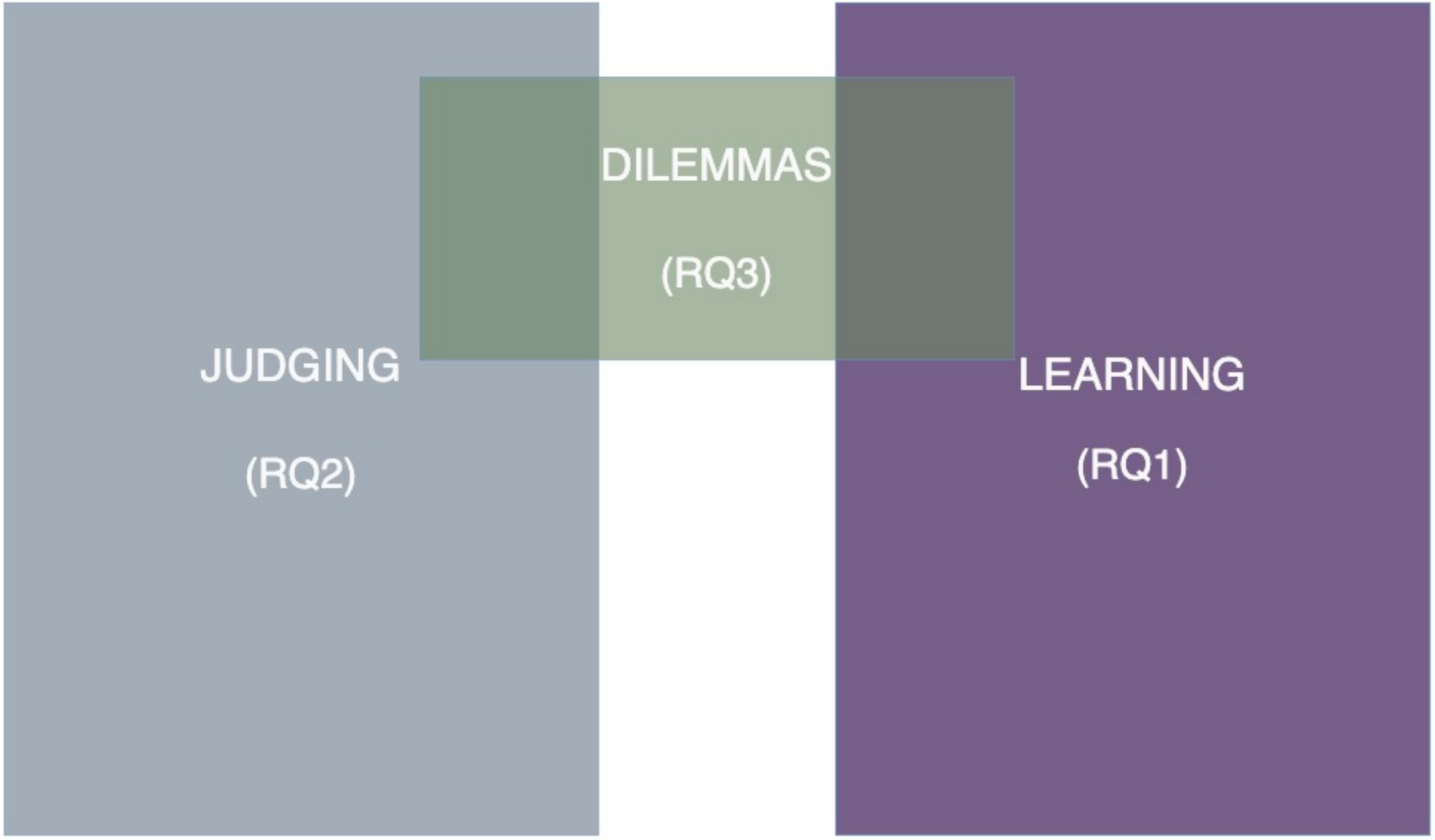
RQ2) Judging

- How to **guide the learning** of agents based on **several moral values**?

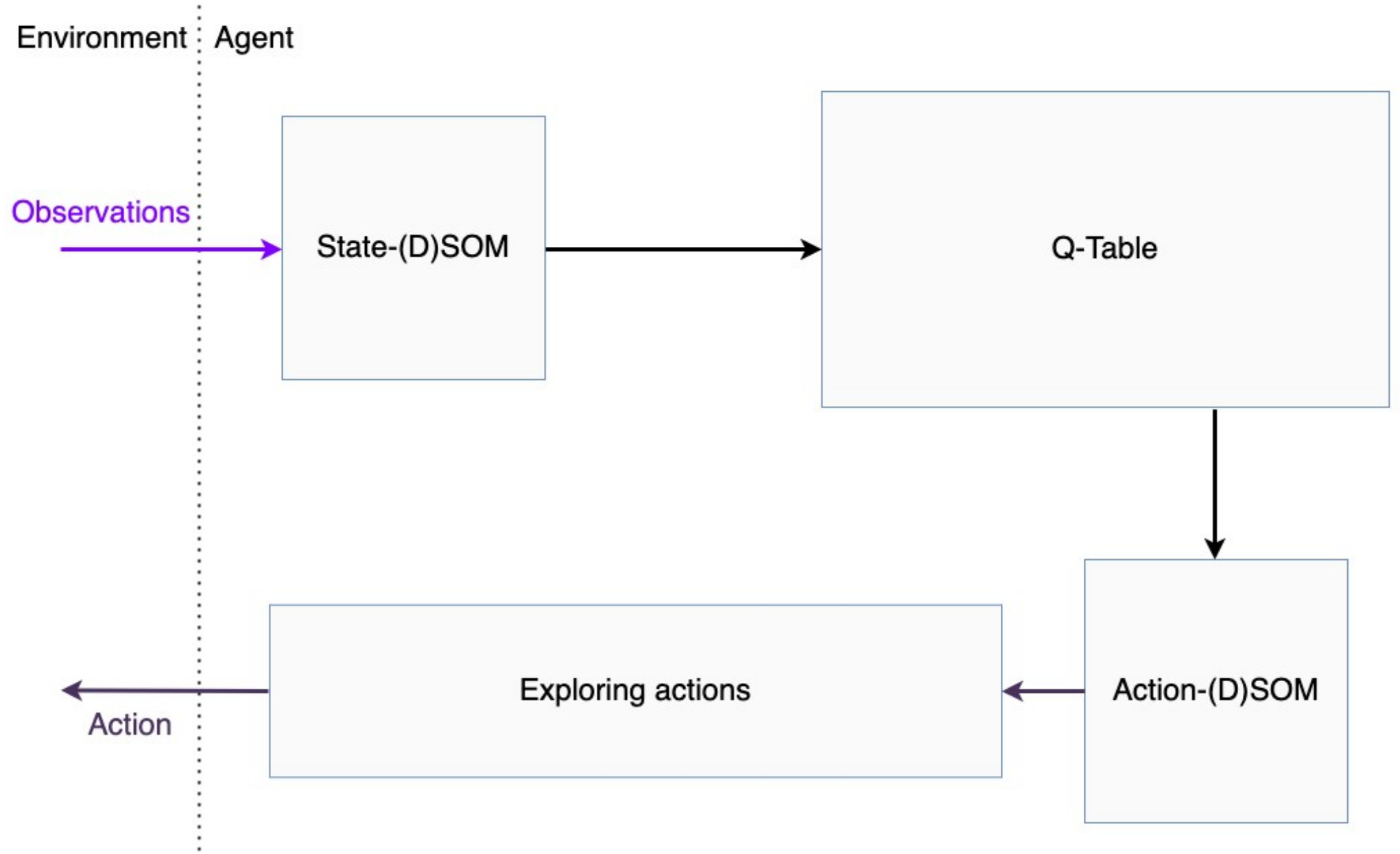
RQ3) Dilemmas

- How to learn to **address dilemmas** in situation **in interaction with human users**?


DILEMMAS
(RQ3)
JUDGING
(RQ2)
LEARNING
(RQ1)


DILEMMAS
(RQ3)
JUDGING
(RQ2)
LEARNING
(RQ1)

Q-(D)SOM : Combining Q-Table and (D)SOMs



Q-(D)SOM : The Q-Table

	Action 1	**Action 2**	**Action 3**	**Action 4**
State 1	$Q_{s1,a1}$	$Q_{s1,a2}$	$Q_{s1,a3}$	$Q_{s1,a4}$
...	...	...	...	...
State 7	$Q_{s7,a1}$	$Q_{s7,a2}$	$Q_{s7,a3}$	$Q_{s7,a4}$
...	...	...	...	...
State 9	$Q_{s9,a1}$	$Q_{s9,a2}$	$Q_{s9,a3}$	$Q_{s9,a4}$

Interest of taking "action 3" in "state 7"

(Watkins *et al.*, 1992)

Q-(D)SOM : The Q-Table

$$(s_t, a_t, s_{t+1}, r_{t+1})$$
$$=$$
$$(s_7, a_3, s_1, 0.8)$$

	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	4	2
...	...	...	...	...
State 9	1	1.5	0.5	0

(Watkins *et al.*, 1992)

Q-(D)SOM : The Q-Table

$(s_t, a_t, s_{t+1}, r_{t+1})$
=
$(s_7, a_3, s_1, 0.8)$

	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	4	2
...	...	...	...	...
State 9	1	1.5	0.5	0

$$Q(s_7, a_3) \leftarrow \alpha \left[r_{t+1} + \gamma max_{a'} Q(s_{t+1}, a') \right] + (1 - \alpha) Q(s_t, a_t)$$

Current learned interest

(Watkins *et al.*, 1992)

Q-(D)SOM : The Q-Table

$(s_t, a_t, s_{t+1}, r_{t+1})$
=
$(s_7, a_3, s_1, 0.8)$

	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	4	2
...	...	...	...	...
State 9	1	1.5	0.5	0

$$Q(s_7, a_3) \leftarrow \alpha \left[r_{t+1} + \gamma max_{a'} Q(s_{t+1}, a') \right] + (1 - \alpha) 4$$

Reward

(Watkins *et al.*, 1992)

Q-(D)SOM : The Q-Table

$(s_t, a_t, s_{t+1}, r_{t+1})$
=
$(s_7, a_3, s_1, 0.8)$

	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	4	2
...	...	...	...	...
State 9	1	1.5	0.5	0

$$Q(s_7, a_3) \leftarrow \alpha\,[0.8 + \gamma max_{a'} Q(s_{t+1}, a')] + (1-\alpha)4$$

Best possible action in next state

(Watkins *et al.*, 1992)

Q-(D)SOM : The Q-Table

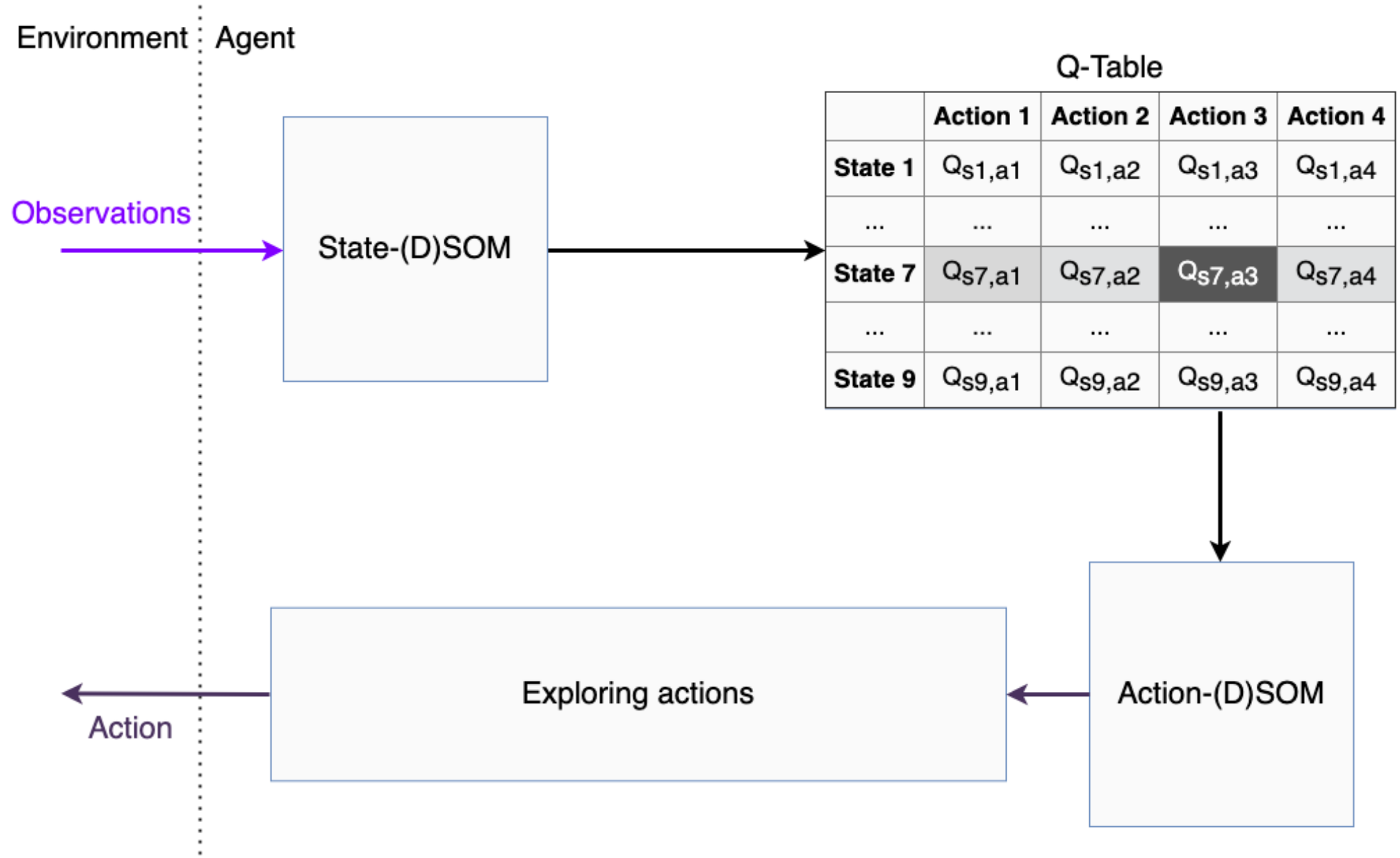
$(s_t, a_t, s_{t+1}, r_{t+1})$
=
$(s_7, a_3, s_1, 0.8)$

	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	**4.78**	2
...	...	...	...	...
State 9	1	1.5	0.5	0

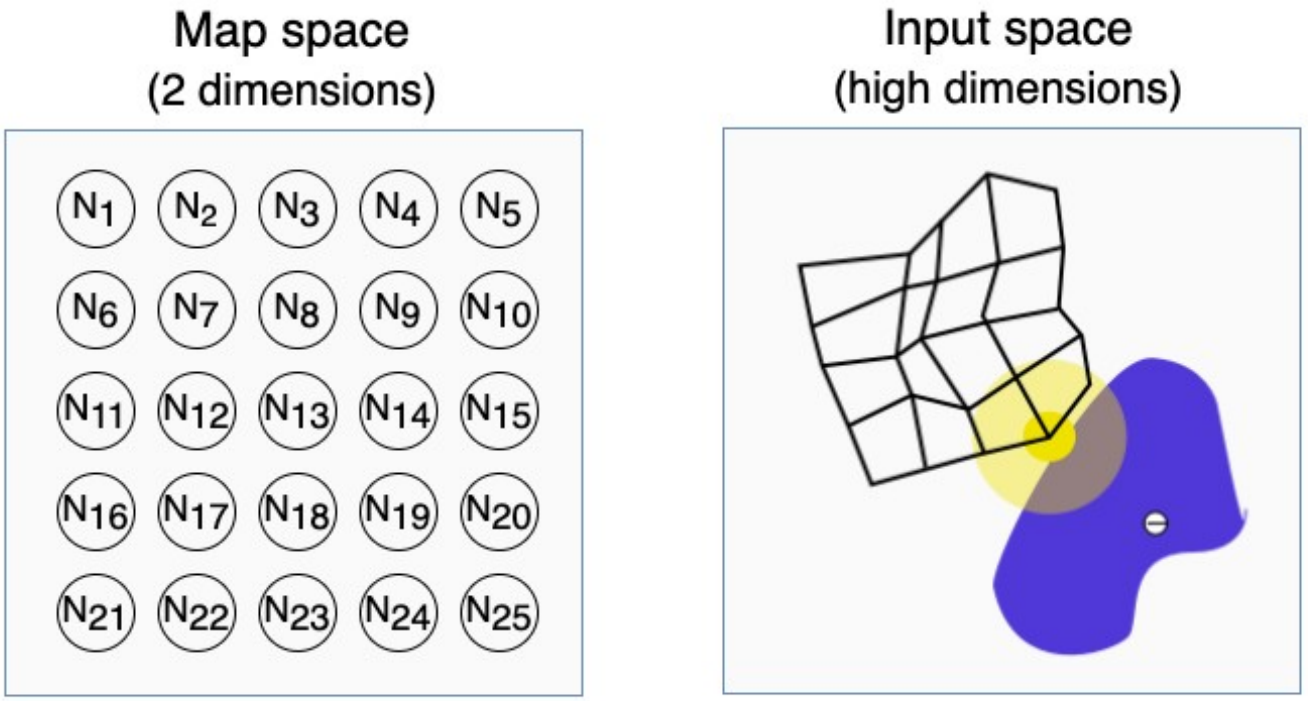
$$Q(s_7, a3) \leftarrow \alpha\,[0.8 + \gamma 5] + (1 - \alpha)4 = 4.78$$

(Watkins *et al.*, 1992)

Q-(D)SOM : Combining Q-Table and (D)SOMs

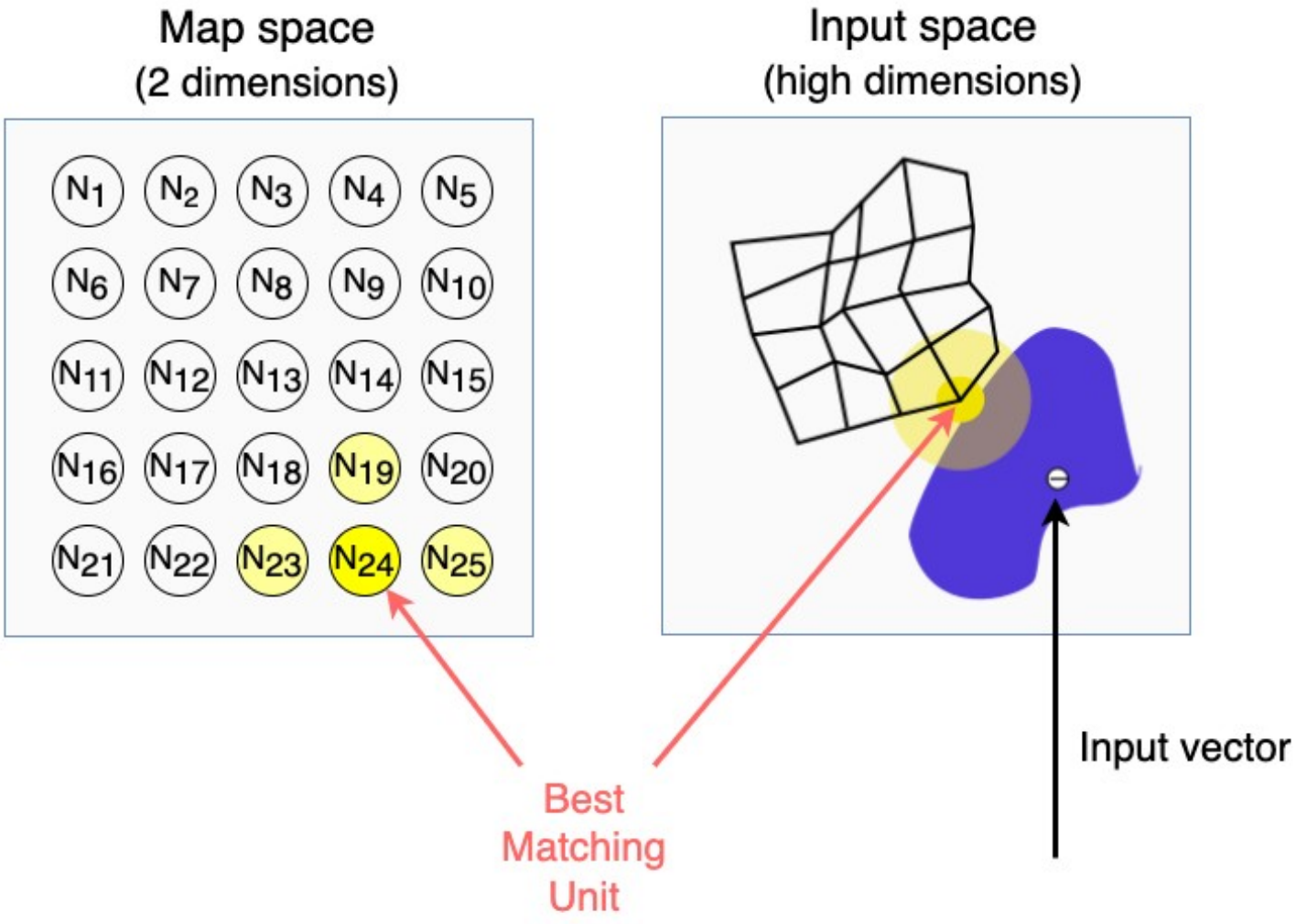


Q-(D)SOM : The Self-Organizing Map



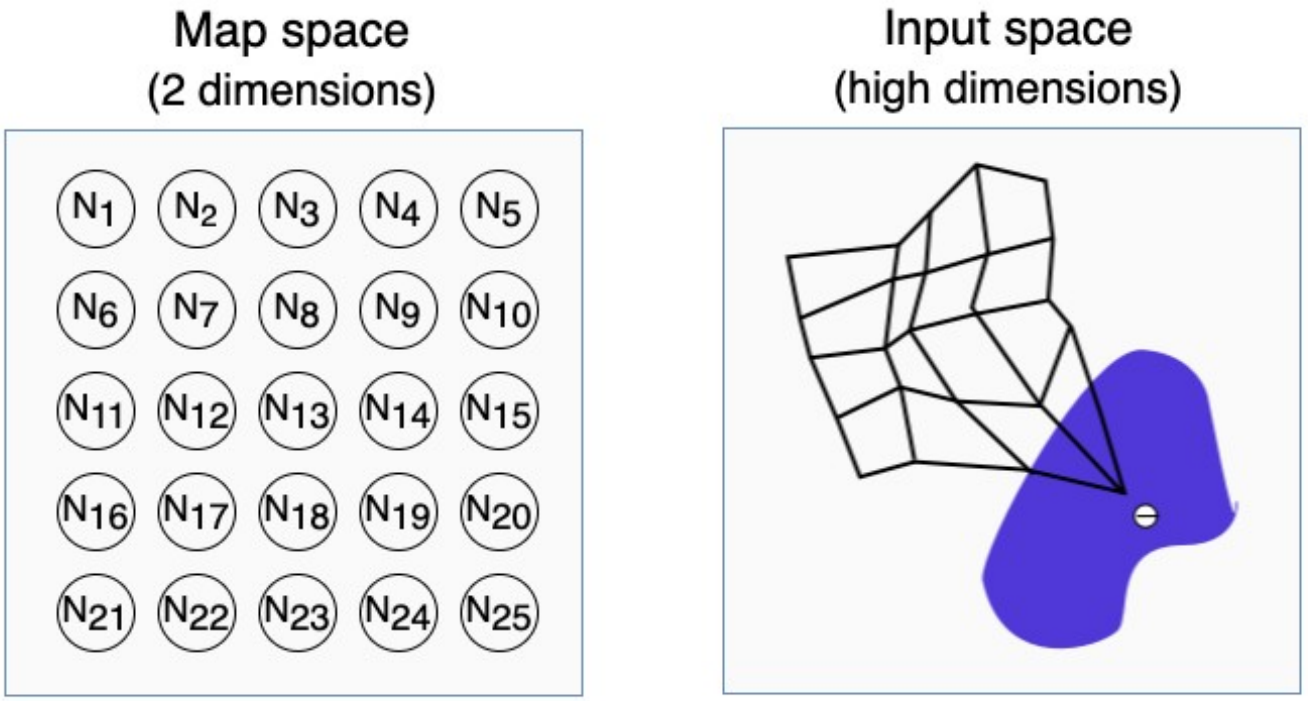
(Kohonen, 1990)

Q-(D)SOM : The Self-Organizing Map



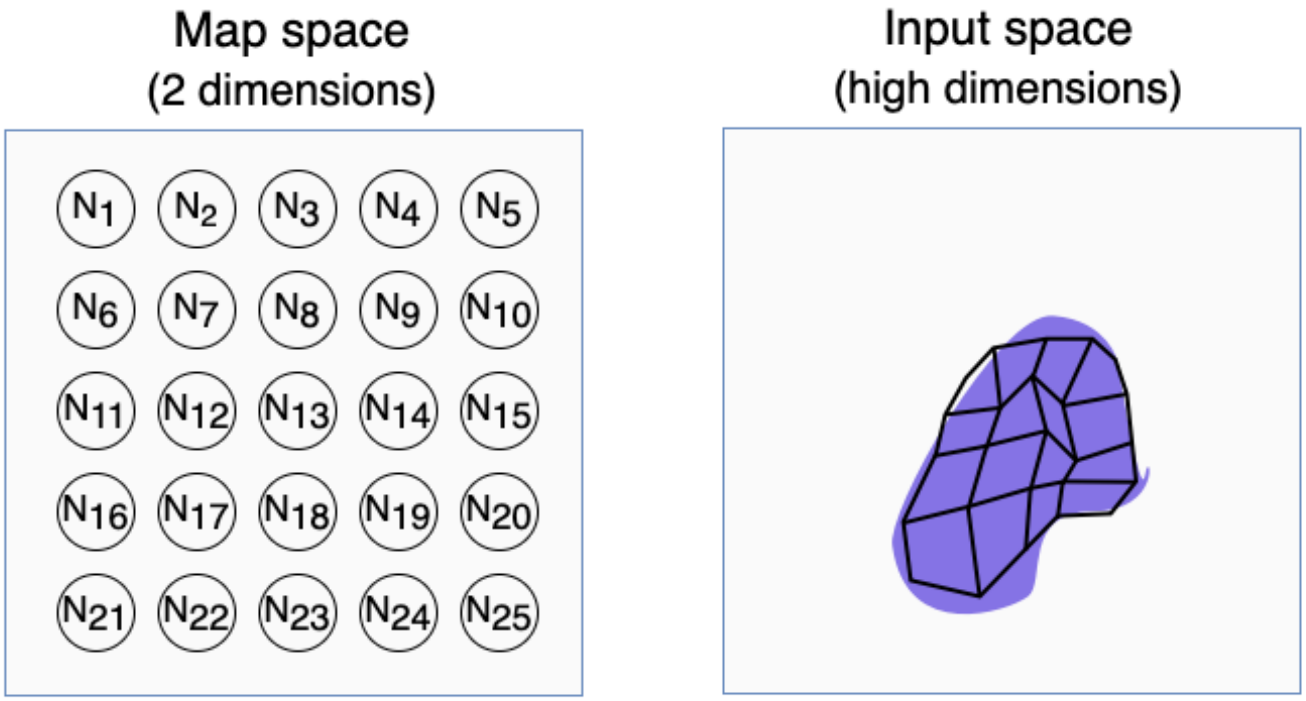
(Kohonen, 1990)

Q-(D)SOM : The Self-Organizing Map



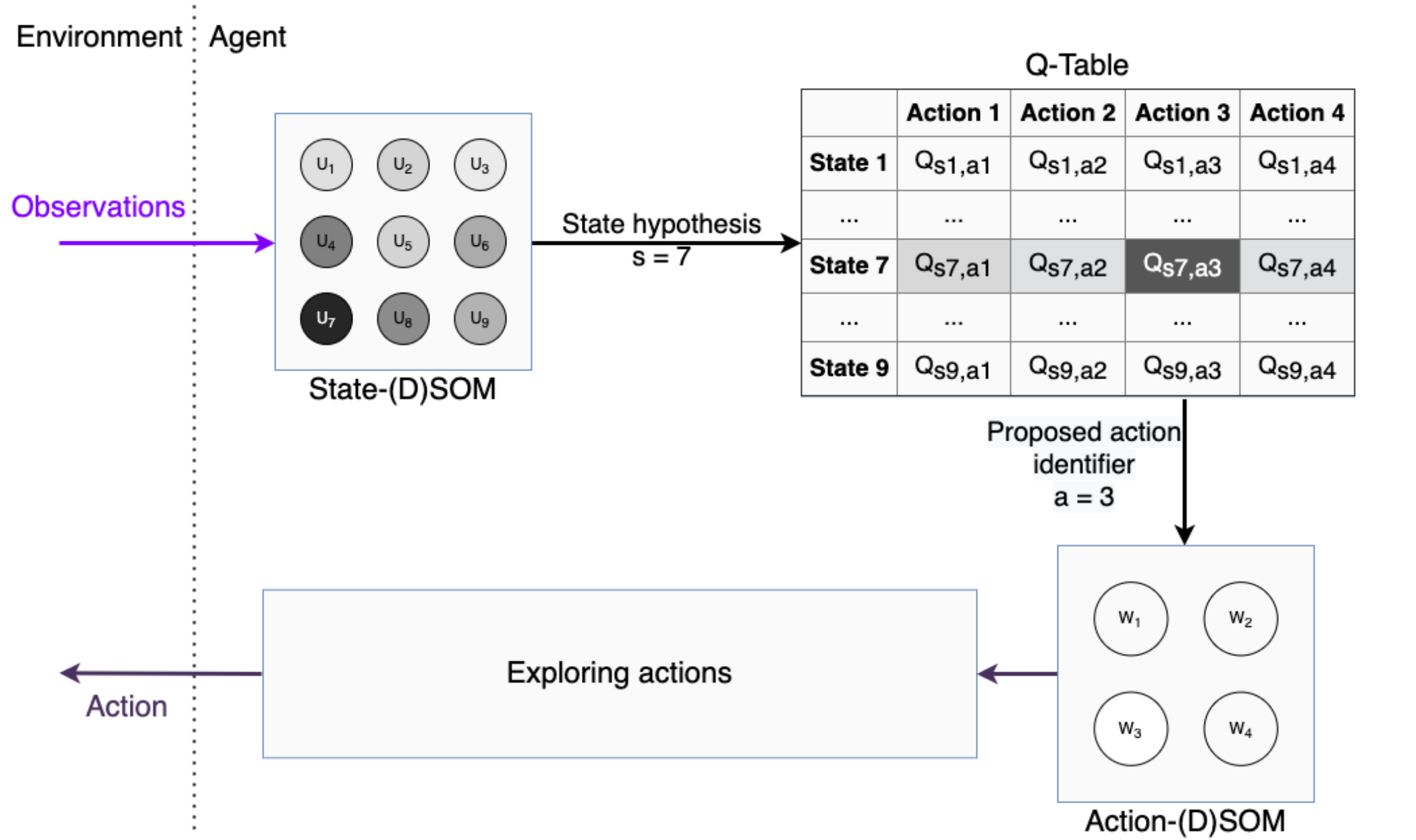
(Kohonen, 1990)

Q-(D)SOM : The Self-Organizing Map

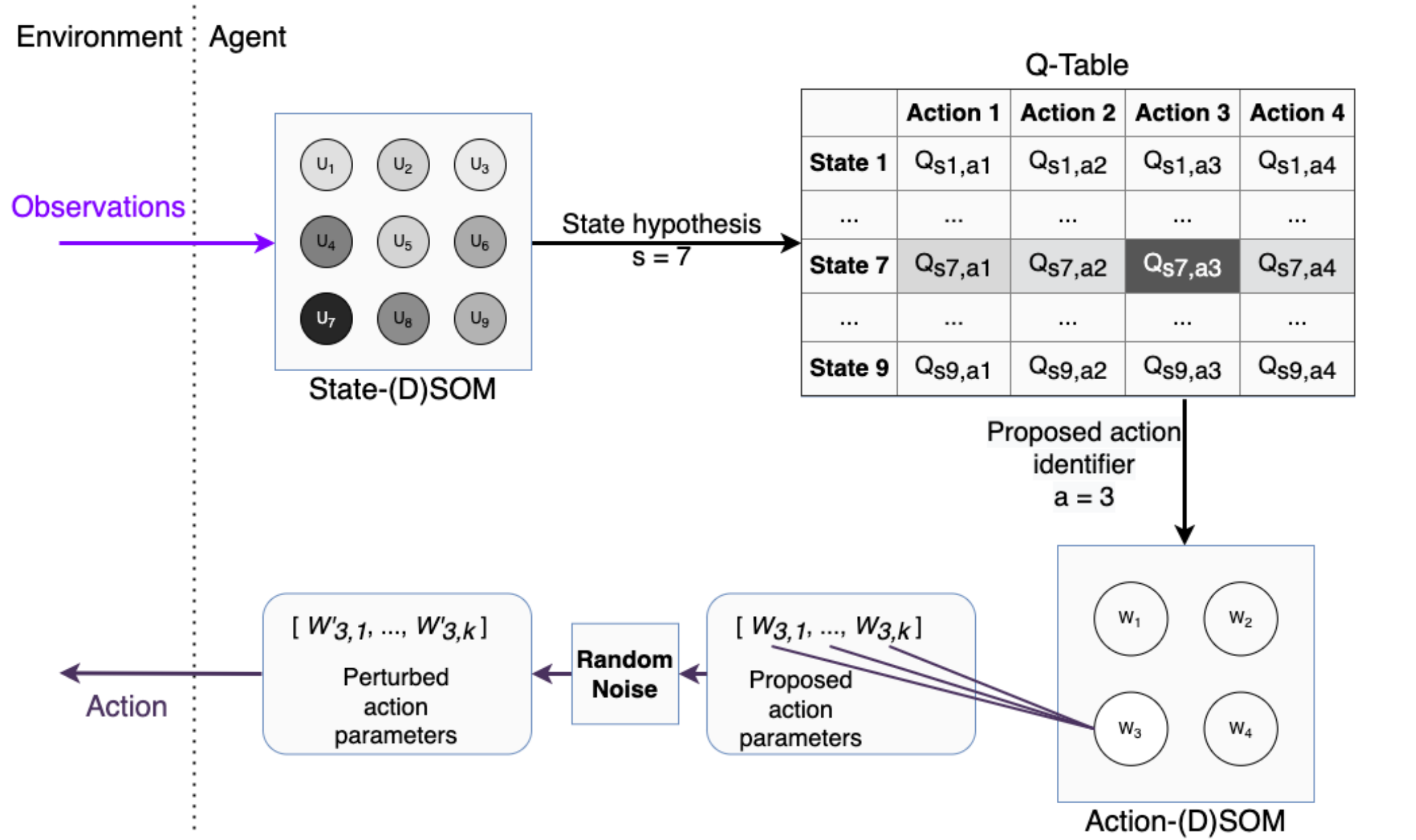


(Kohonen, 1990)

Q-(D)SOM : Combining Q-Table and (D)SOMs



Q-(D)SOM : Combining Q-Table and (D)SOMs



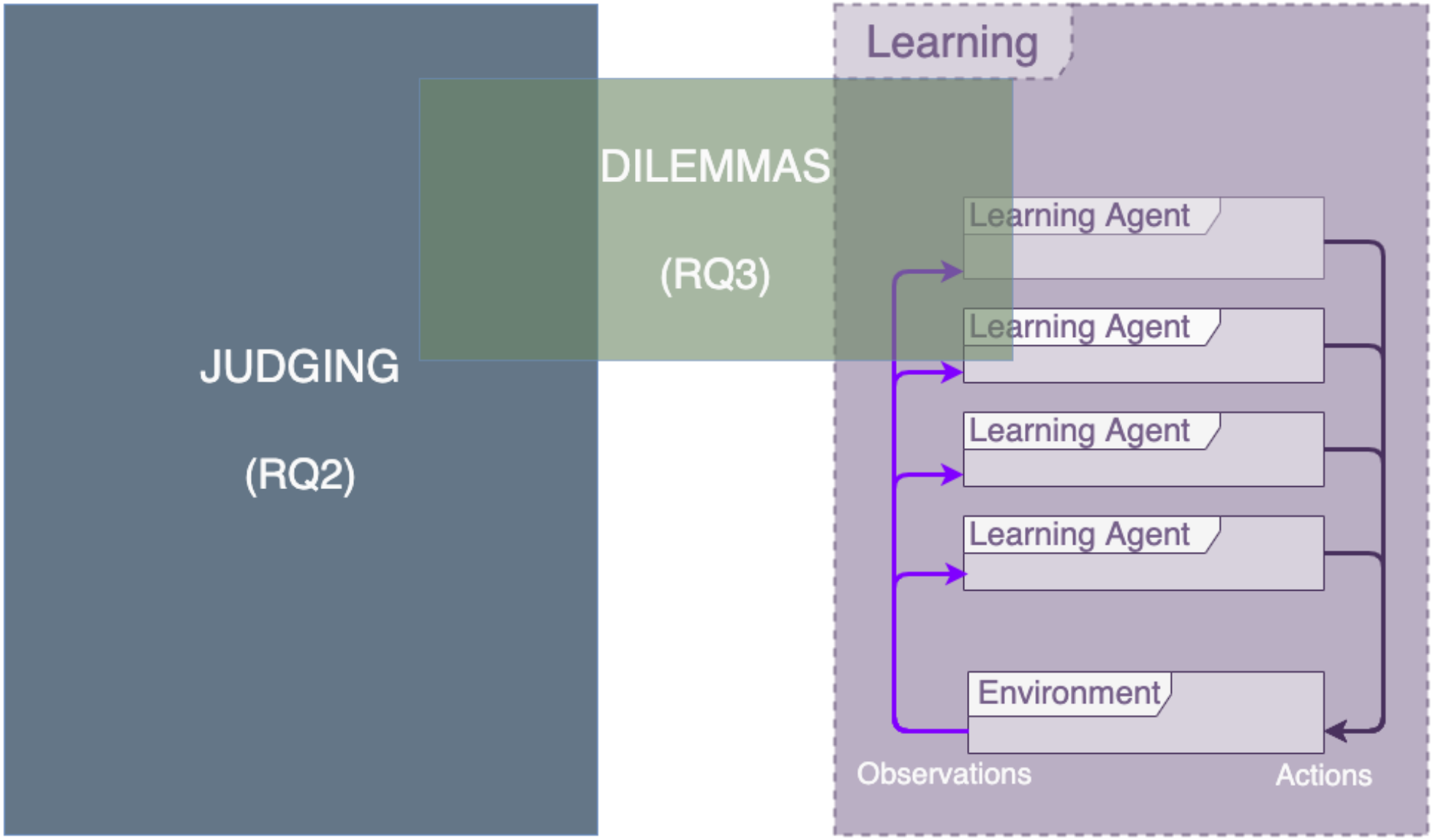
What we have achieved

Objective	Mechanism
Handle complex environments – Multiple persons	• Multi-agents • Individual observations
Handle complex environments – Multiple values	
Handle complex environments – Multiple situations	• (D)SOMs
Adapt to shifting ethical consensus	• Non-convergence • DSOMs
Learn behaviours with non-dilemma situations	• Learning algorithms
Specify desired behaviour	
Learn behaviours with dilemma situations	

What we still lack

Correctly designing the reward function

- Adding, removing, or updating multiple moral values
- Allowing domain experts to participate
- Readability by non-AI experts


JUDGING
(RQ2)
DILEMMAS
(RQ3)
Learning
Learning Agent
Learning Agent
Learning Agent
Learning Agent
Environment
Observations
Actions

Advantages of symbolic reward functions

Numeric reward functions

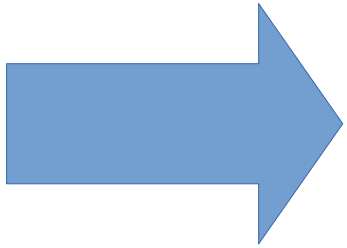
R = 1 - distance

(-) Difficult to integrate expert knowledge
(-) Can be difficult to read or understand for human users

Advantages of symbolic reward functions

Numeric reward functions

R = 1 - distance



(-) Difficult to integrate expert knowledge
(-) Can be difficult to read or understand for human users

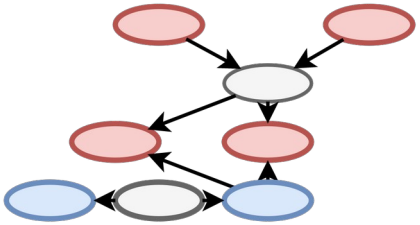
Advantages of symbolic reward functions

Numeric reward functions

R = 1 - distance

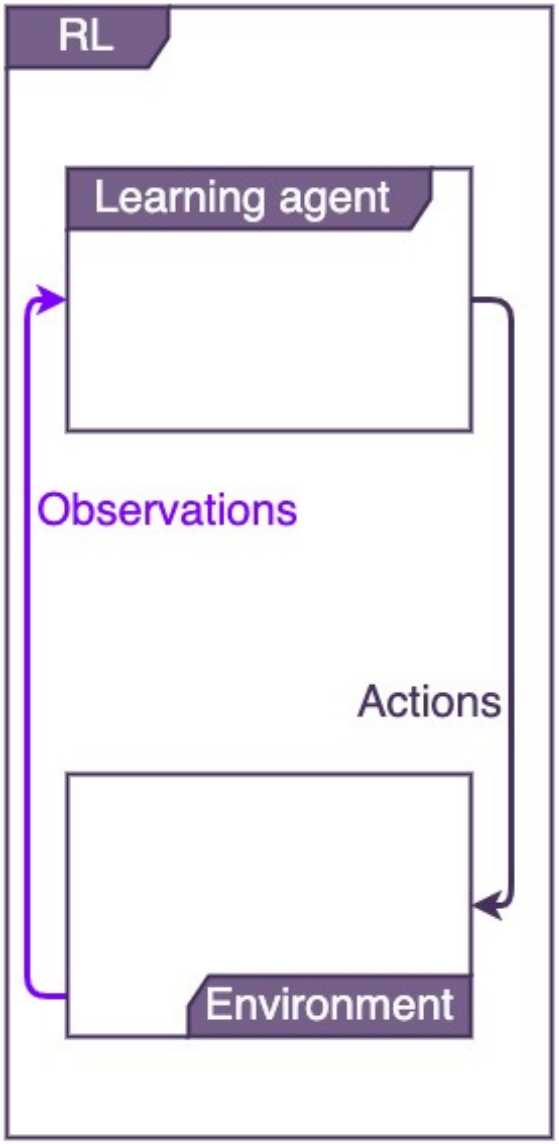
(-) Difficult to integrate expert knowledge
(-) Can be difficult to read or understand for human users

Symbolic judgments for reward functions

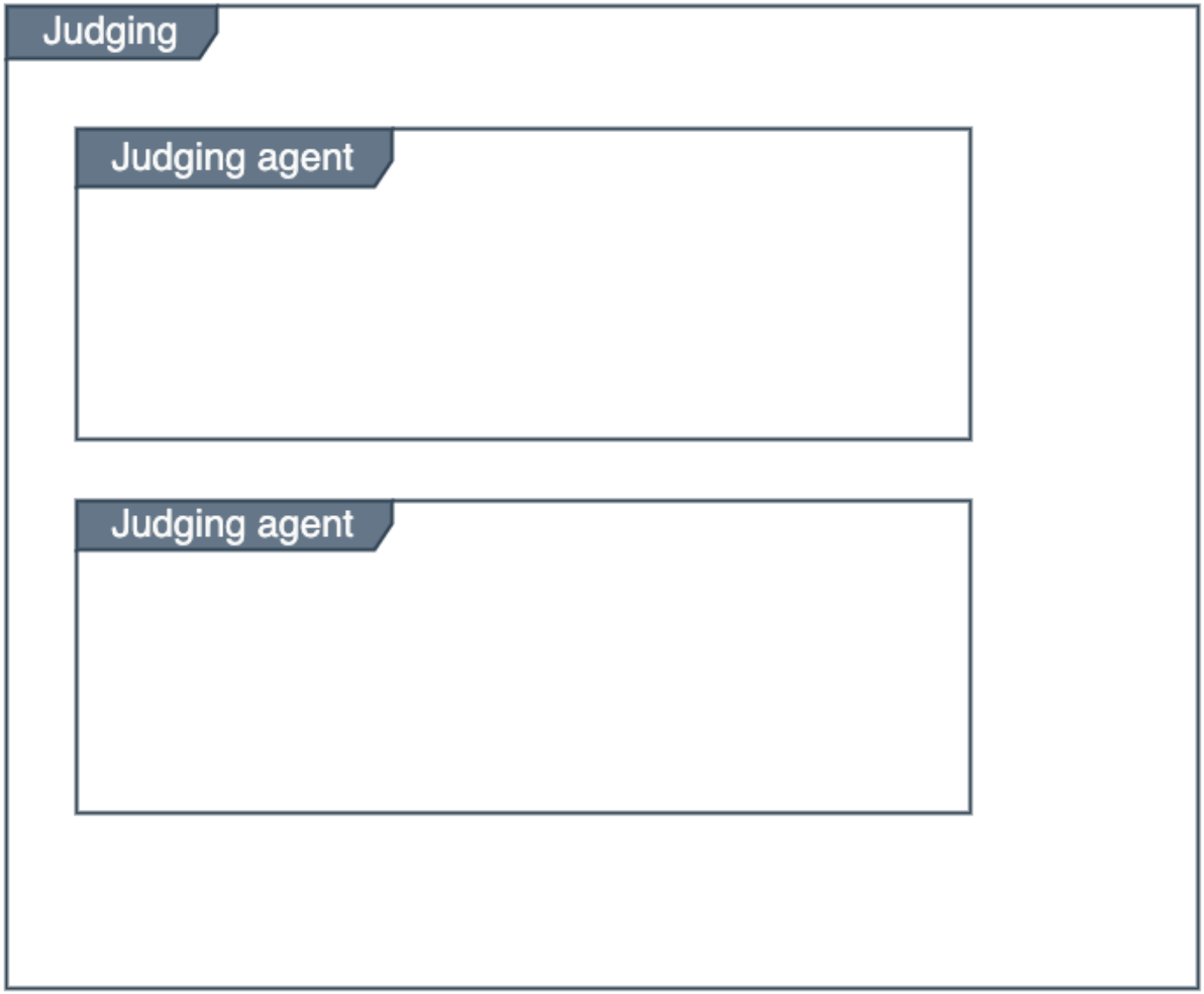
R = 

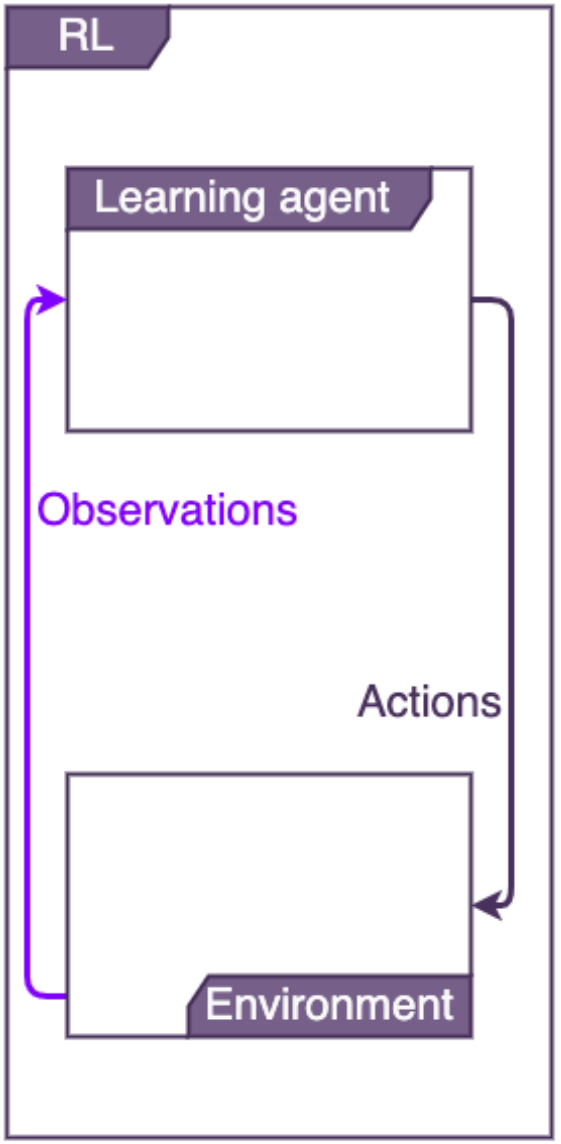
(+) Easier to integrate expert knowledge
(+) Easier to understand or justify rewards
(+) Agentification for complex mechanisms

Judgment architecture

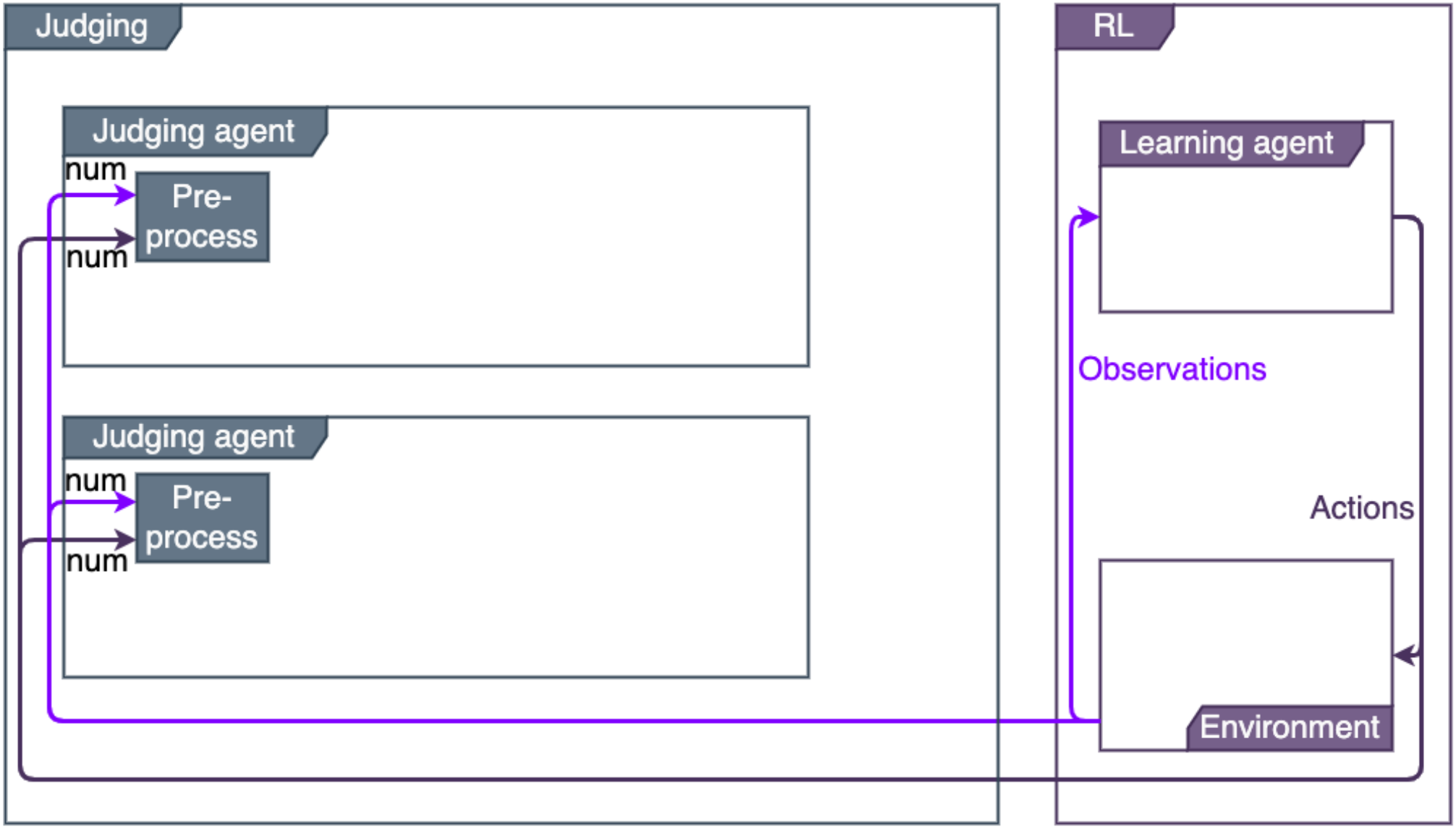


Judgment architecture

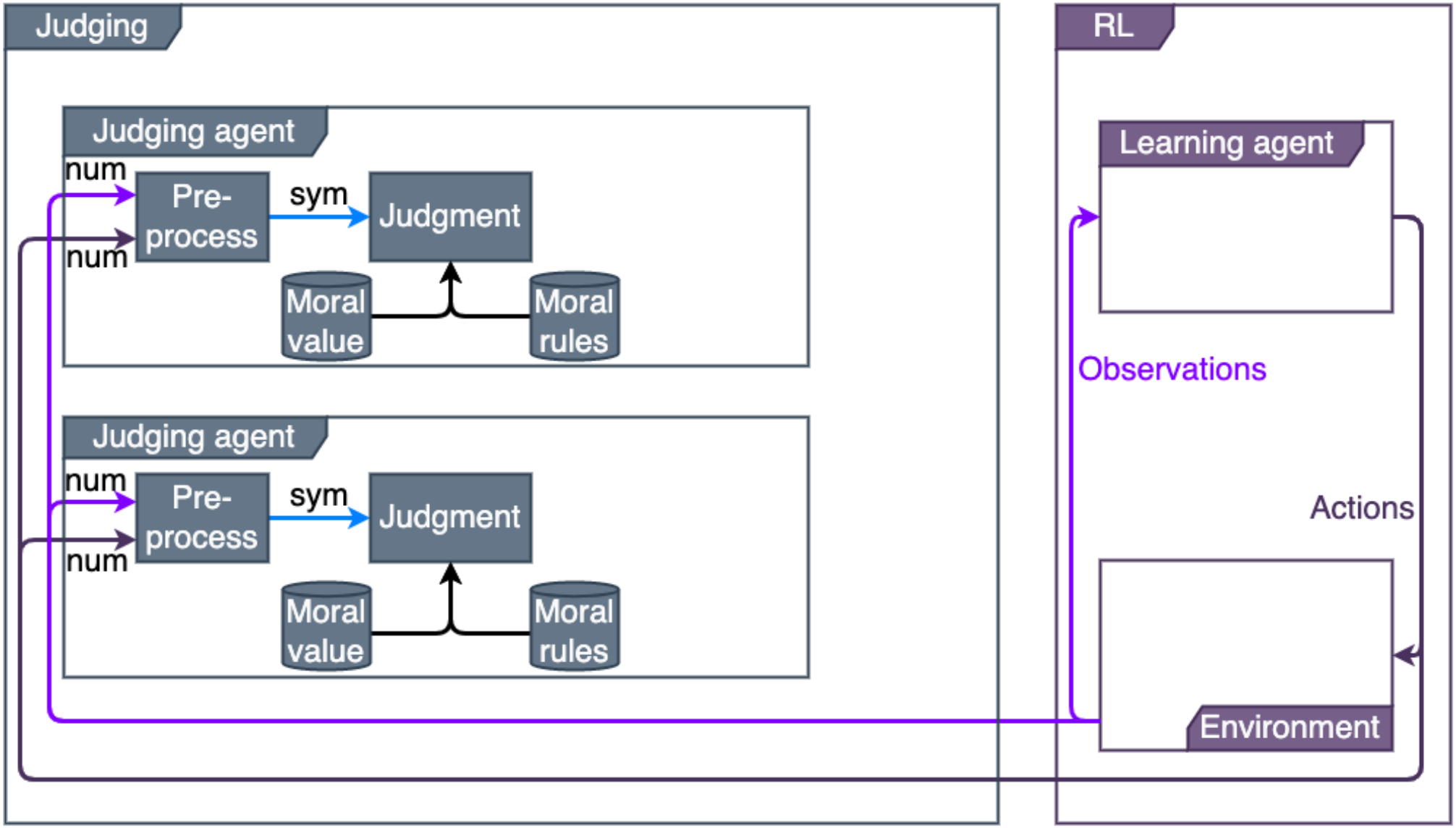




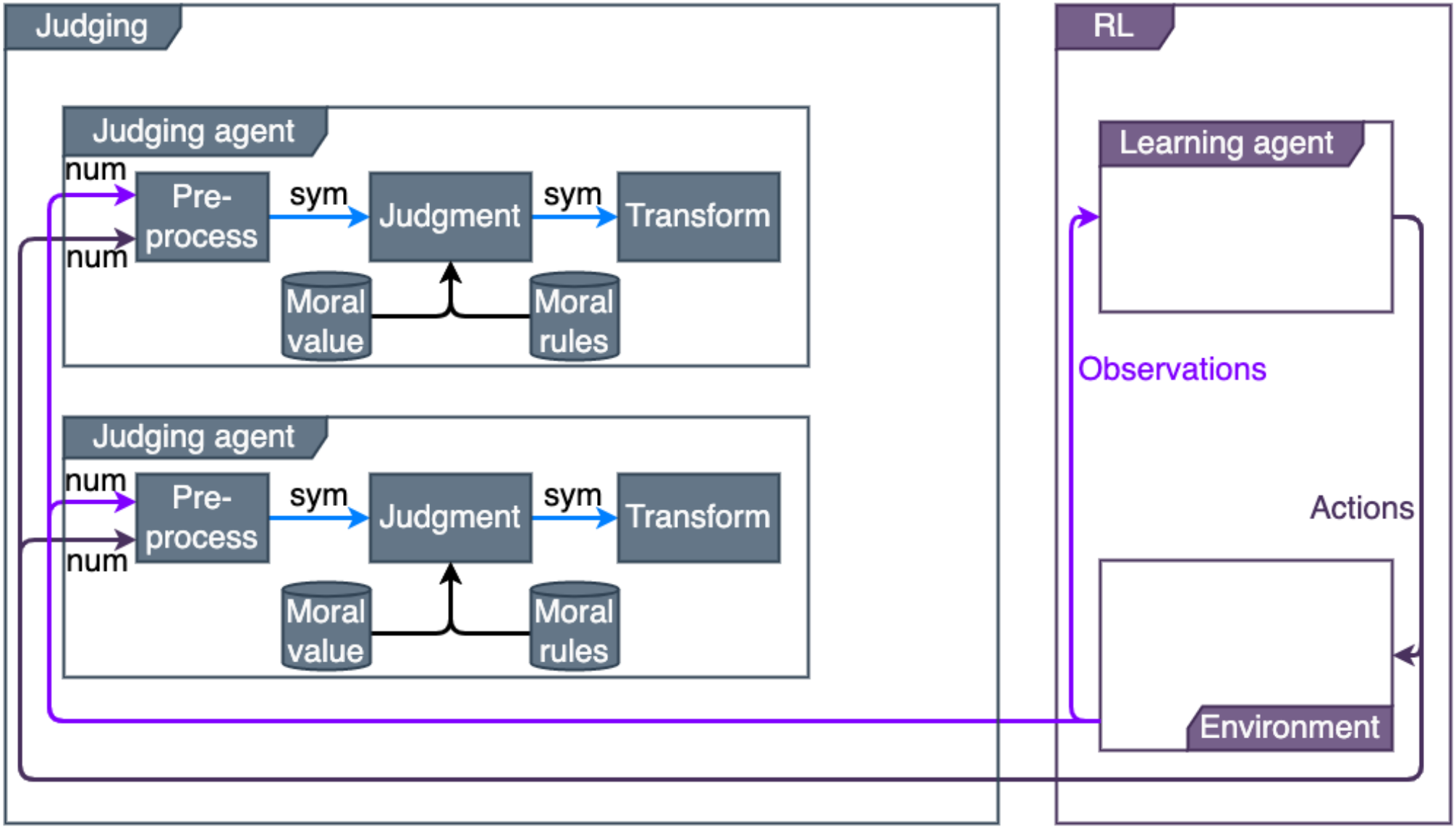
Judgment architecture



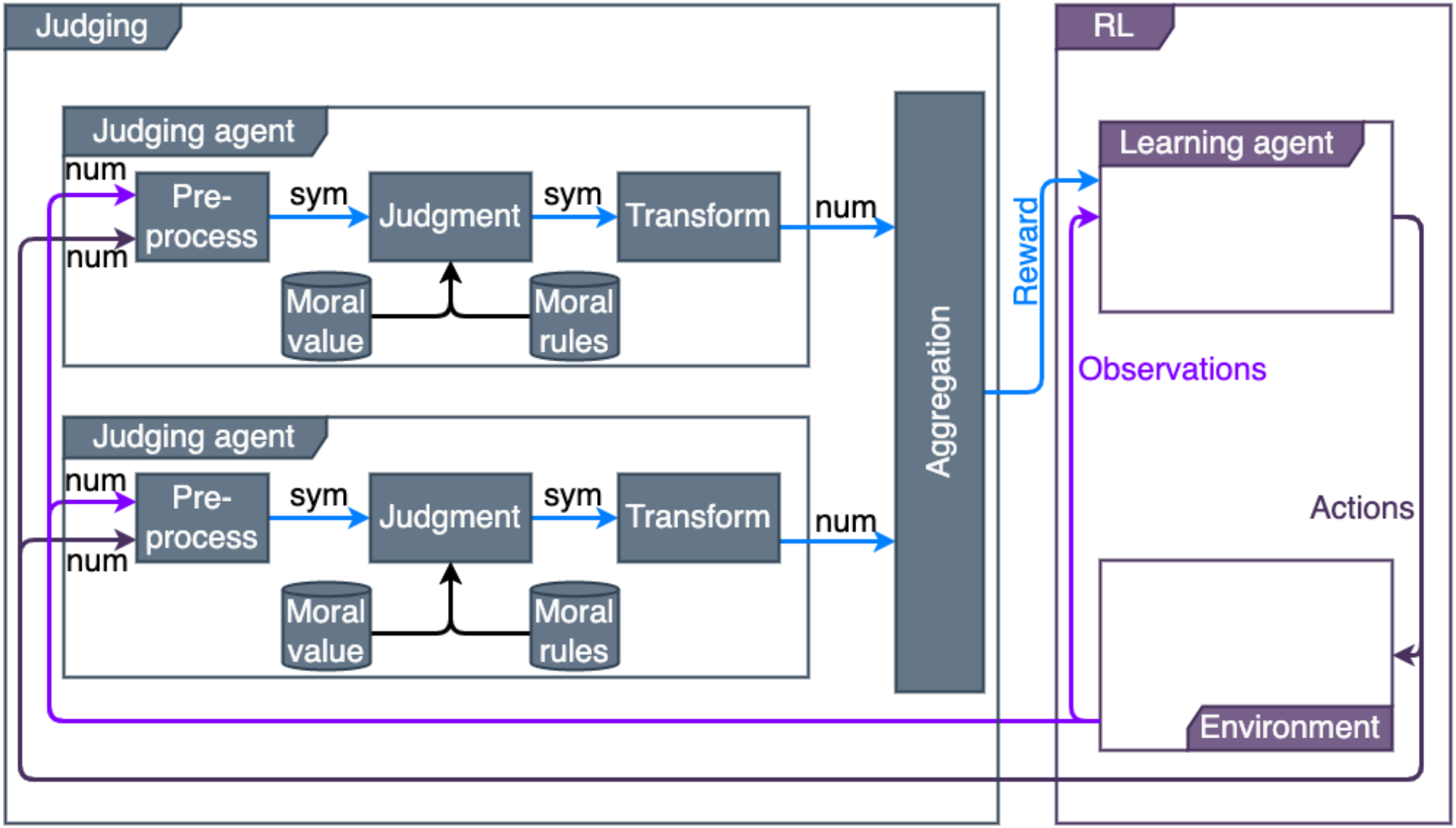
Judgment architecture



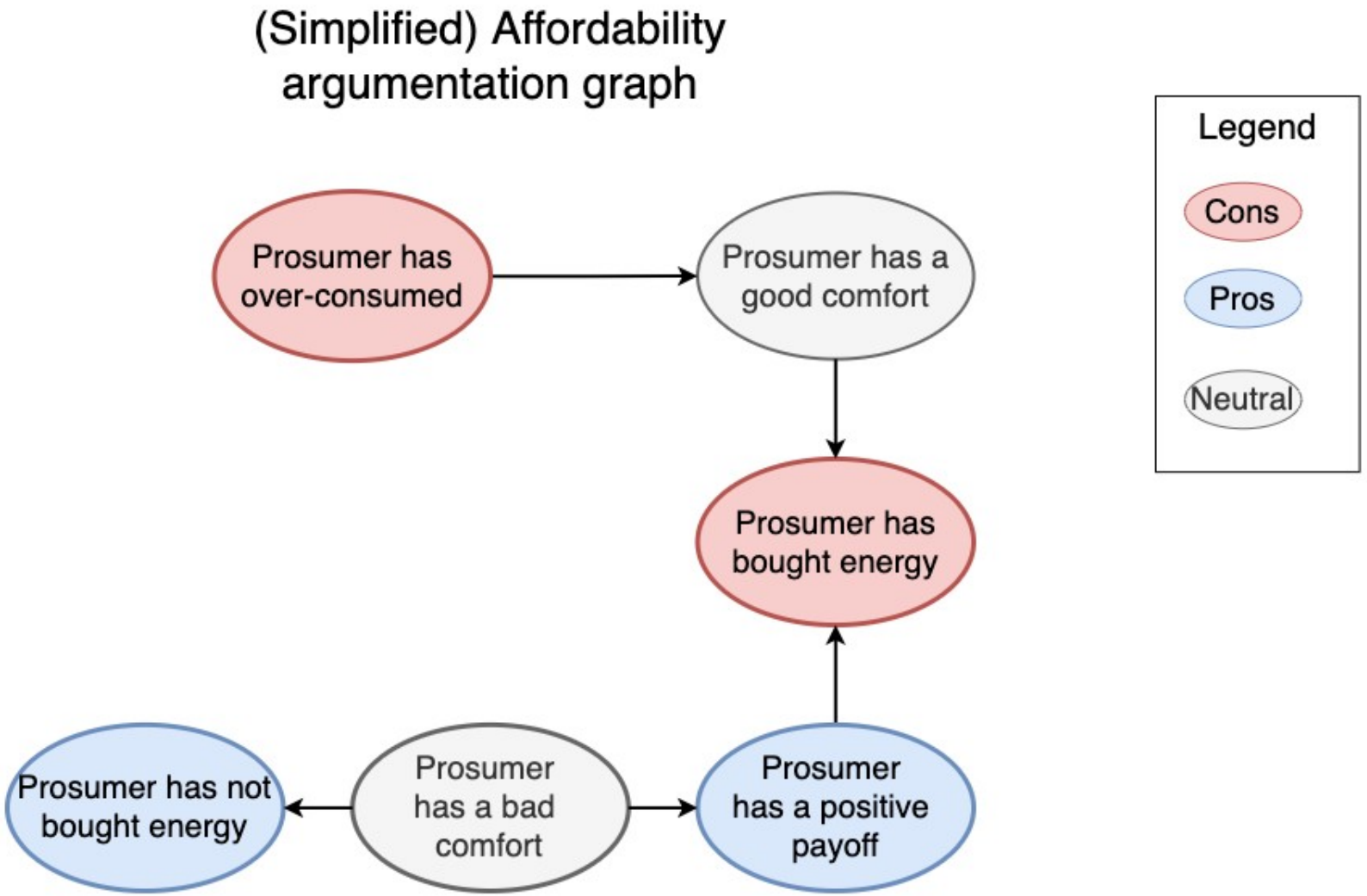
Judgment architecture



Judgment architecture

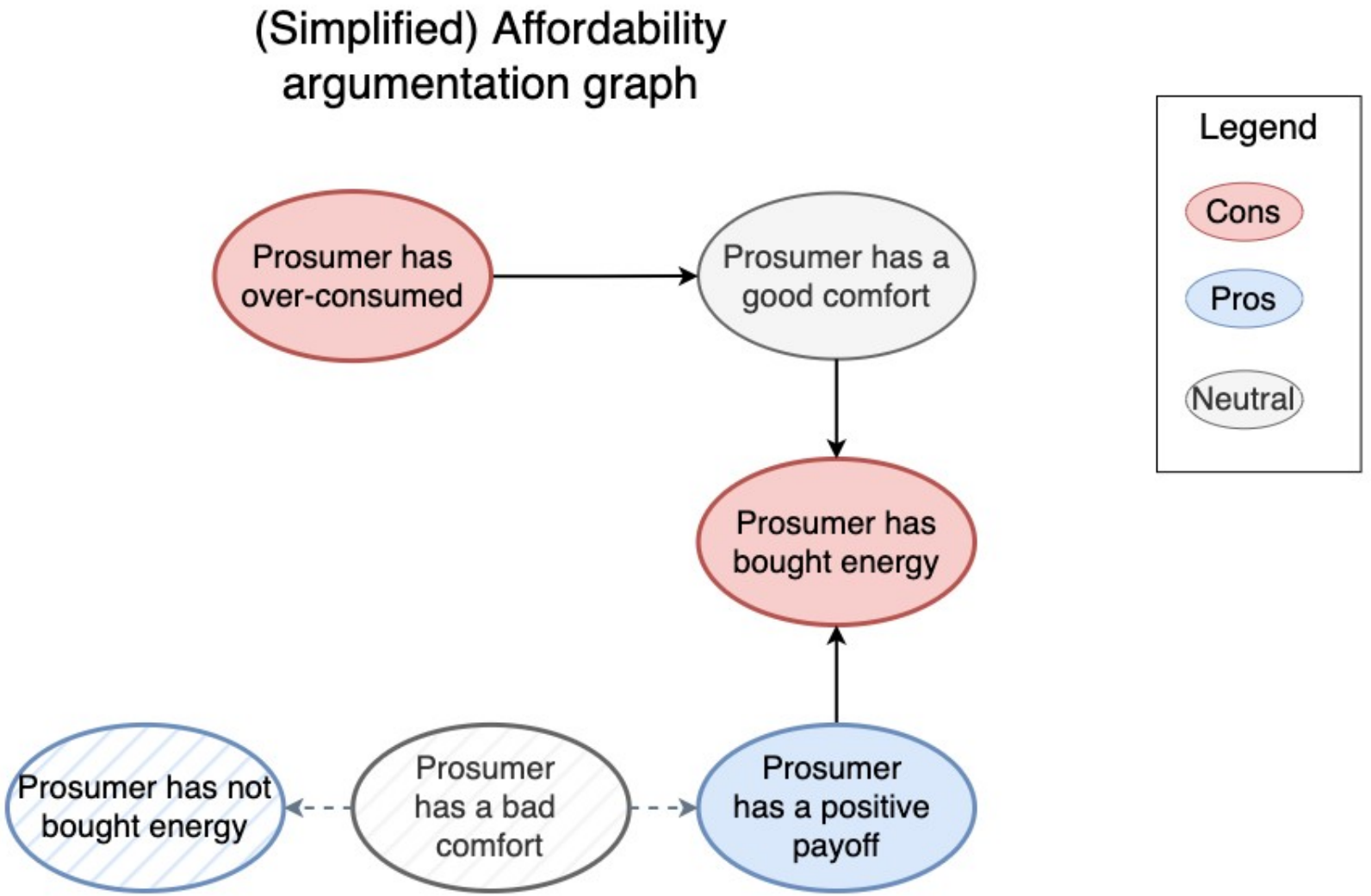


Argumentation-based Judgment



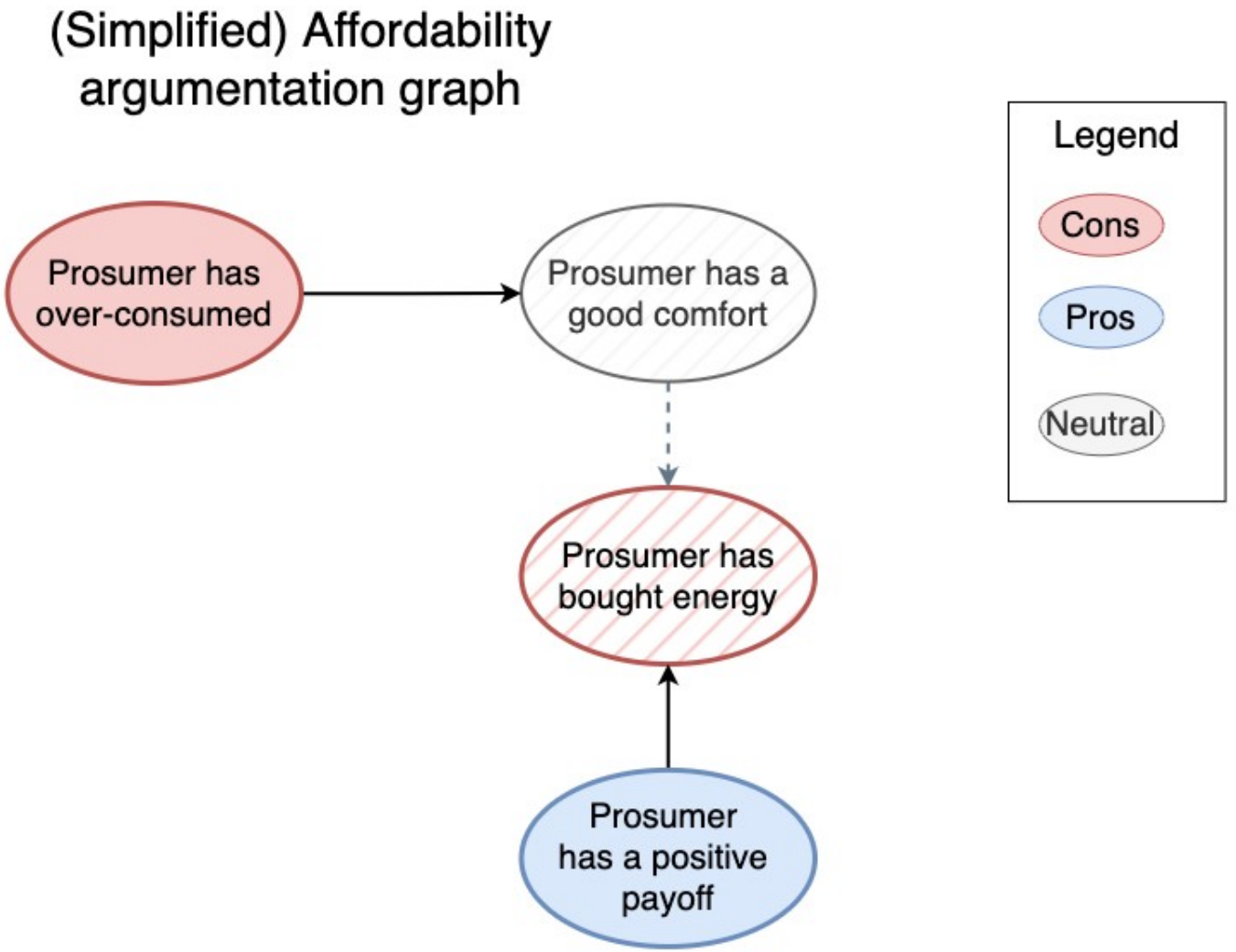
(Amgoud *et al.*, 2009)

Argumentation-based Judgment



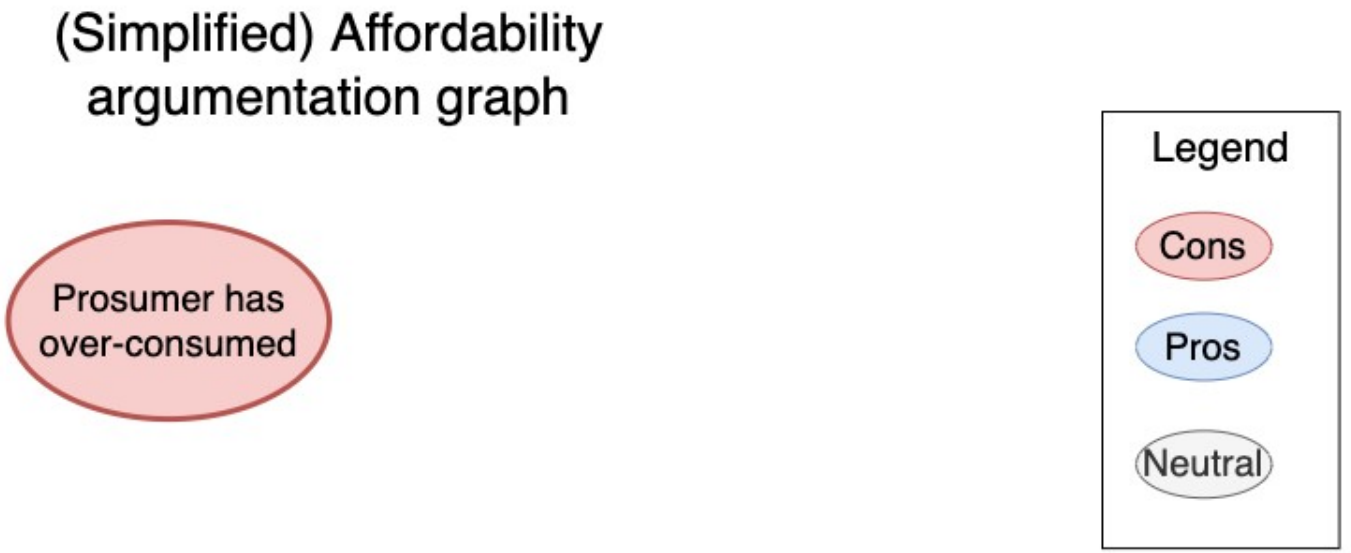
(Amgoud *et al.*, 2009)

Argumentation-based Judgment



(Amgoud *et al.*, 2009)

Argumentation-based Judgment



(Amgoud *et al.*, 2009)

What we have achieved

Objective	Mechanism
Handle complex environments – Multiple persons	• Multi-agents • Individual observations
Handle complex environments – Multiple values	• Multiple judging agents
Handle complex environments – Multiple situations	• (D)SOMs
Adapt to shifting ethical consensus	• Non-convergence • DSOMs • Agentified reward functions
Learn behaviours with non-dilemma situations	• Learning algorithms
Specify desired behaviour	• Symbolic judgments
Learn behaviours with dilemma situations	

What we still lack

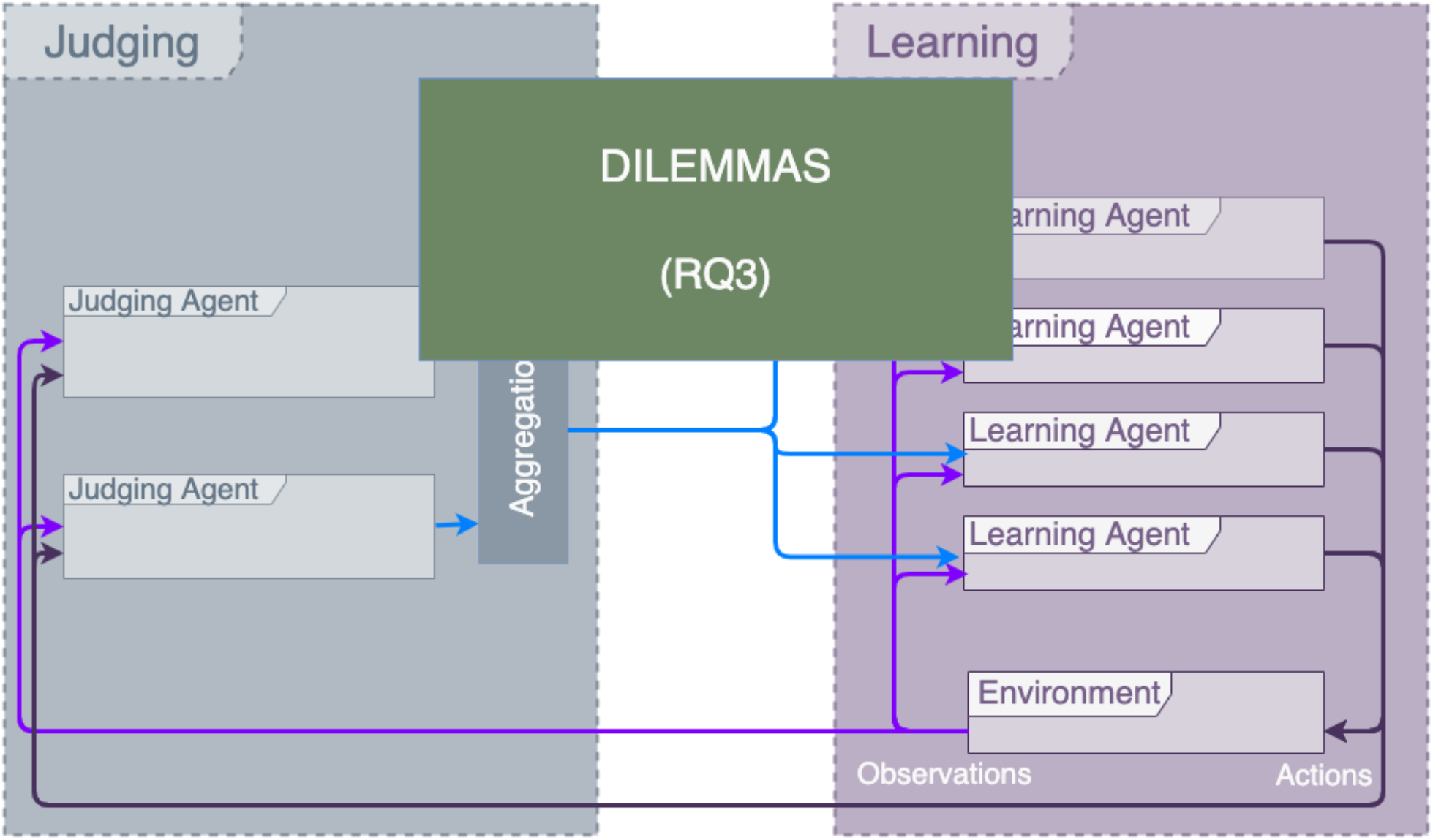
Explicit dilemmas

- Hidden dilemmas because of aggregation

average(1, 1, 1, 0.2) = 0.8

Human preferences

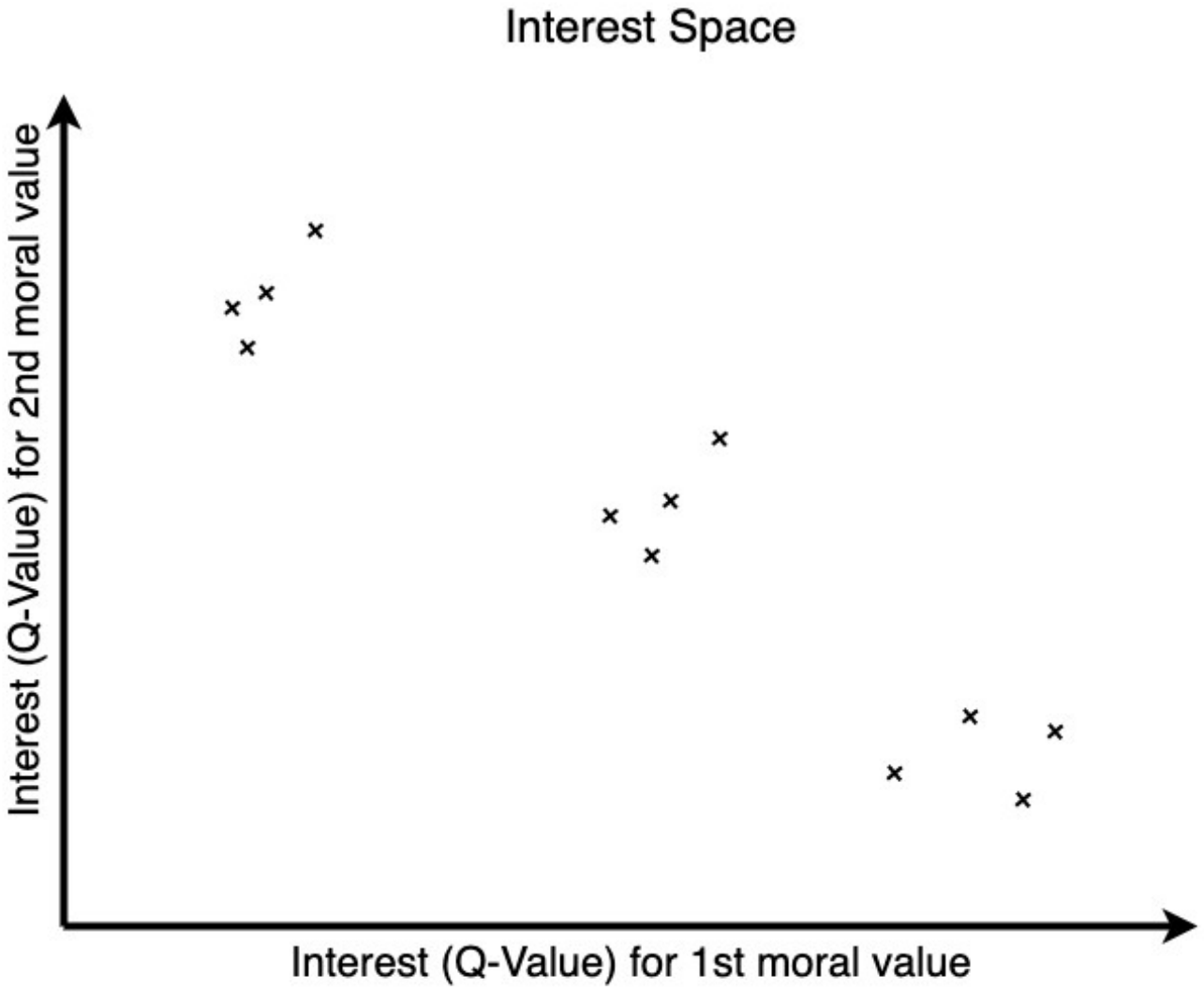
- Favouring a moral value


Judging
Learning
DILEMMAS
(RQ3)
Judging Agent
Judging Agent
Aggregatio
arning Agent
arning Agent
Learning Agent
Learning Agent
Environment
Observations
Actions

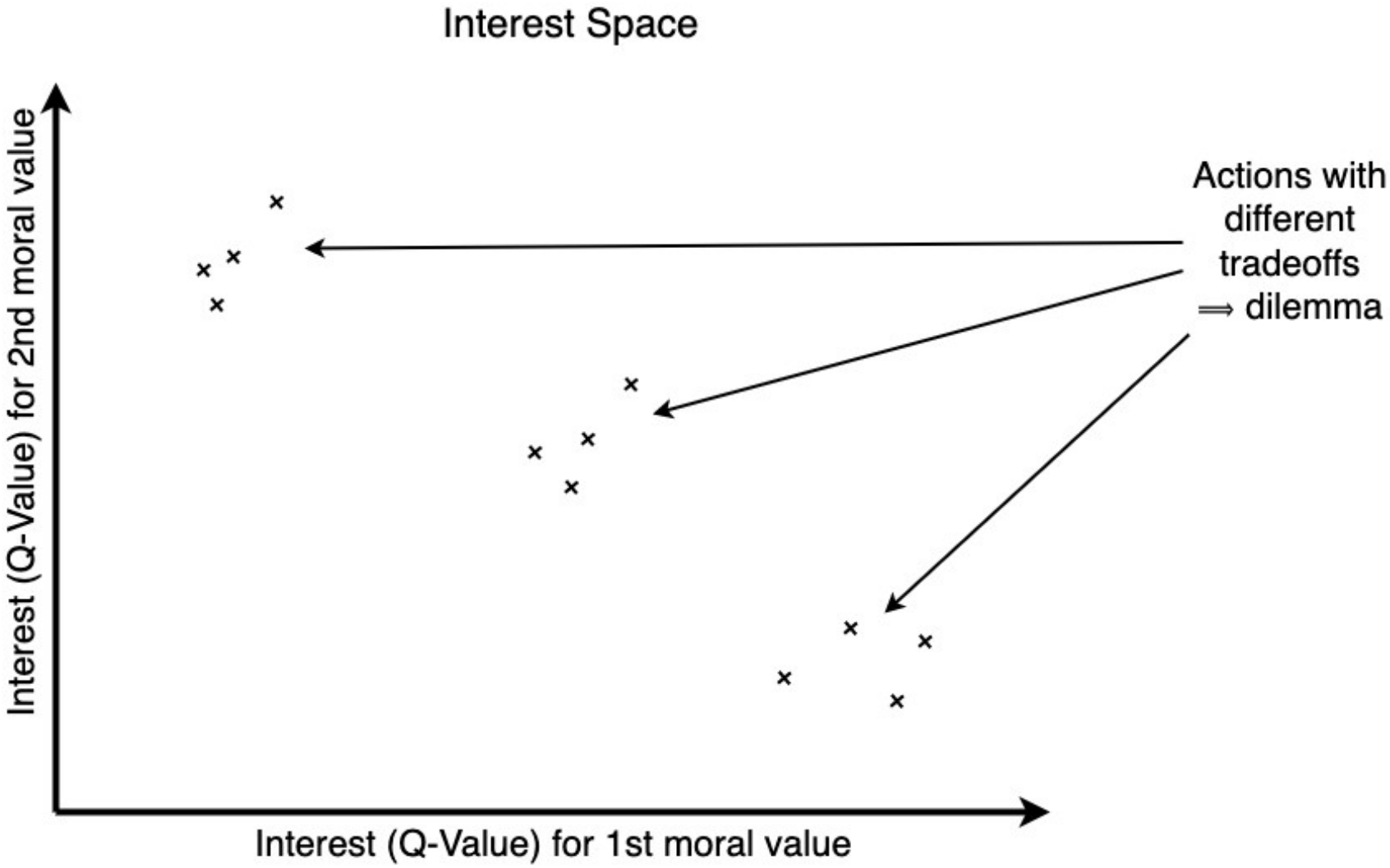
Desired properties for multi-objective RL

- Explicit identification of dilemmas
- Take human preferences into account
- Human-actionable preferences
- “Manageable” for humans

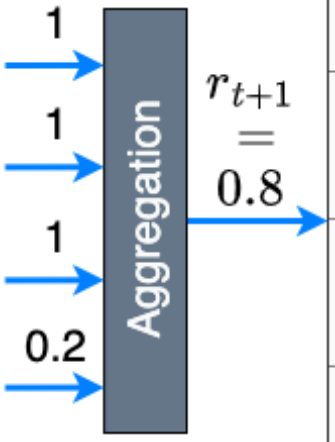
Dilemmas in multi-objective space



Dilemmas in multi-objective space



The multi-objective Q-Table



	Action 1	Action 2	Action 3	Action 4
State 1	3	5	3.5	3
...	...	...	...	...
State 7	1	0.5	4	2
...	...	...	...	...
State 9	1	1.5	0.5	0

$$Q(s_t, a_t) \leftarrow \alpha \left[r_{t+1} + \gamma max_{a'} Q(s_{t+1}, a') \right] + (1 - \alpha) Q(s_t, a_t)$$

The multi-objective Q-Table

$r_{t+1} = [1, 1, 1, 0.2]$

	Action 1	Action 2	Action 3	Action 4
State 1	[3, 4, 3.5, 3]	[4, 3, 5, 4]	[3.5, 3.5, 4, 4]	[5, 3, 3.5, 4]
...	...	...	...	...
State 7	[3.5, 0, 3, 1]	[0.5, 5, 2, 3]	[3, 1, 4, 0.5]	[2, 2, 2, 2]
...	...	...	...	...
State 9	[1, 1, 0.5, 2]	[2, 0.5, 0, 1]	[0.5, 0.5, 0.5, 1]	[1, 0, 2, 0]

1 → State 1

1 → ...

1 → State 7

0.2 → ...

$$\forall k \in [[1, m]] Q(s_t, a_t, k) \leftarrow \alpha \left[r_{t+1,k} + \gamma max_{a'} Q(s_{t+1}, a', k) \right] + (1 - \alpha) Q(s_t, a_t, k)$$

The multi-objective Q-Table

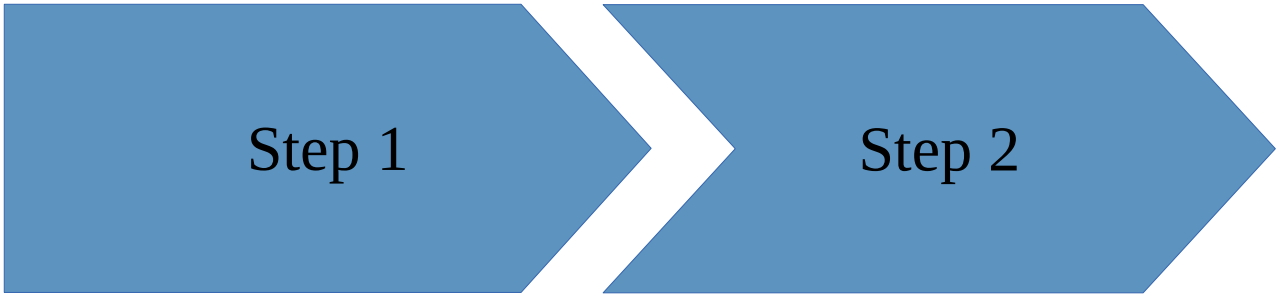
	Action 1	Action 2	Action 3	Action 4
State 1	[3, 4, 3.5, 3]	[4, 3, 5, 4]	[3.5, 3.5, 4, 4]	[5, 3, 3.5, 4]
...	...	...	...	...
State 7	[3.5, 0, 3, 1]	[0.5, 5, 2, 3]	[3, 1, 4, 0.5]	[2, 2, 2, 2]
...	...	...	...	...
State 9	[1, 1, 0.5, 2]	[2, 0.5, 0, 1]	[0.5, 0.5, 0.5, 1]	[1, 0, 2, 0]

$Q(s_7, a_3, m_1)$ = interest of taking "action 3" in "state 7" for the "1st moral value"

Our algorithm in 3 steps

Learning
interesting actions

Our algorithm in 3 steps



Learning interesting actions

Identifying dilemmas

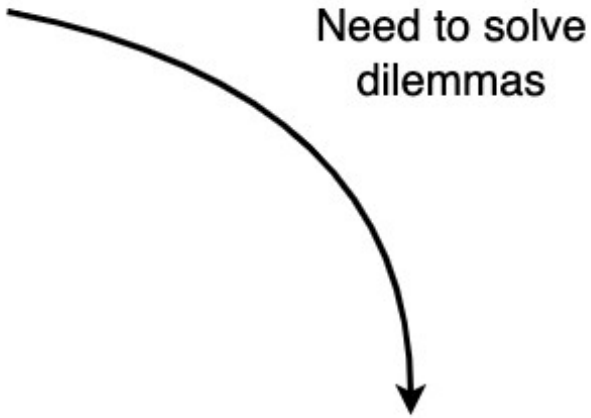
Our algorithm in 3 steps

Learning interesting actions

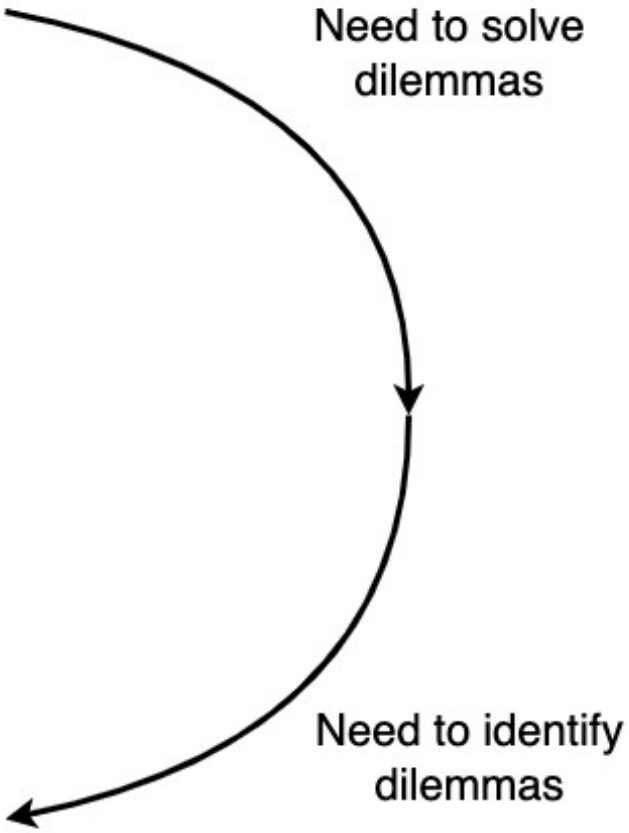
Identifying dilemmas

Learning human preferences

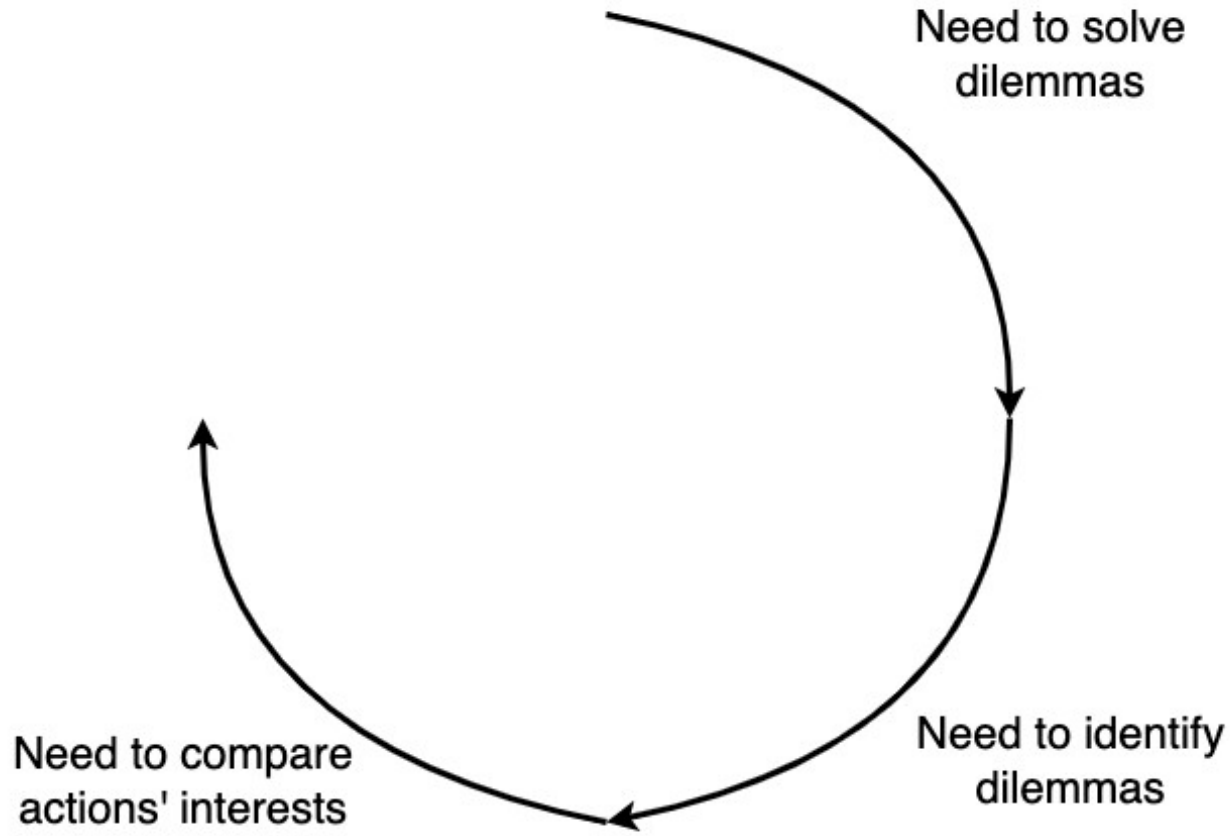
Step 1. Learning interesting actions



Step 1. Learning interesting actions

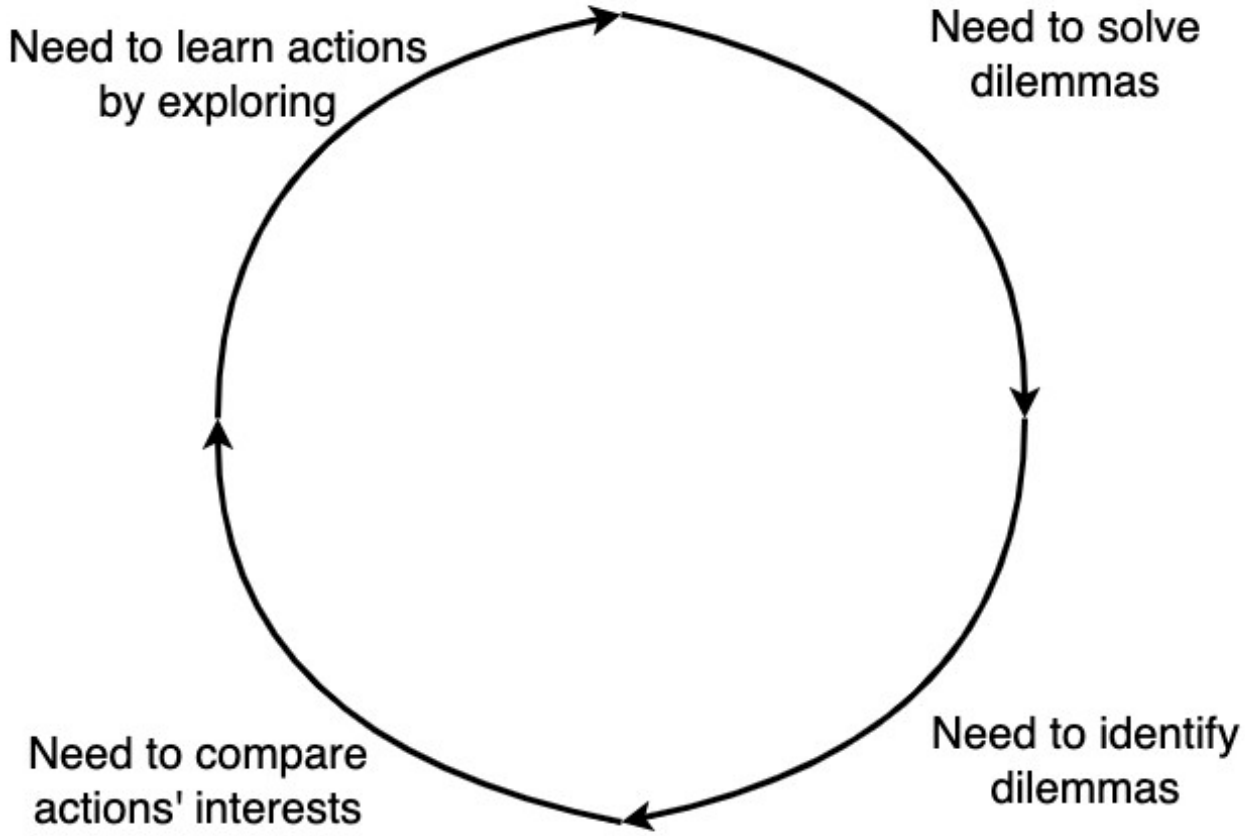


Step 1. Learning interesting actions



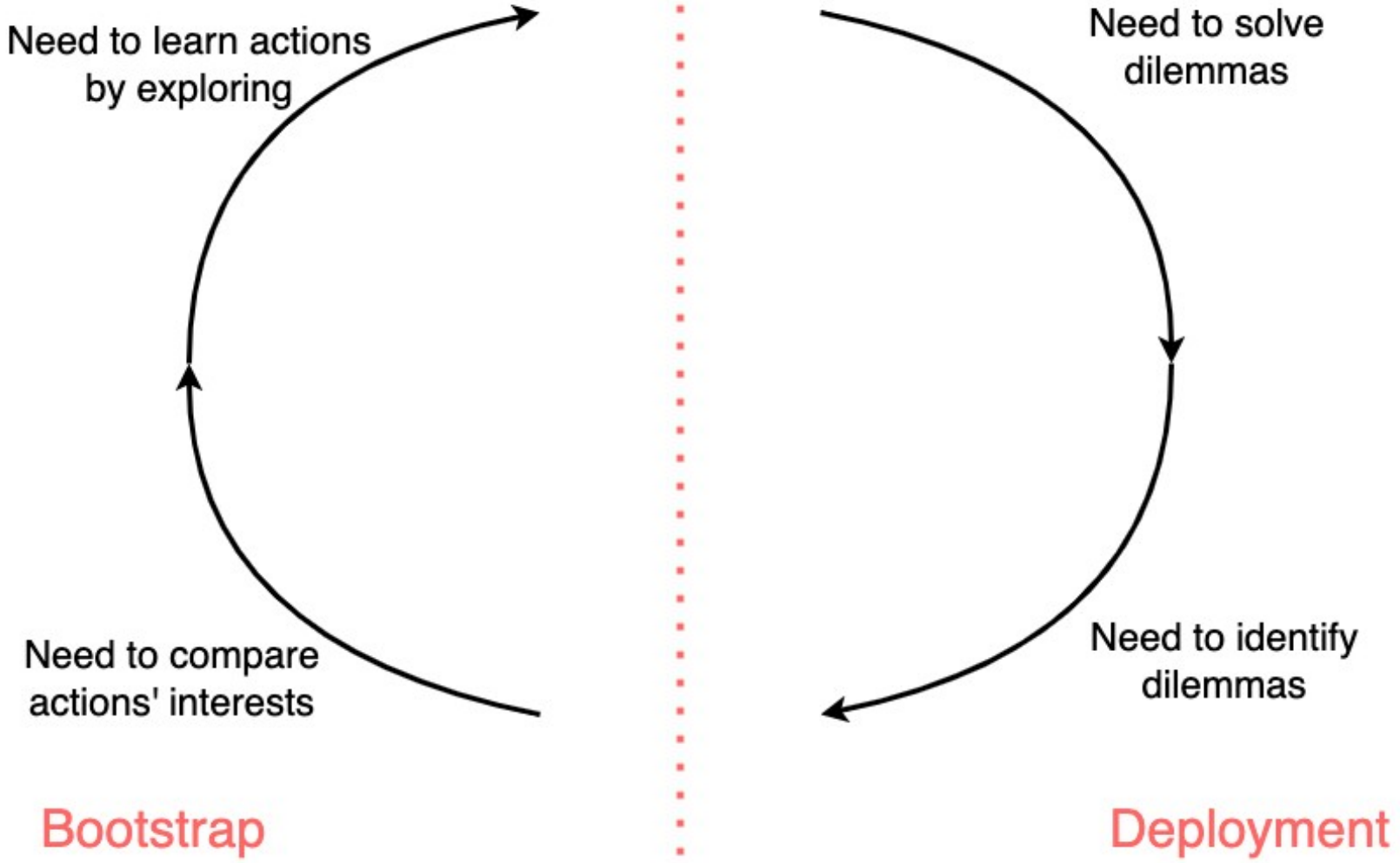
$$[1,0] \overset{?}{>} [0,1]$$

Step 1. Learning interesting actions

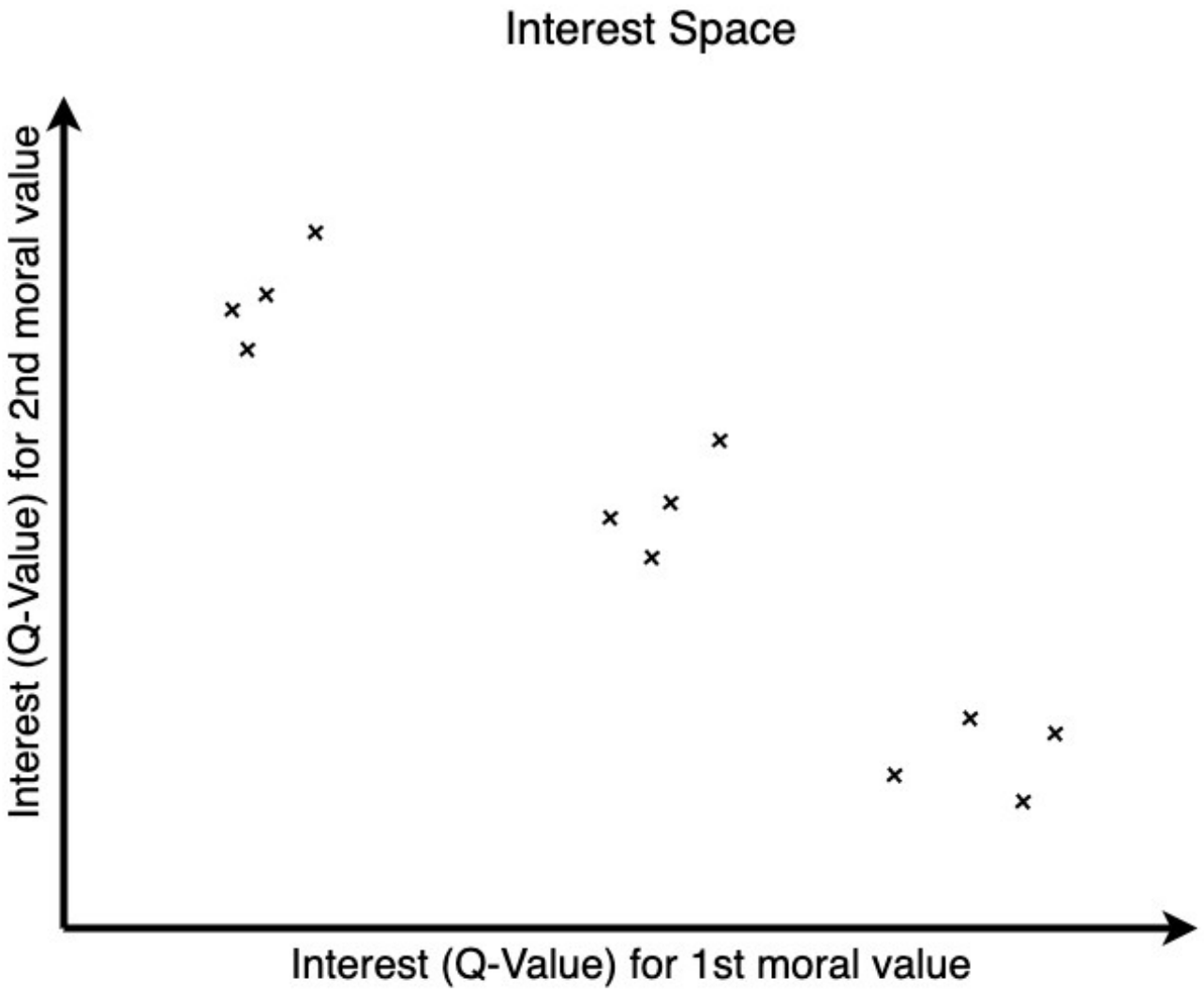


$[1, 0] \overset{?}{>} [0, 1]$

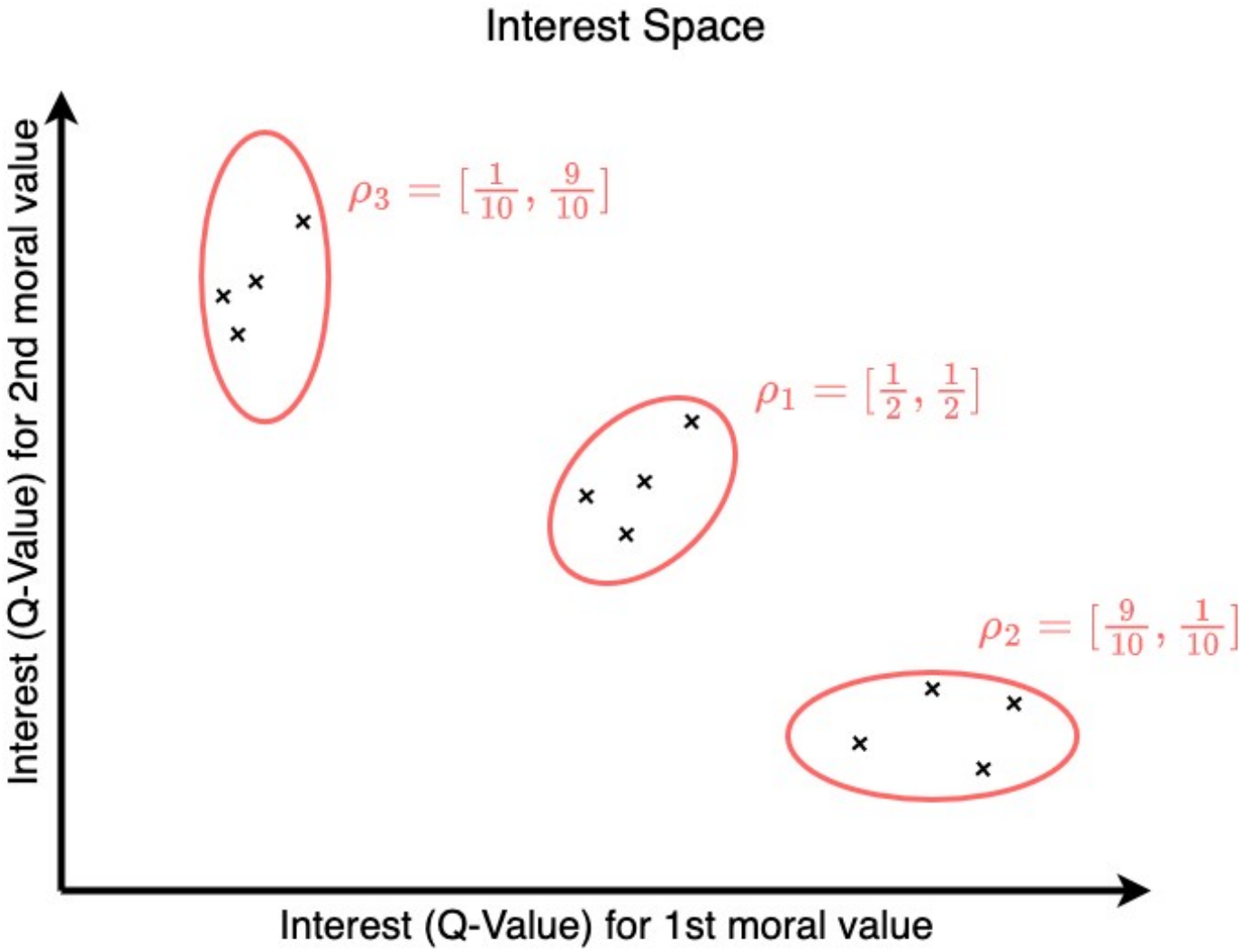
Step 1. Learning interesting actions



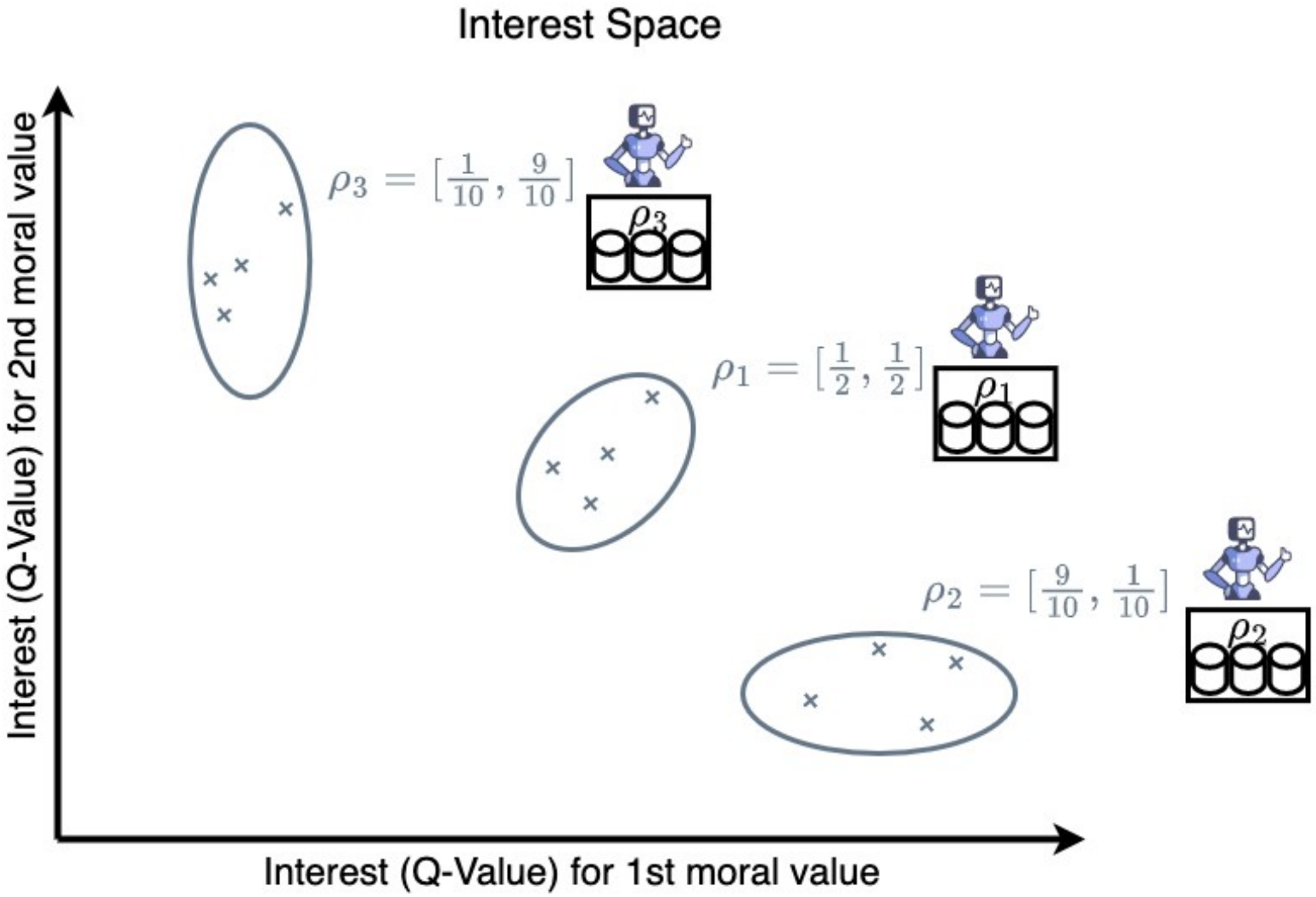
Step 1. Learning interesting actions



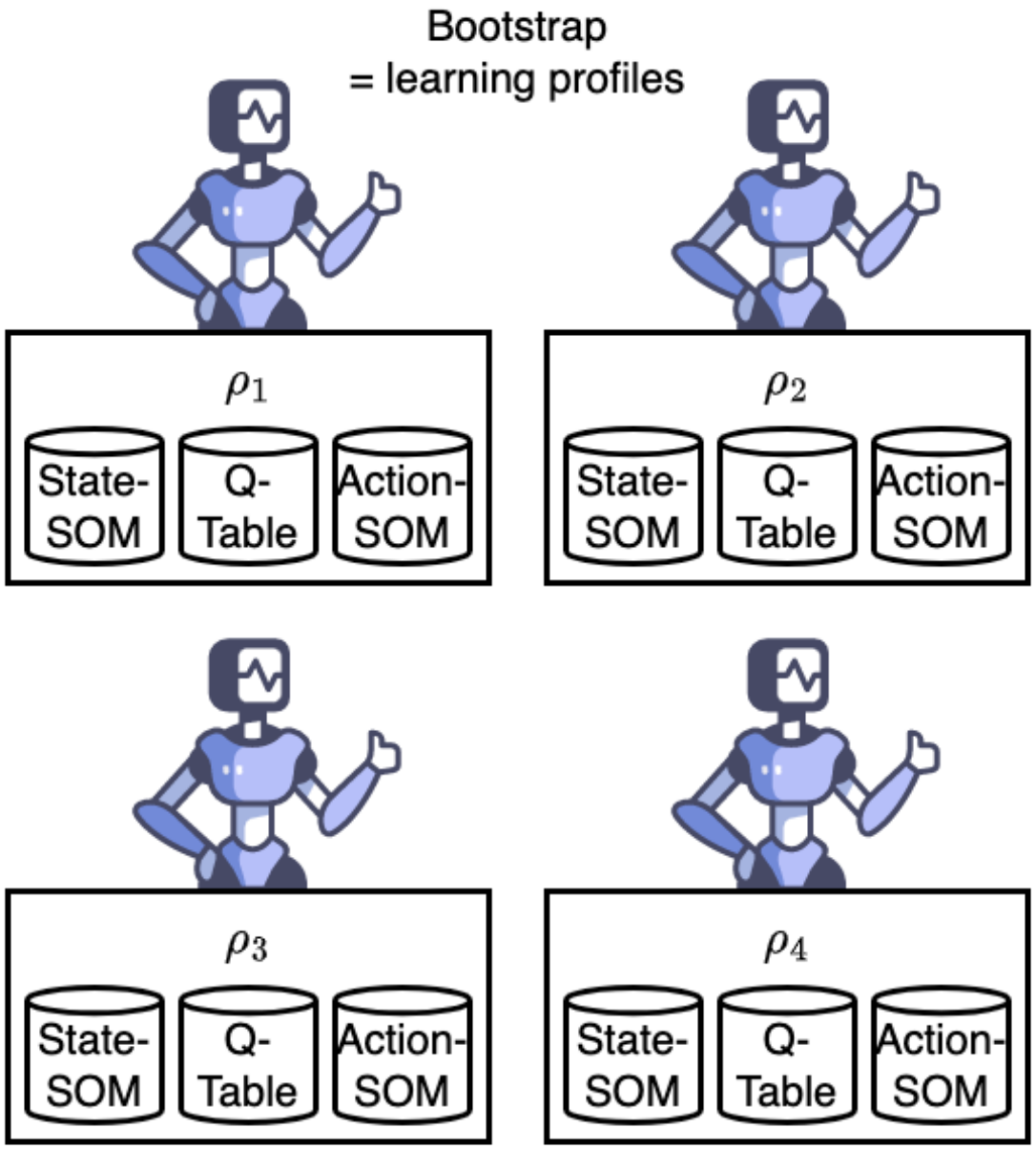
Step 1. Learning interesting actions



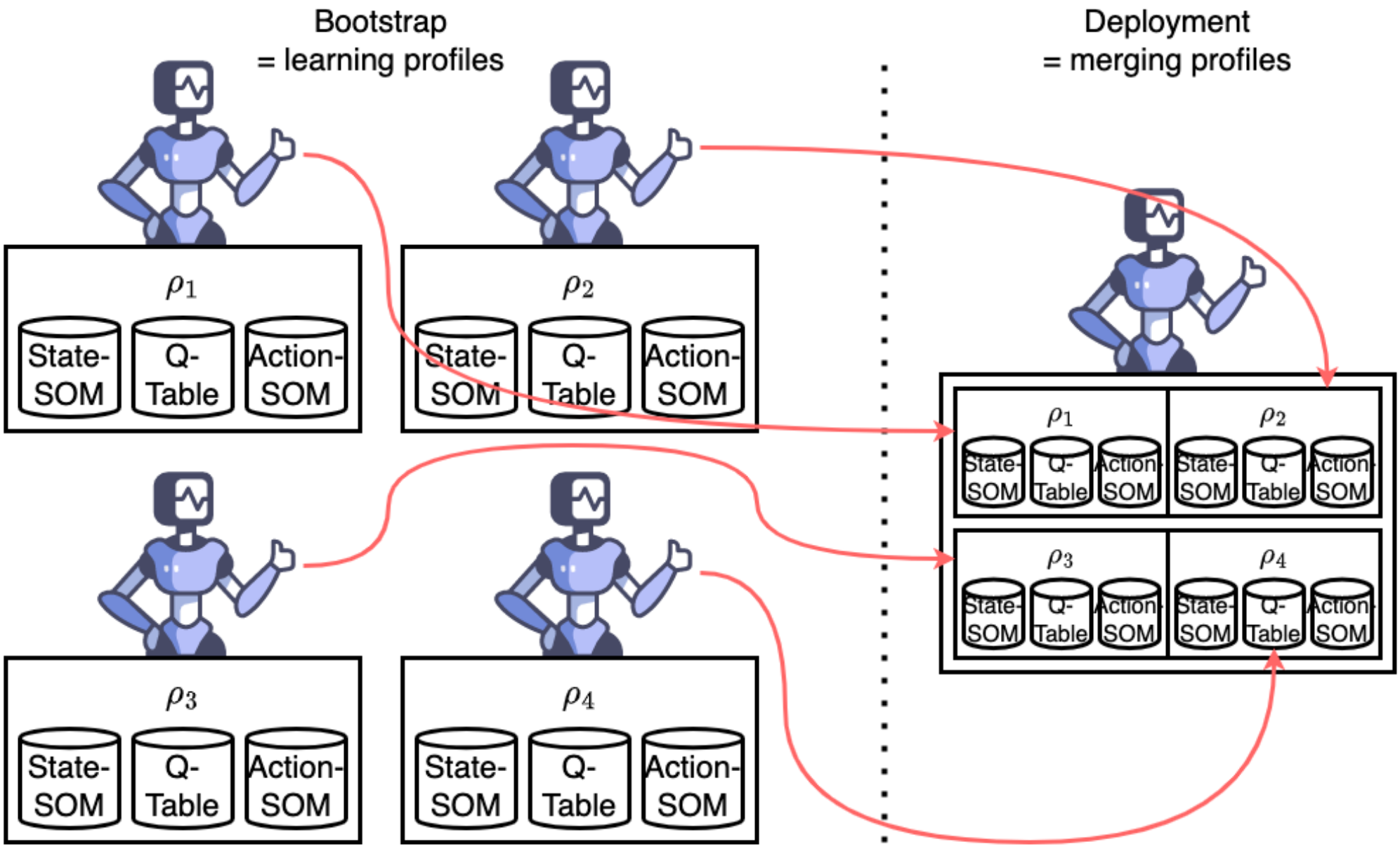
Step 1. Learning interesting actions



Step 1. Learning interesting actions



Step 1. Learning interesting actions



Step 2. Identifying dilemmas

Profile/Action	Interests
p_1, a_1	[3, 4, 3.5, 3]
p_1, a_2	[2.5, 3.5, 3, 3.5]
...	...
p_2, a_4	[5, 3, 2.5, 3]
...	...
p_3, a_7	[3, 4, 3, 3.5]
...	...
p_4, a_5	[2, 1, 6, 0.5]
...	...
p_5, a_9	[1.5, 2, 3, 3]

All alternatives from all profiles

Step 2. Identifying dilemmas

Profile/Action	Interests
p_1, a_1	[3, 4, 3.5, 3]
p_1, a_2	[2.5, 3.5, 3, 3.5]
...	...
p_2, a_4	[5, 3, 2.5, 3]
...	...
p_3, a_7	[3, 4, 3, 3.5]
...	...
p_4, a_5	[2, 1, 6, 0.5]
...	...
p_5, a_9	**[1, 2, 3.5, 3]**

Pareto-dominates

All alternatives from all profiles

Step 2. Identifying dilemmas

Profile/Action	Interests
p_1, a_1	[3, 4, 3.5, 3]
p_2, a_4	[5, 3, 2.5, 3]
p_3, a_7	[3, 4, 3, 3.5]
p_4, a_5	[2, 1, 6, 0.5]

Pareto-optimal alternatives

Not dominated by any other

Step 2. Identifying dilemmas

Profile/Action	Interests	Theoreticals
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]

Normal Q-Value

Q-Value if the reward was maximal

Step 2. Identifying dilemmas

Profile/Action	Interests	Theoreticals	Ratio
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$\left[\frac{3}{6}, \frac{4}{6}, \frac{3.5}{6}, \frac{3}{6}\right]$
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$\left[\frac{5}{7}, \frac{3}{7}, \frac{2.5}{7}, \frac{3}{7}\right]$
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]	$\left[\frac{3}{5}, \frac{4}{5}, \frac{3}{5}, \frac{3.5}{5}\right]$
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$\left[\frac{2}{7}, \frac{1}{7}, \frac{6}{7}, \frac{0.5}{7}\right]$

Interests / Theoreticals

Step 2. Identifying dilemmas

Profile/Action	Interests	Theoreticals	Ratio
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$[50\%, 66\%, 58\%, 50\%]$
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$[71\%, 42\%, 35\%, 42\%]$
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]	$[60\%, 80\%, 60\%, 65\%]$
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$[28\%, 14\%, 85\%, 7\%]$

Human thresholds ζ_1

$[50\%, 75\%, 50\%, 60\%]$

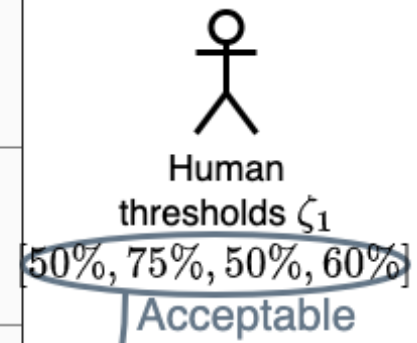
Acceptable

Interests / Theoreticals

Step 2. Identifying dilemmas

Profile/Action	Interests	Theoreticals	Ratio
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$[50\%, 66\%, 58\%, 50\%]$
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$[71\%, 42\%, 35\%, 42\%]$
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]	$[60\%, 80\%, 60\%, 65\%]$
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$[28\%, 14\%, 85\%, 7\%]$

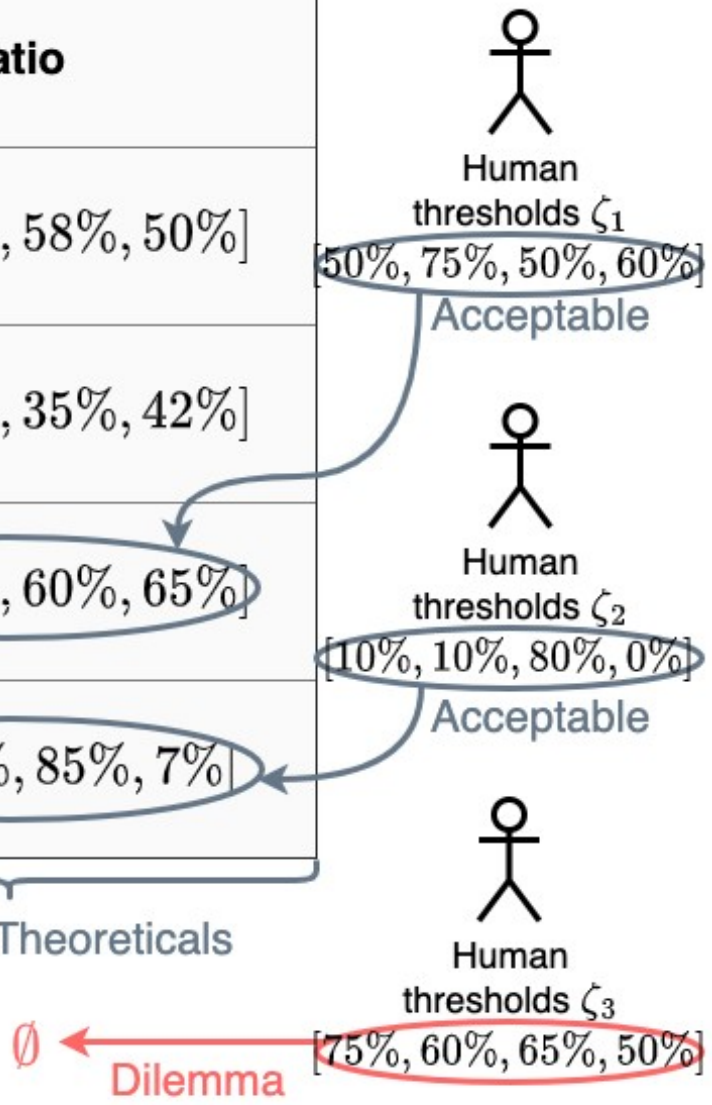
Interests / Theoreticals



Step 2. Identifying dilemmas

Profile/Action	Interests	Theoreticals	Ratio
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$[50\%, 66\%, 58\%, 50\%]$
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$[71\%, 42\%, 35\%, 42\%]$
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]	$[60\%, 80\%, 60\%, 65\%]$
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$[28\%, 14\%, 85\%, 7\%]$

Interests / Theoreticals



Step 2. Identifying dilemmas

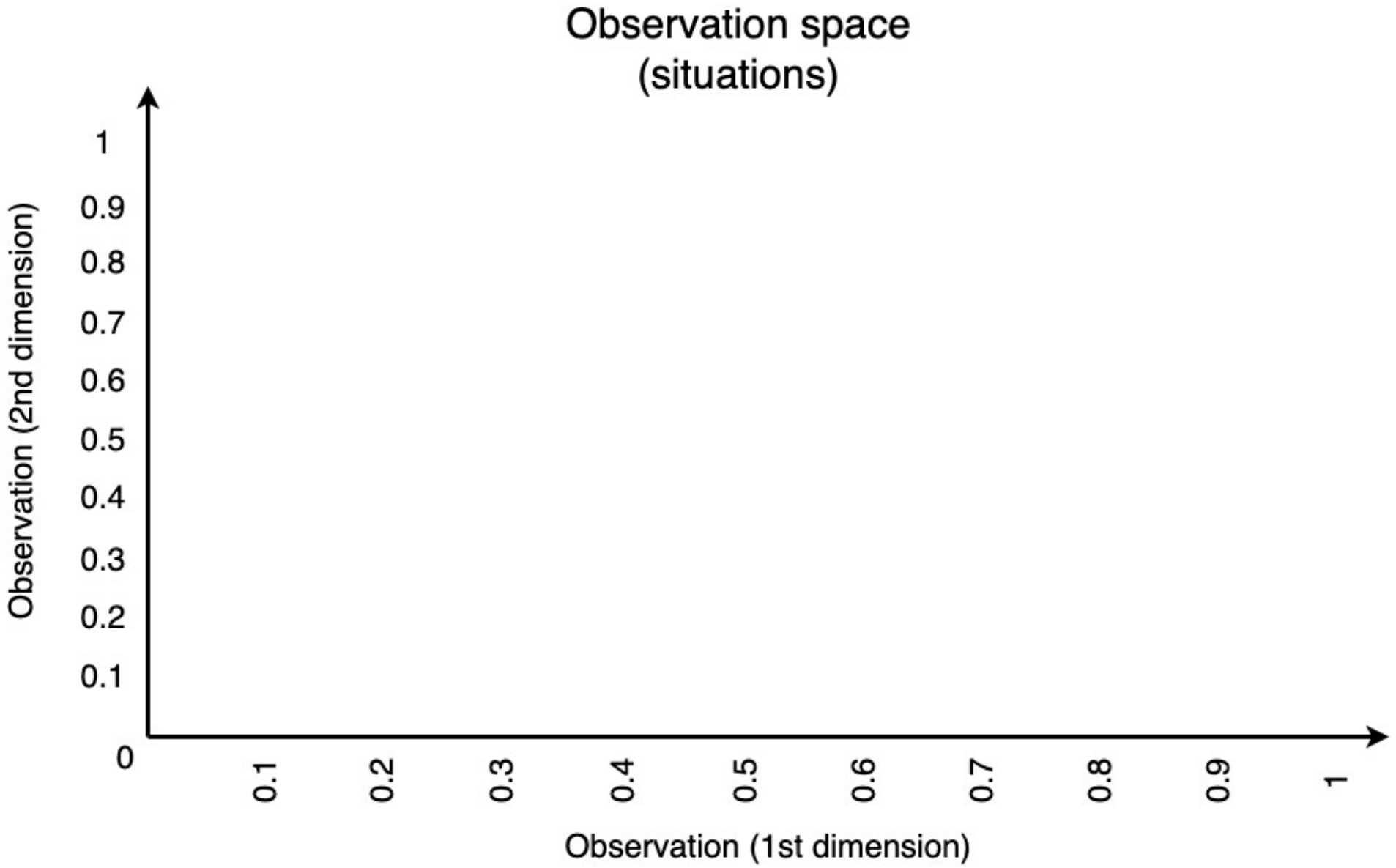
Profile/Action	Interests	Theoreticals	Ratio	Action Parameters
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$[50\%, 66\%, 58\%, 50\%]$	[0.8, 0.1, 0.7, 0.2, 0.5, 0.9]
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$[71\%, 42\%, 35\%, 42\%]$	[0.2, 0.6, 0.3, 0.4, 0.9, 0.1]
p_3, a_7	[3, 4, 3, 3.5]	[5, 5, 5, 5]	$[60\%, 80\%, 60\%, 65\%]$	[0.8, 0.0, 0.7, 0.2, 0.5, 0.9]
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$[28\%, 14\%, 85\%, 7\%]$	[0.0, 0.0, 0.1, 1.0, 0.1, 0.2]

Step 2. Identifying dilemmas

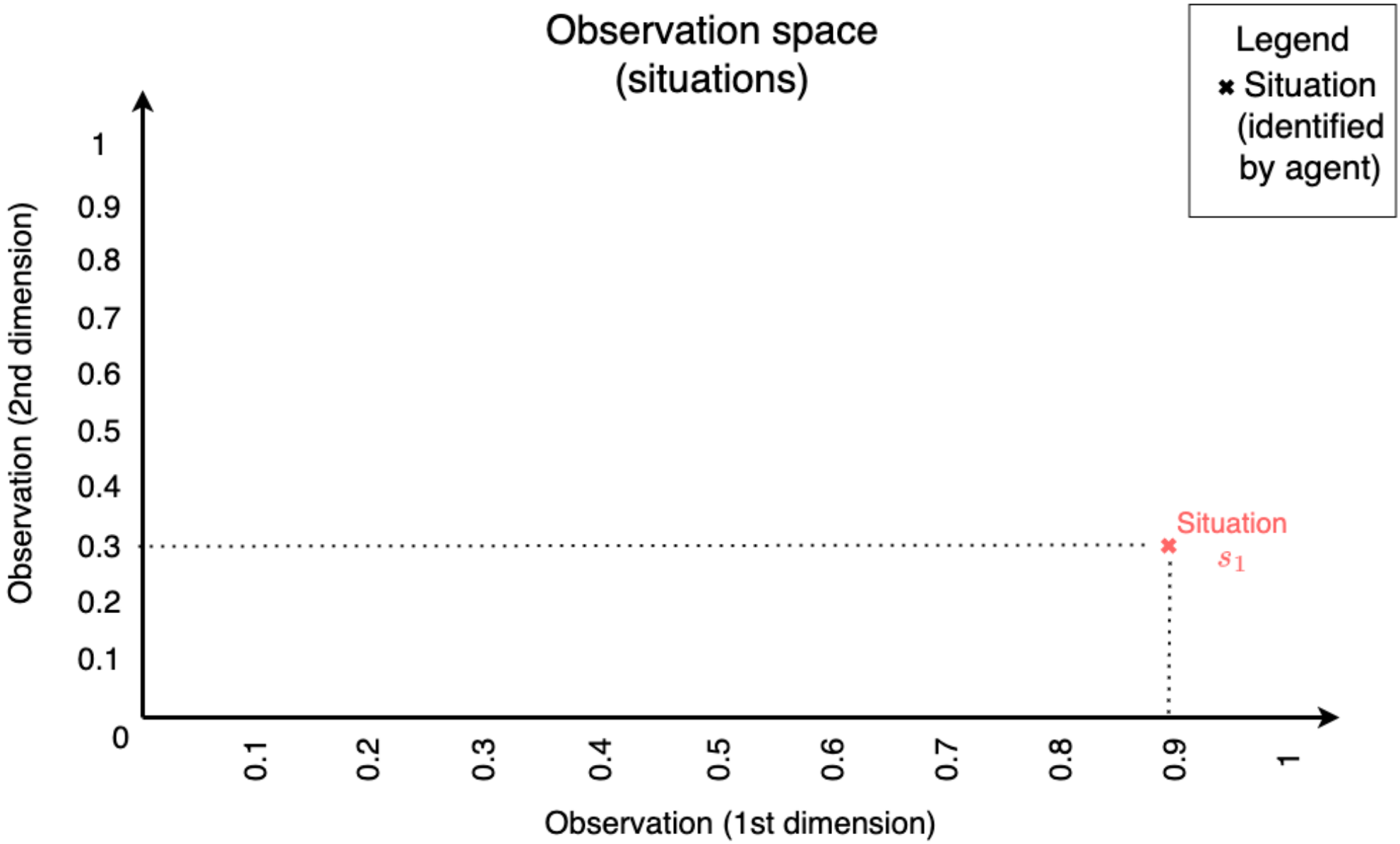
Profile/Action	Interests	Theoreticals	Ratio	Action Parameters
p_1, a_1	[3, 4, 3.5, 3]	[6, 6, 6, 6]	$[50\%, 66\%, 58\%, 50\%]$	[0.8, 0.1, 0.7, 0.2, 0.5, 0.9]
p_2, a_4	[5, 3, 2.5, 3]	[7, 7, 7, 7]	$[71\%, 42\%, 35\%, 42\%]$	[0.2, 0.6, 0.3, 0.4, 0.9, 0.1]
p_4, a_5	[2, 1, 6, 0.5]	[7, 7, 7, 7]	$[28\%, 14\%, 85\%, 7\%]$	[0.0, 0.0, 0.1, 1.0, 0.1, 0.2]

Similar actions removed

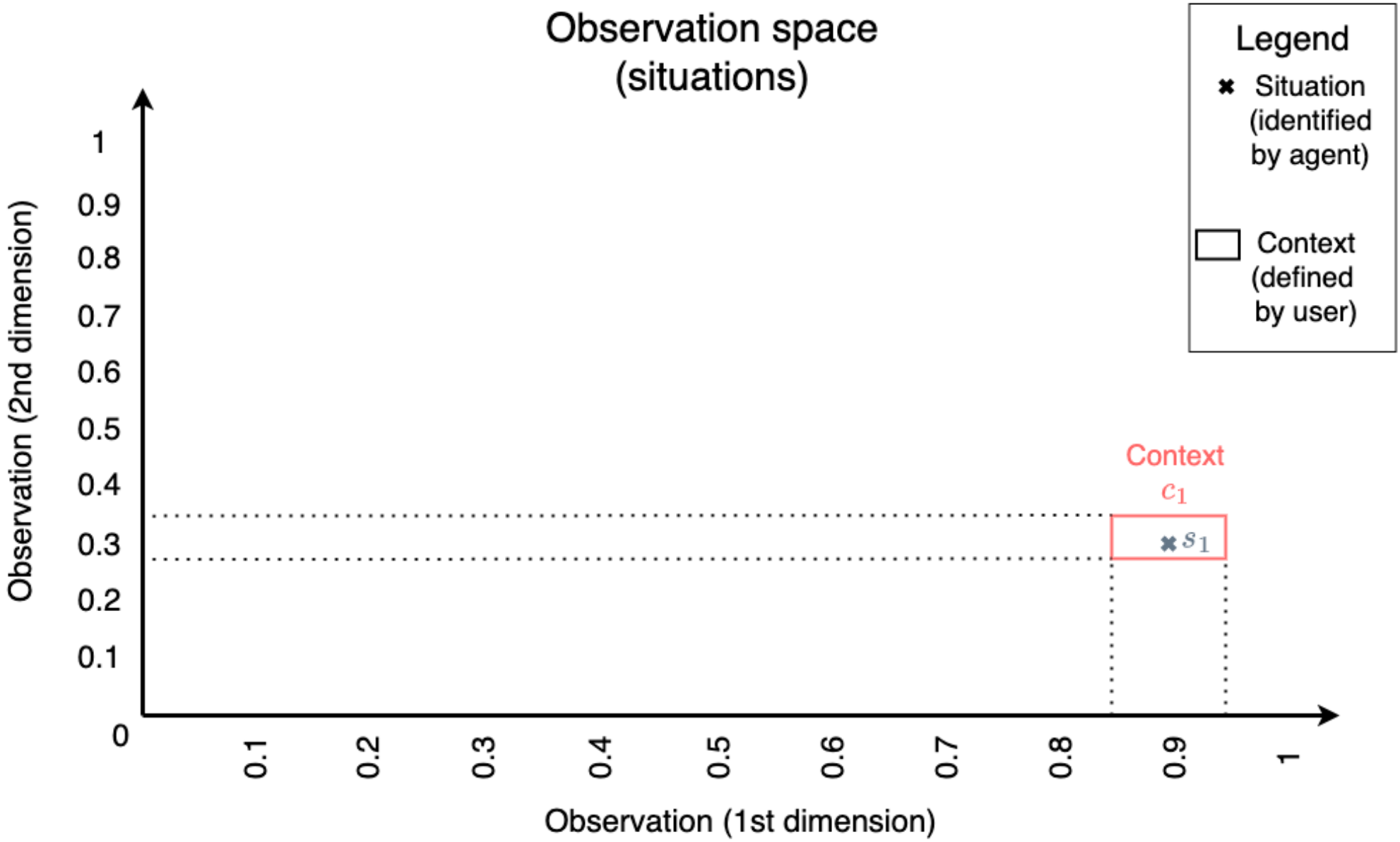
Step 3. Learning human preferences



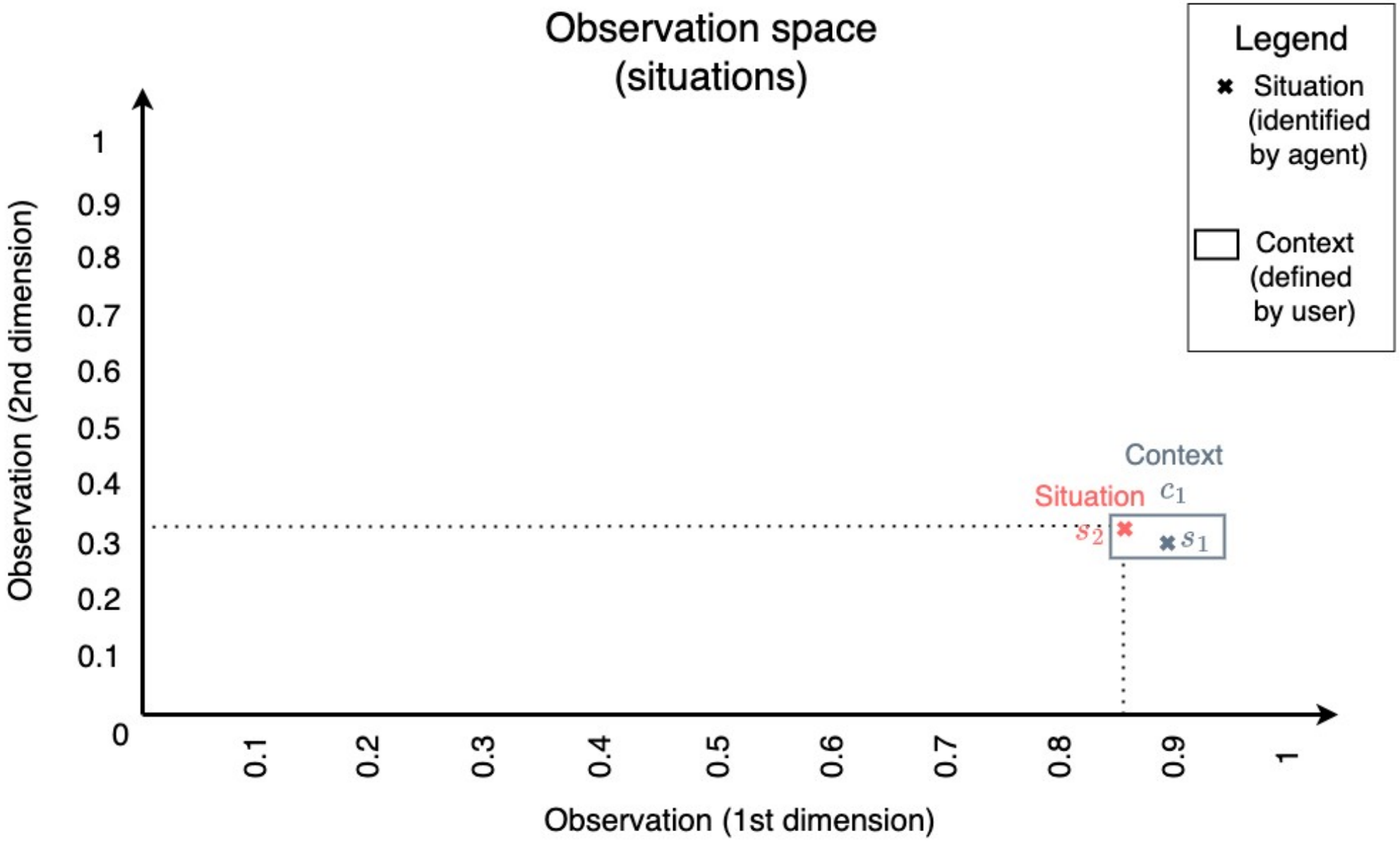
Step 3. Learning human preferences



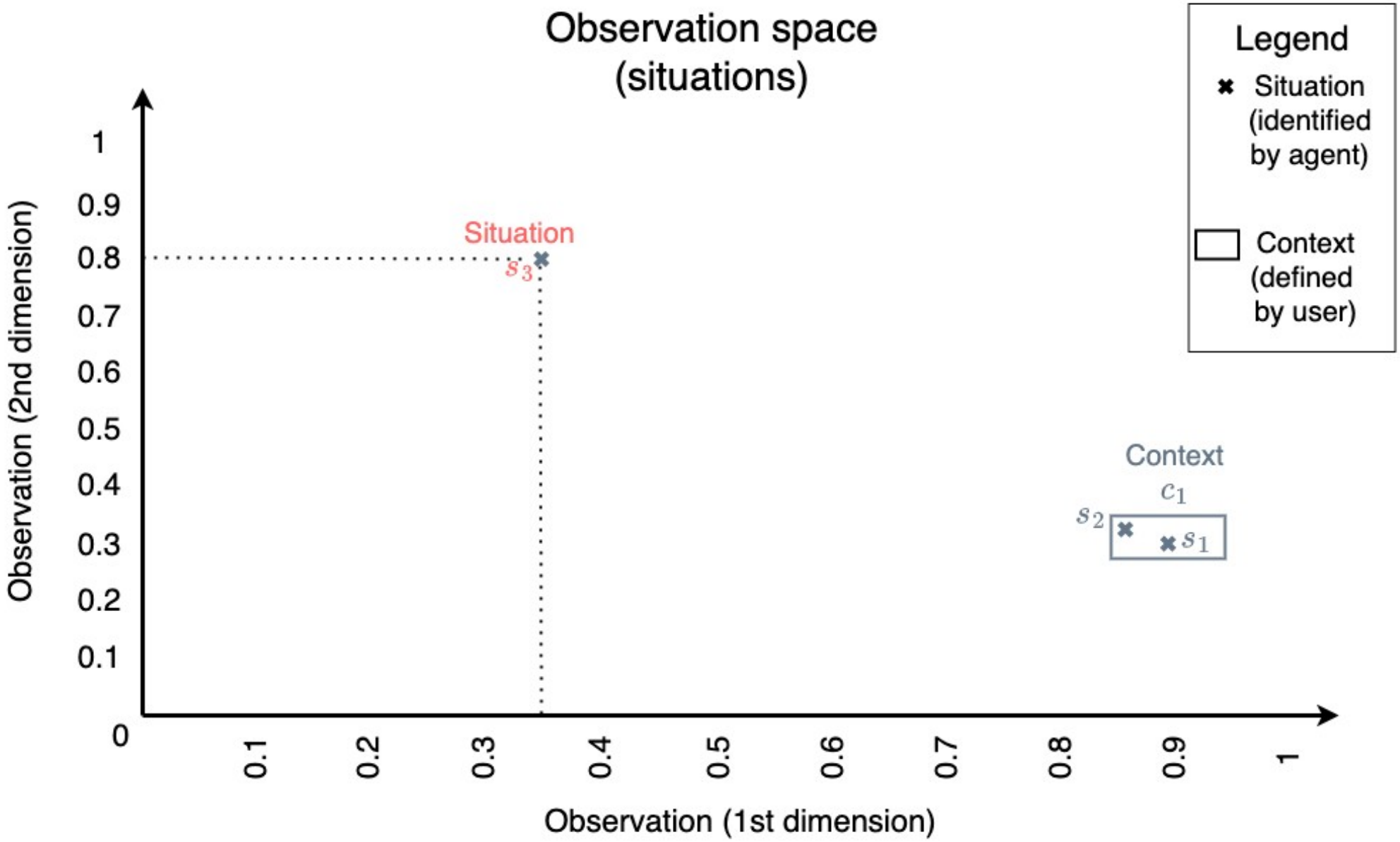
Step 3. Learning human preferences



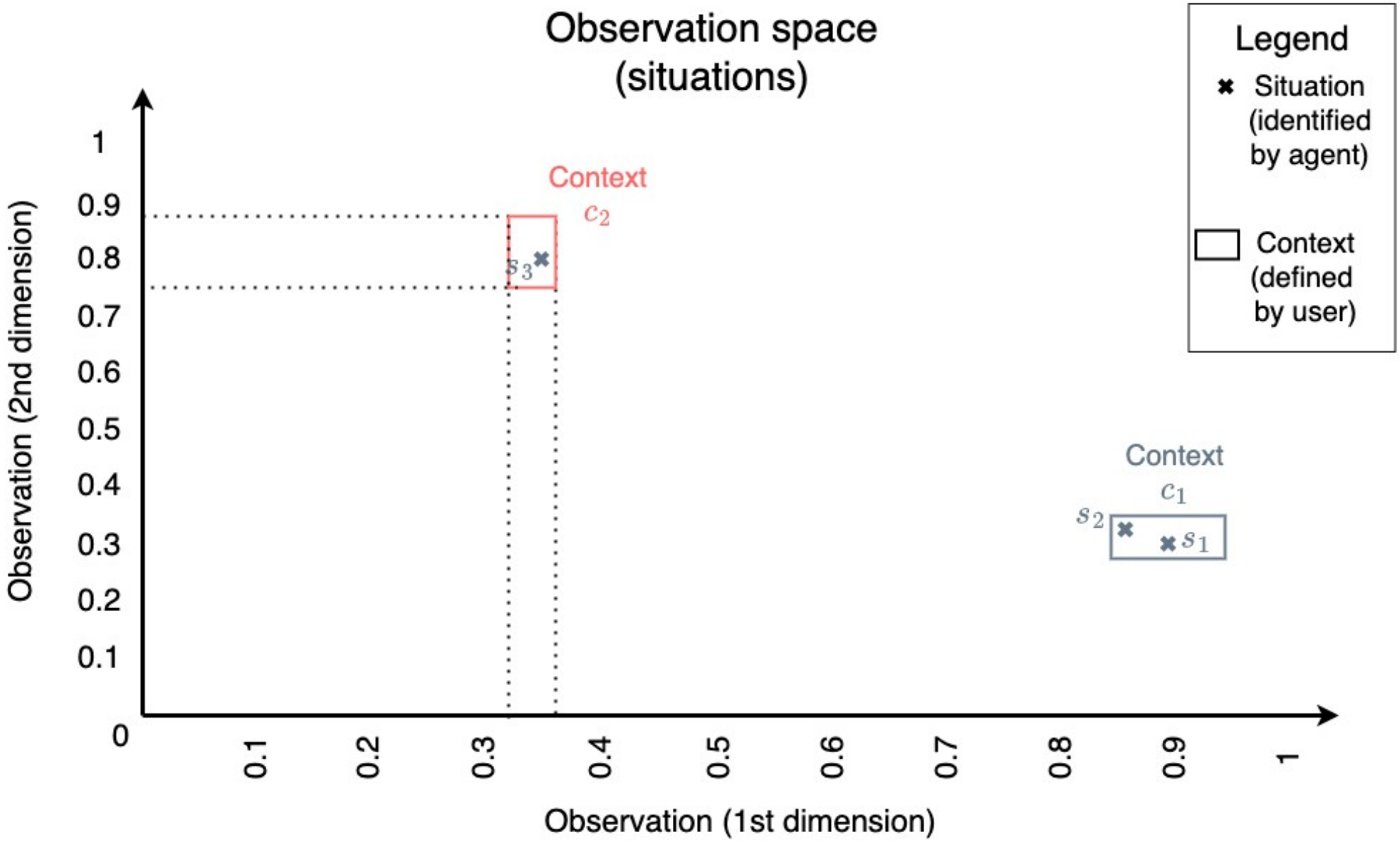
Step 3. Learning human preferences



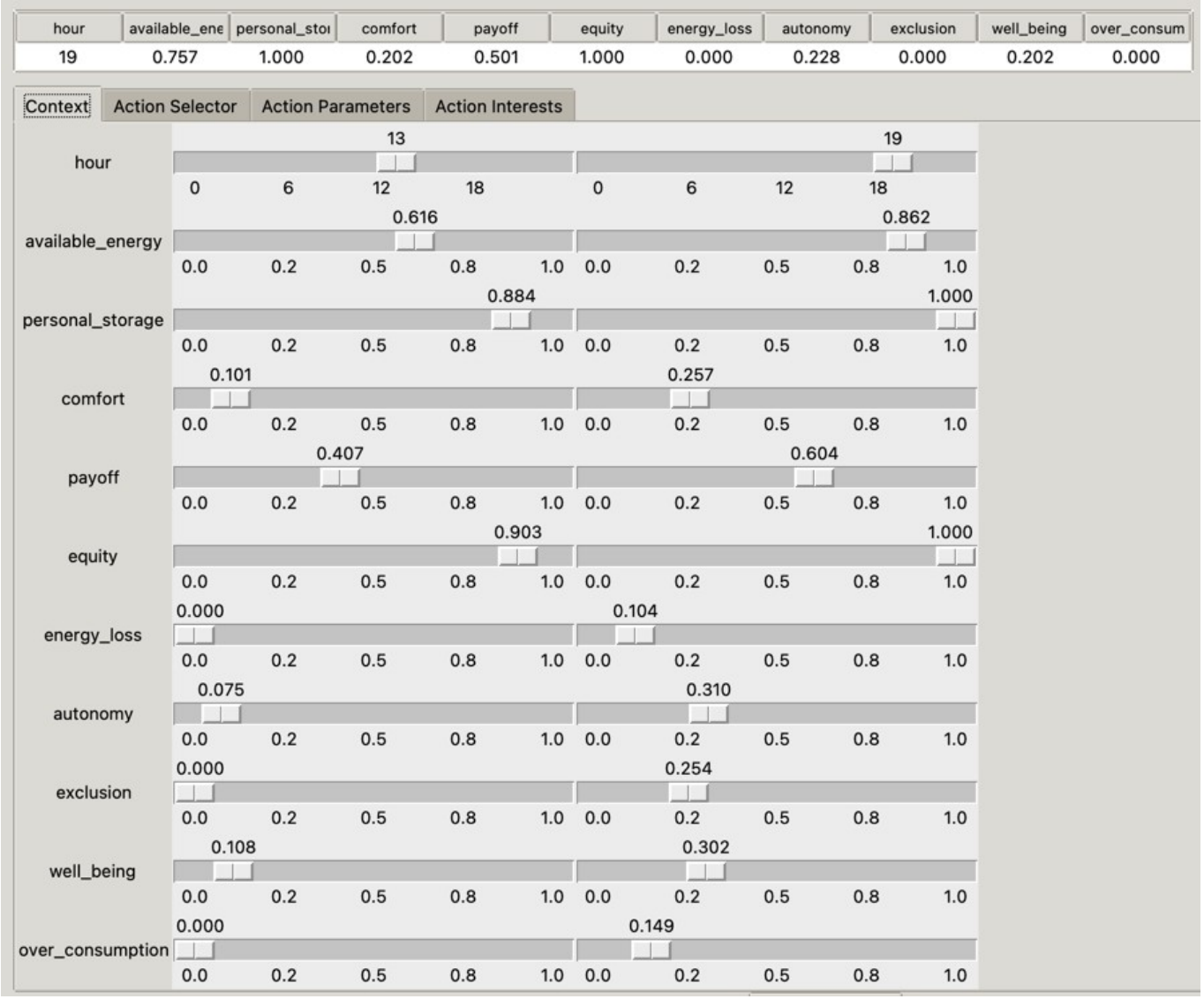
Step 3. Learning human preferences



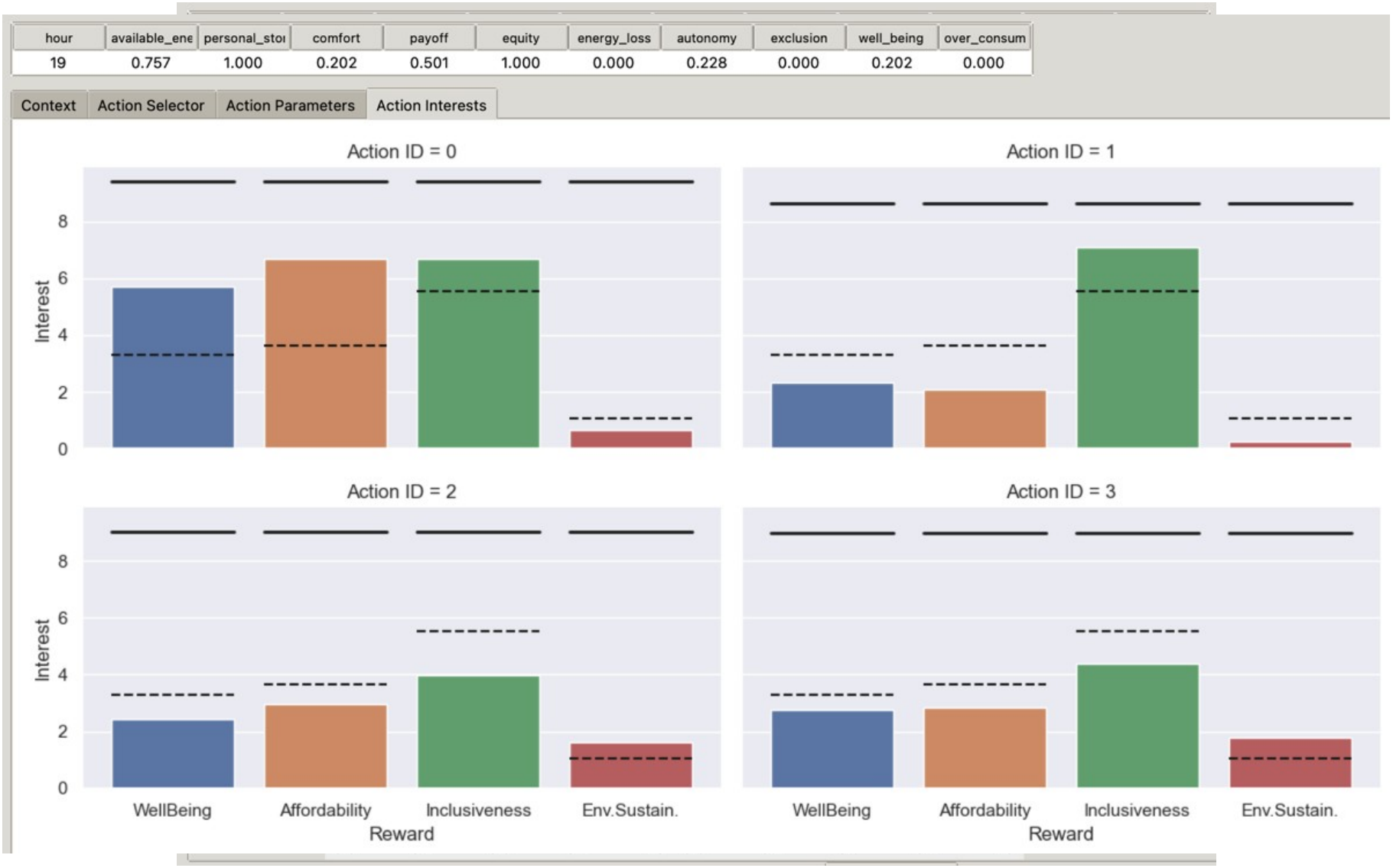
Step 3. Learning human preferences



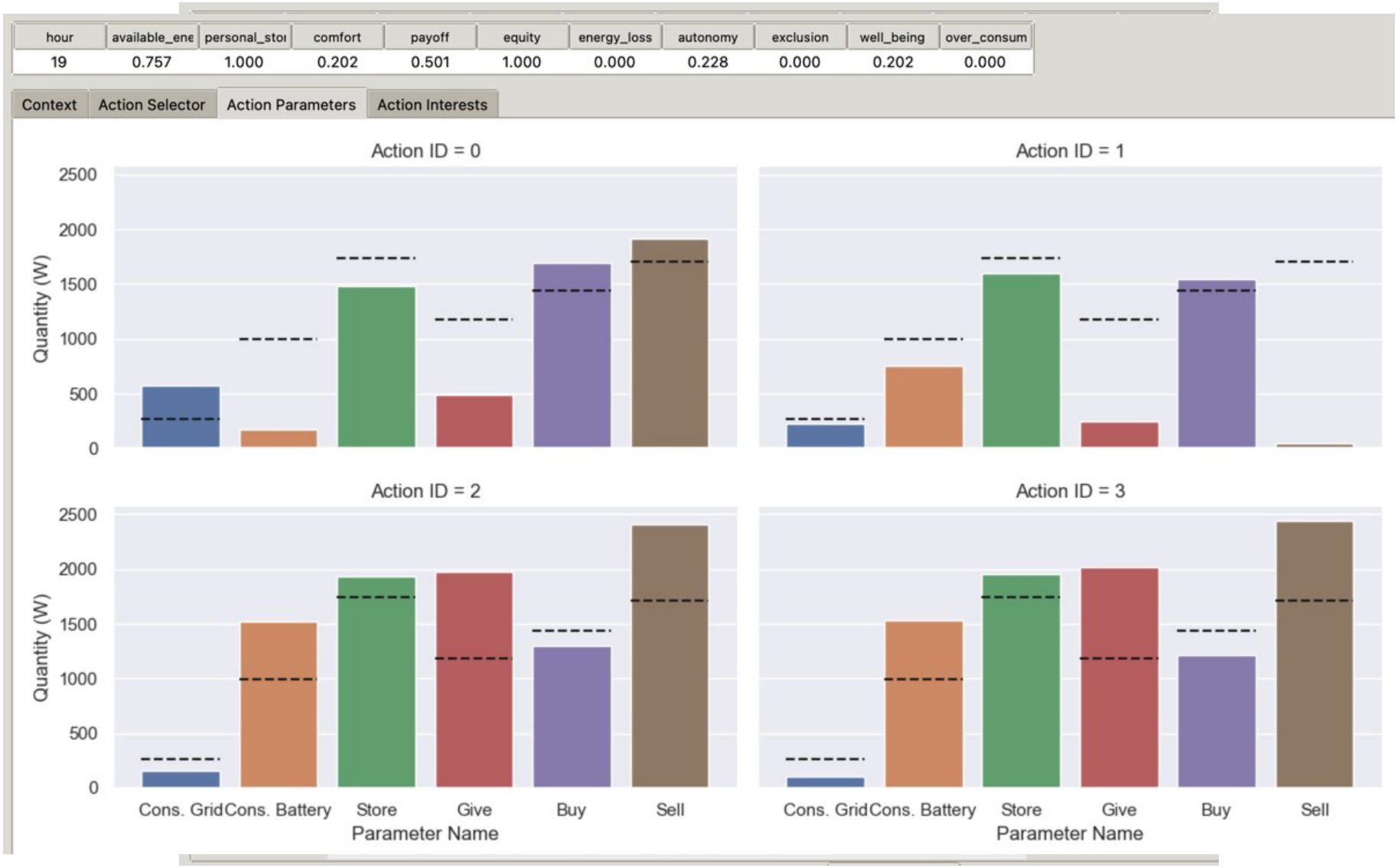
Human user interface



Human user interface



Human user interface



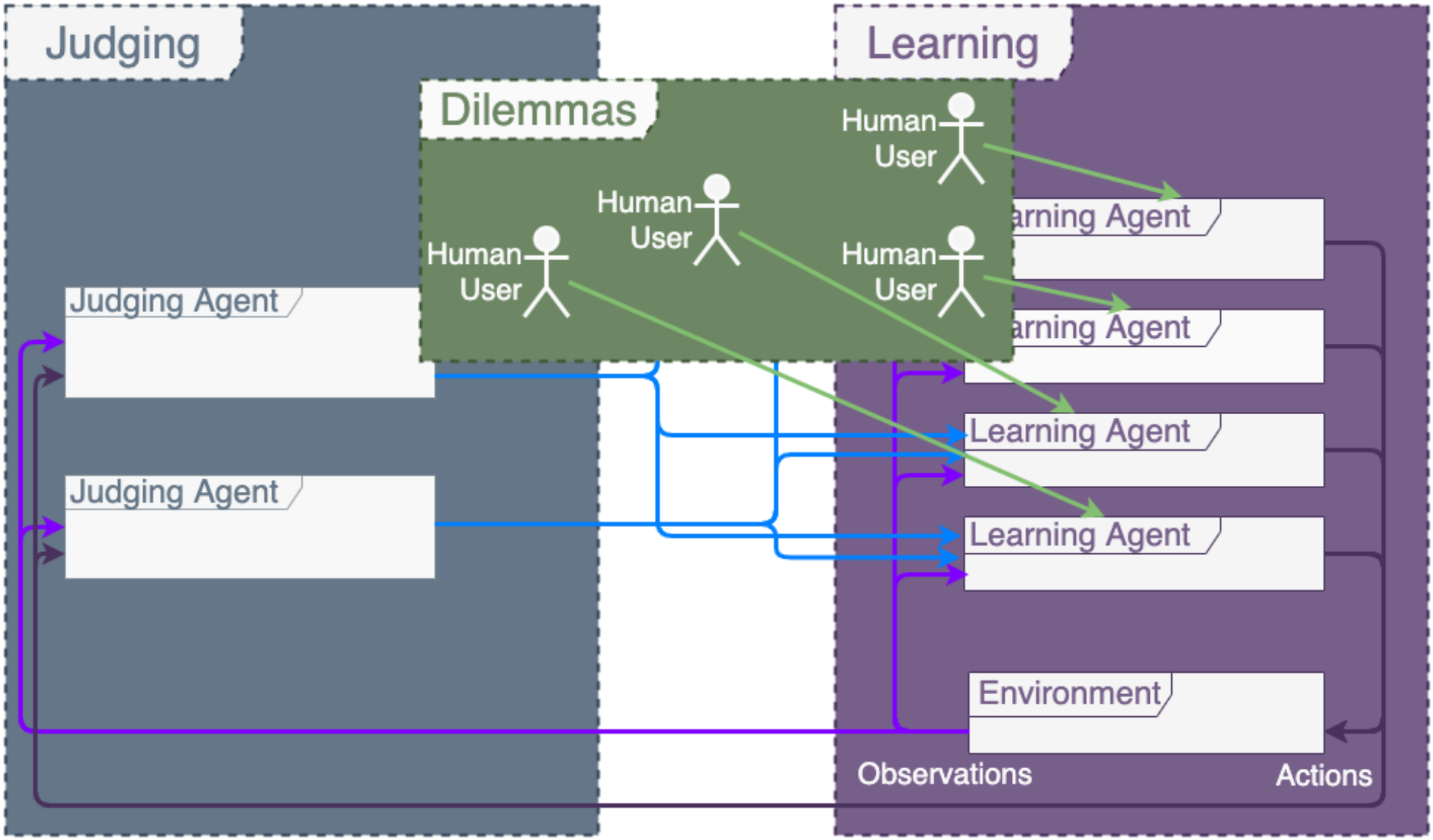
What we have achieved

Objective	Mechanism
Handle complex environments – Multiple persons	• Multi-agents • Individual observations
Handle complex environments – Multiple values	• Multiple judging agents
Handle complex environments – Multiple situations	• (D)SOMs
Adapt to shifting ethical consensus	• Non-convergence • DSOMs • Agentified reward functions
Learn behaviours with non-dilemma situations	• Learning algorithms
Specify desired behaviour	• Symbolic judgments
Learn behaviours with dilemma situations	• Explicit identification • Grouping into similar contexts

What we still lack

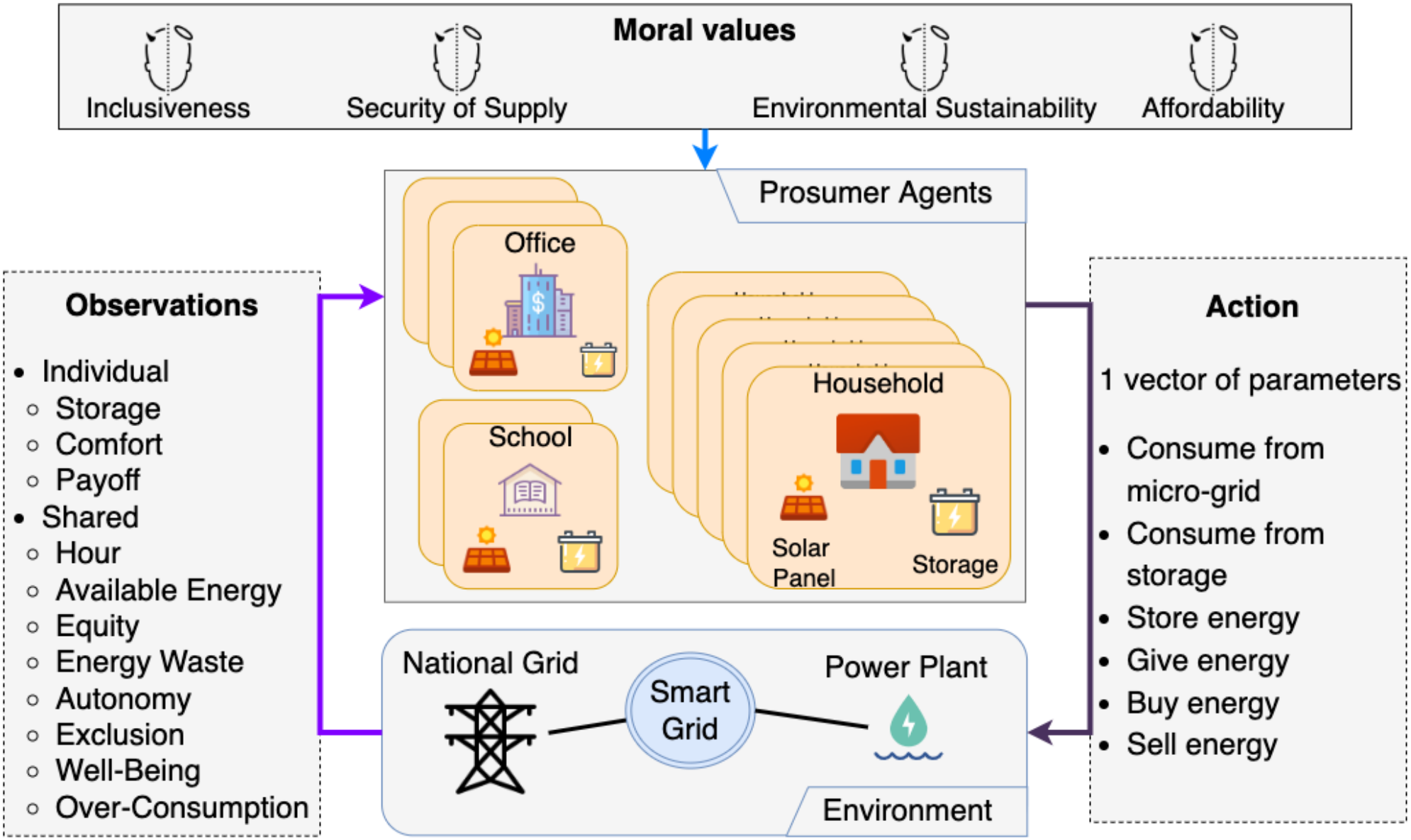
Impaired adaptability (ethical consensus shifting)

- Because of separation into bootstrap + deployment
  - Few leads, e.g., re-bootstrapping regularly
  - Potentially more intelligent mechanisms?


Judging
Learning
Dilemmas
Human User
Human User
Human User
Human User
Judging Agent
Judging Agent
arning Agent
arning Agent
Learning Agent
Learning Agent
Environment
Observations
Actions

Experiments

Smart Grid as a use-case



Different types of agents

Household

Battery capacity: 500Wh

Action range: 2,500Wh

Office

Battery capacity: 2,500Wh

Action range: 14,100Wh

School

Battery capacity: 10,000Wh

Action range: 205,000Wh

Different types of agents

Household

Battery capacity: 500Wh

Action range: 2,500Wh

Office

Battery capacity: 2,500Wh

Action range: 14,100Wh

School

Battery capacity: 10,000Wh

Action range: 205,000Wh

Different types of agents

Household

Battery capacity: 500Wh

Action range: 2,500Wh

Office

Battery capacity: 2,500Wh

Action range: 14,100Wh

School

Battery capacity: 10,000Wh

Action range: 205,000Wh

Number of agents

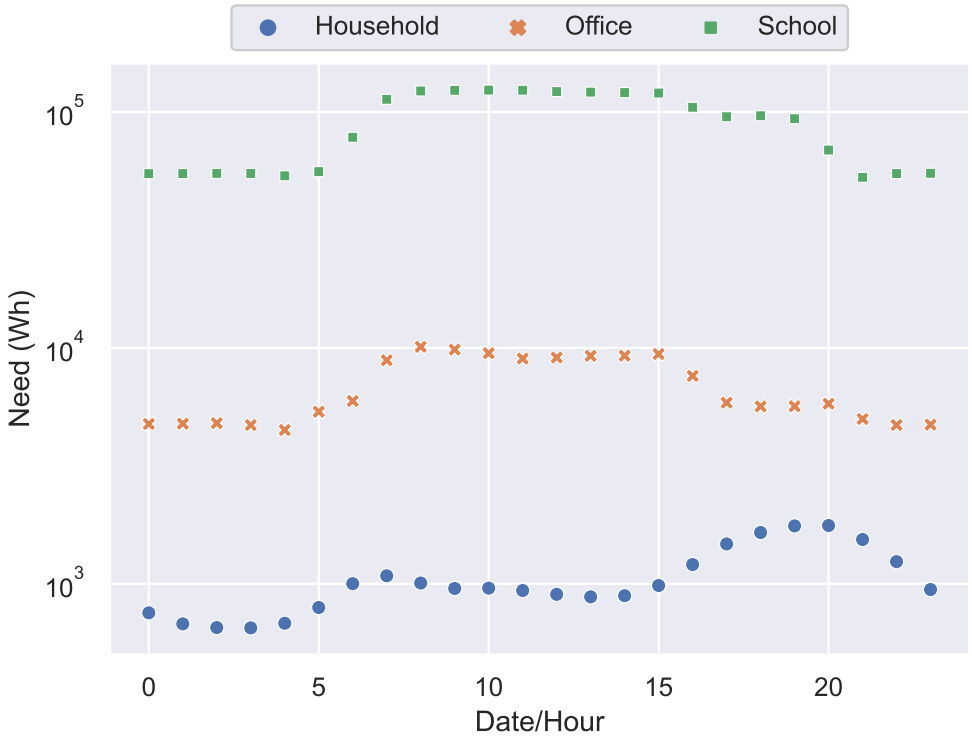
- Scenario “Small”
  - 20 Households
  - 5 Offices
  - 1 School

Number of agents

- Scenario “Small”
  - 20 Households
  - 5 Offices
  - 1 School

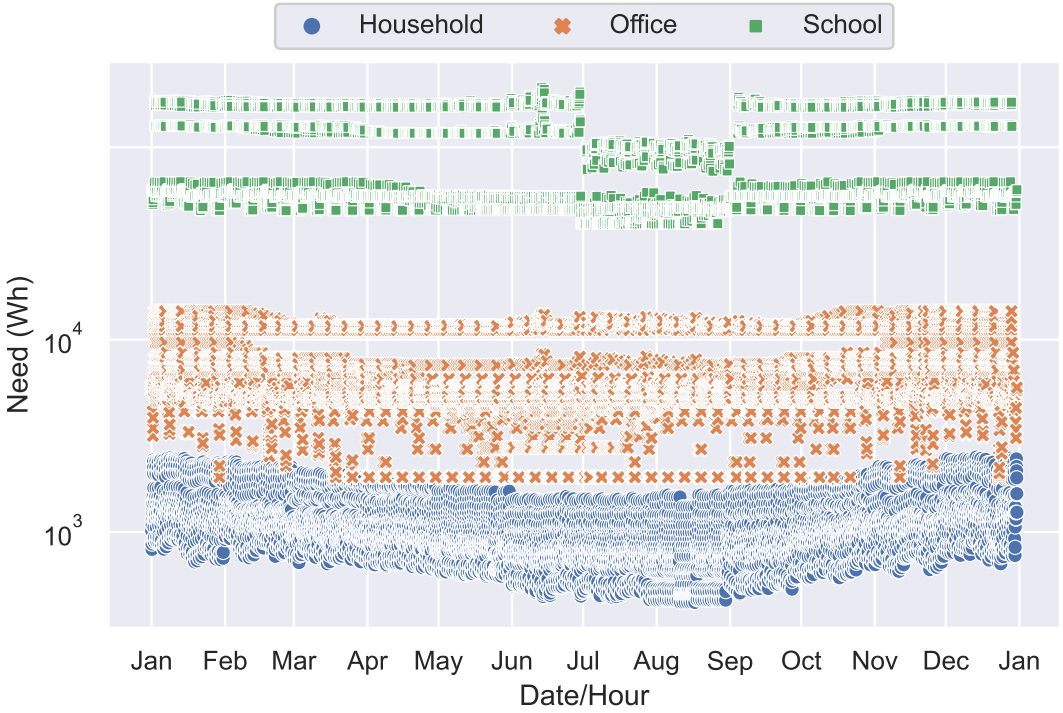
- Scenario “**Medium**”
  - 80 Households
  - 19 Offices
  - 1 School

Datasets for simulating needs

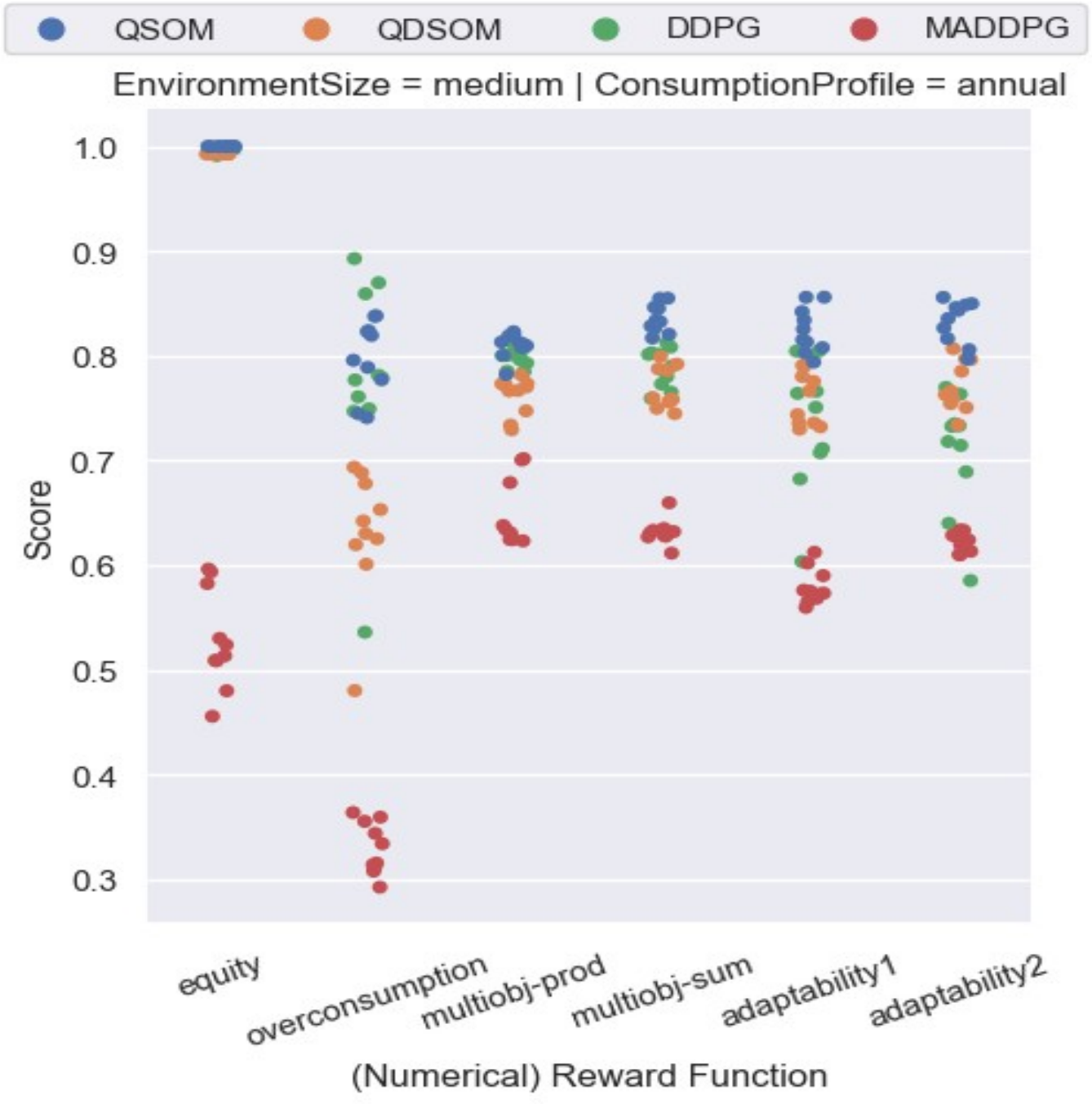


Scenario “Daily”

Scenario “**Annual**”

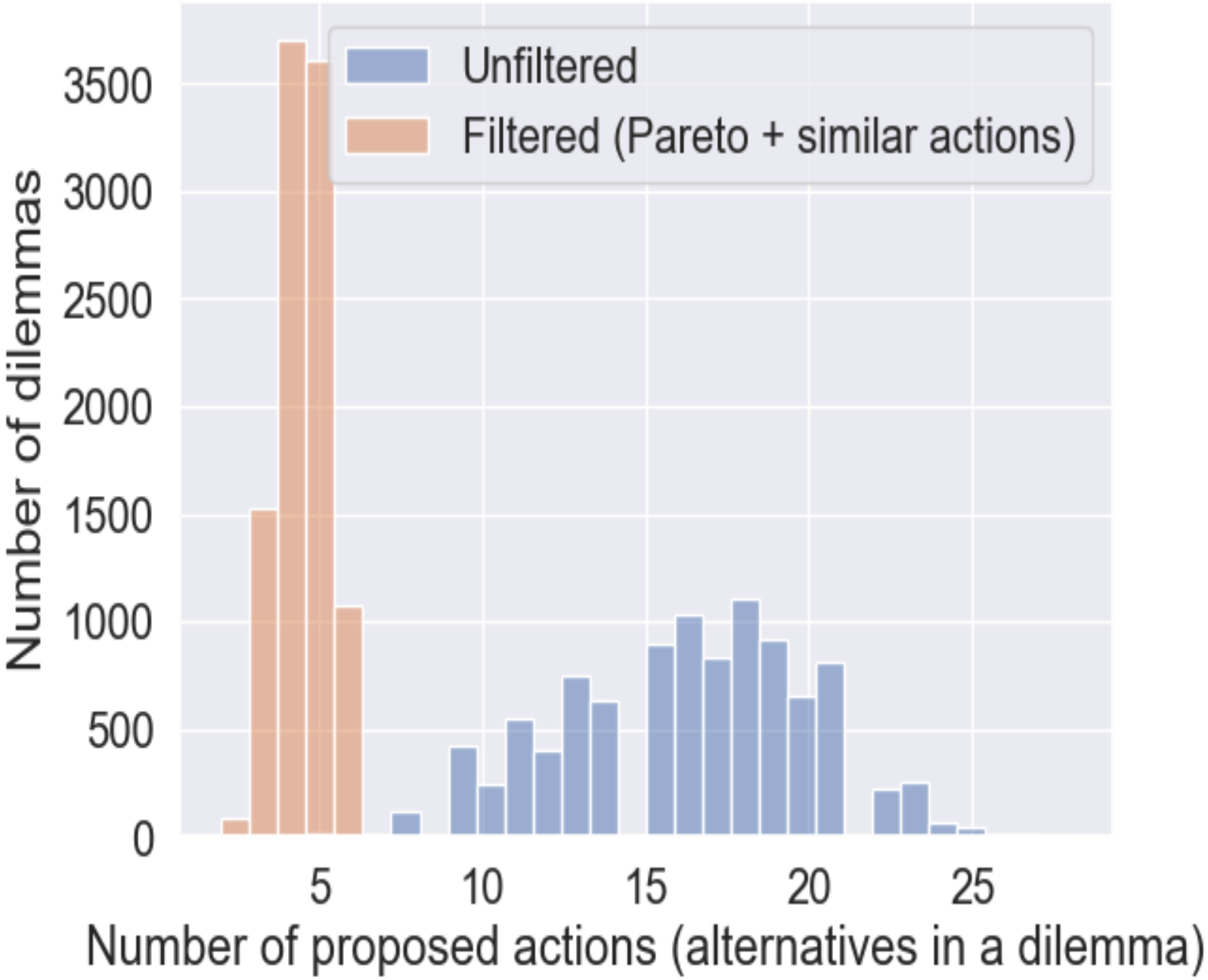


Learning – Q-SOM statistically outperforms others

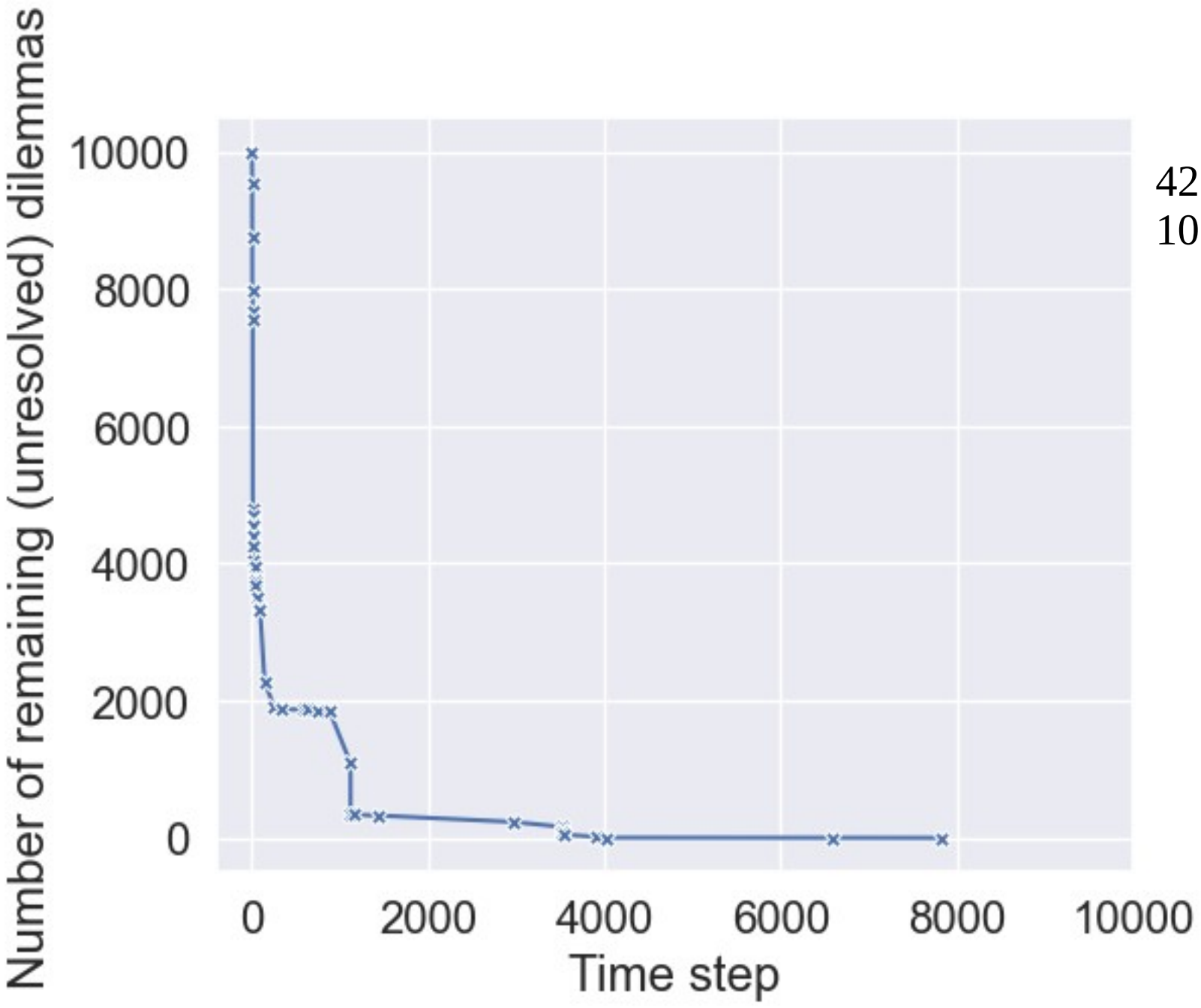


Judging – Agents adapt to multiple moral values

Dilemmas – Number of proposed actions is manageable



Number of unresolved dilemmas diminishes with time



42 interactions
10,000 dilemmas

Conclusion

Synthesis

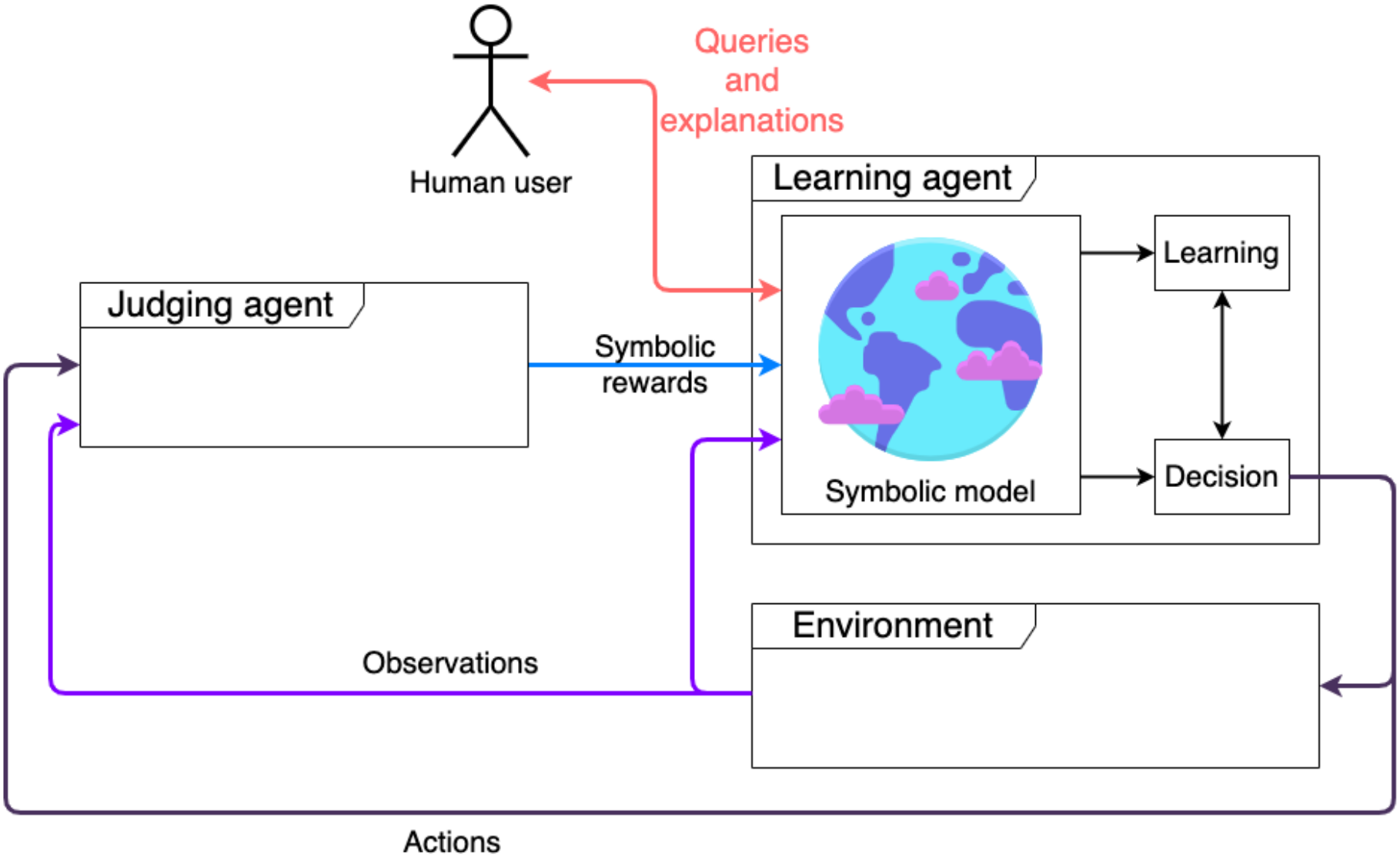
- Reinforcement Learning algorithms to learn behaviours, using continuous representations, able to adapt
- Hybrid learning through symbolic judgments in specific agents to make learning agents integrate ethical considerations
- Putting the user back in the loop, with a focus on dilemmas and contextualized preferences
- Validation on a Smart Grid simulator, with real datasets

A few limitations

- Learning could be improved
  - Multi-agent => joint-actions
  - Interesting actions => automatic partitioning of interest space

- Human preferences
  - Different definitions of preferences
  - Experiments with users

Perspective - Explainability



Perspective - Co-construction

- Human users learning from agents
  - e.g., change in preferences => which impact on society?

Perspective - Co-construction

- Human users learning from agents
  - e.g., change in preferences => which impact on society?

- Judging agents with long-term reasoning over learning agents
  - e.g., purposefully returning a lower reward because they know they could have done better

Perspective - Co-construction

- Human users learning from agents
  - e.g., change in preferences => which impact on society?

- Judging agents with long-term reasoning over learning agents
  - e.g., purposefully returning a lower reward because they know they could have done better

- Meta-judge selecting which moral values to take into account
  - e.g., automated curriculum learning

Thank you for your attention

References

- Amgoud, Leila, and Henri Prade. “Using Arguments for Making and Explaining Decisions.” Artificial Intelligence 173, no. 3 (2009): 413–36.
- Anderson, Michael, and Susan Leigh Anderson. Machine Ethics. Cambridge University Press, 2011.
- Anderson, Michael, Susan Leigh Anderson, and Vincent Berenz. “A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm.” Proceedings of the IEEE 107, no. 3 (2019): 526–40.
- Bernstein, Daniel S., Robert Givan, Neil Immerman, and Shlomo Zilberstein. “The Complexity of Decentralized Control of Markov Decision Processes.” Mathematics of Operations Research 27, no. 4 (2002): 819–40.
- Bremner, Paul, Louise A. Dennis, Michael Fisher, and Alan F. Winfield. “On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots.” Proceedings of the IEEE 107, no. 3 (2019): 541–61.

References

- Cointe, Nicolas, Grégory Bonnet, and Olivier Boissier. “Ethical Judgment of Agents’ Behaviors in Multi-Agent Systems.” In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, 1106–14. AAMAS ’16. Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems, 2016.

- Dignum, Virginia. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, Artificial Intelligence: Foundations, Theory, and Algorithms (Springer International Publishing, 2019).

- Kohonen, Teuvo. “The Self-Organizing Map.” Proceedings of the IEEE 78, no. 9 (1990): 1464–80.

- Sutton, Richard S, and Andrew G Barto. Reinforcement Learning: An Introduction. 2nd ed. MIT Press, 2018.

- Watkins, Christopher J. C. H., and Peter Dayan. “Q-Learning.” Machine Learning 8, no. 3 (1992): 279–92.

References

- Wu, Yueh-Hua, and Shou-De Lin. “A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents.” ArXiv:1712.04172 [Cs], 2017.