MACHINE ETHICS AND NORMATIVE SYSTEMS

TOWARDS USER IN THE LOOP

Rémy Chaput, PhD

2023/02/21

Seminar of the Individual and Collective Reasoning Group — University of Luxembourg

https://rchaput.github.io/talk/icr-lu-2023/

CONTEXT

What is Machine Ethics?

Why do we care?

INCREASING NUMBER OF DEPLOYED AI SYSTEMS

- Examples: loan decisions; automatic hiring; …
- Impact on human (daily) lives
- ⇒ Several concerns from society
  - **Ethical considerations**
  - Explainability
  - Trust
  - …

WHAT IS MACHINE ETHICS

- Incorporating algorithmic capabilities for ethical decision-making
- Artificial agents able to reason about norms and values
- Learning behaviours that are aligned with human values

Related to Dignum's "*Ethics By Design*"

Dignum, Virginia. Responsible artificial intelligence: how to develop and use AI in a responsible way. Cham: Springer, 2019.

MACHINE ETHICS AND NORMATIVE SYSTEMS

A brief state of the art

TOP-DOWN, BOTTOM-UP, AND HYBRID APPROACHES

- **Top-down**
  - Formalizing existing ethical principles
  - E.g., Kant's Categorical Imperative, Aquinas' Doctrine of Double Effect, …
  - ⇒ Symbols and normative systems
  - Great for including expert knowledge, ensuring that the system remains within bounds
  - But more difficult to adapt to new, unknown, or conflictual situations

Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial morality: Top-down, bottom-up, and hybrid approaches." Ethics and

TOP-DOWN, BOTTOM-UP, AND HYBRID APPROACHES

- **Bottom-up**
  - Learning a new principle from interactions
  - E.g., supervised learning, reinforcement learning (RL), and inverse RL
  - ⇒ Learning systems
  - Great for adapting to specific data (different cultures)
  - But harder to explore / assess the learned principle

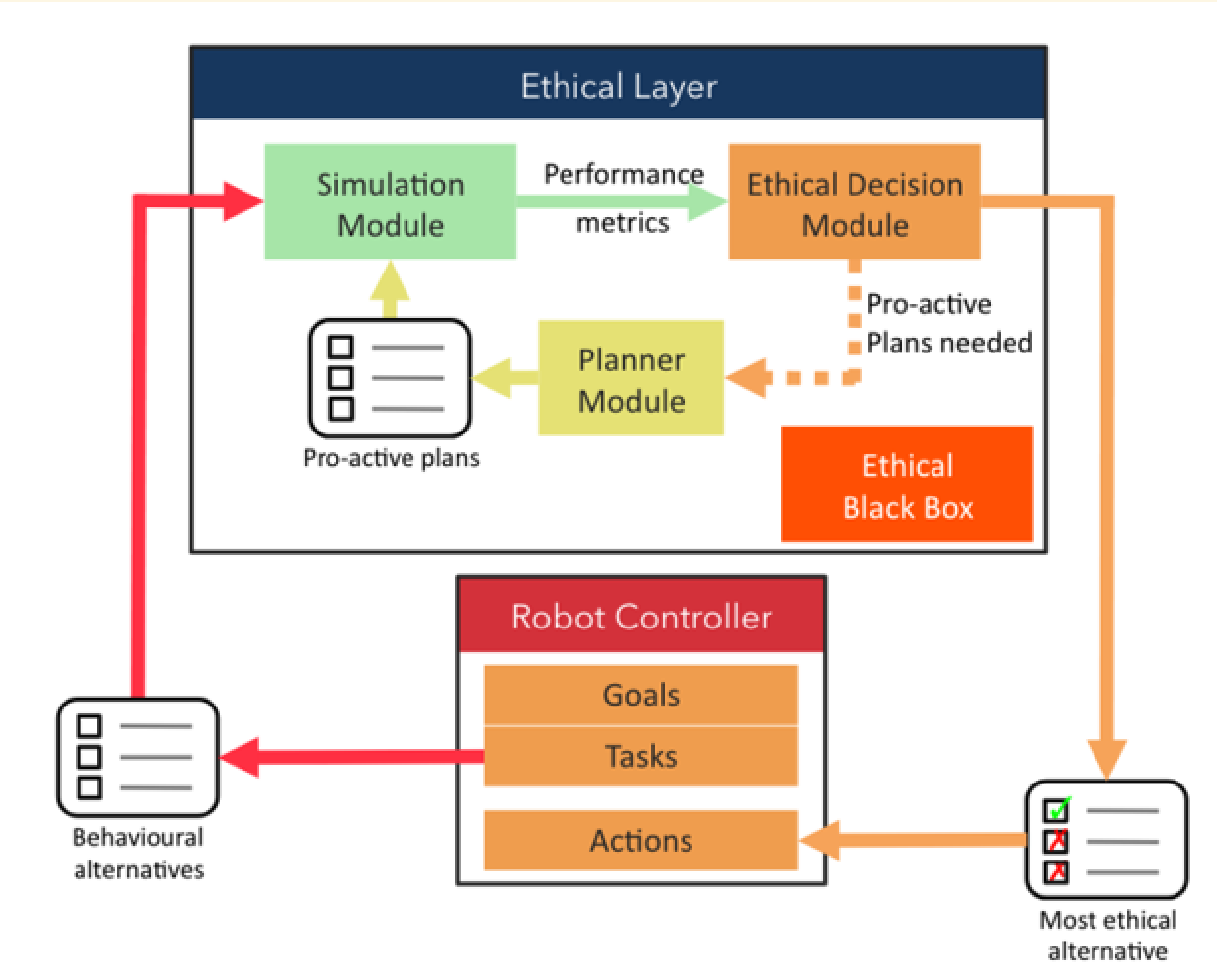
Allen, Colin, Iva Smit, and Wendell Wallach. “Artificial morality: Top-down, bottom-up,

TOP-DOWN, BOTTOM-UP, AND HYBRID APPROACHES

- **Hybrid**
  - Combines advantages of both **Top-down** and **Bottom-up** approaches
  - E.g., learning constrained by norms

Allen, Colin, Iva Smit, and Wendell Wallach. “Artificial morality: Top-down, bottom-up,

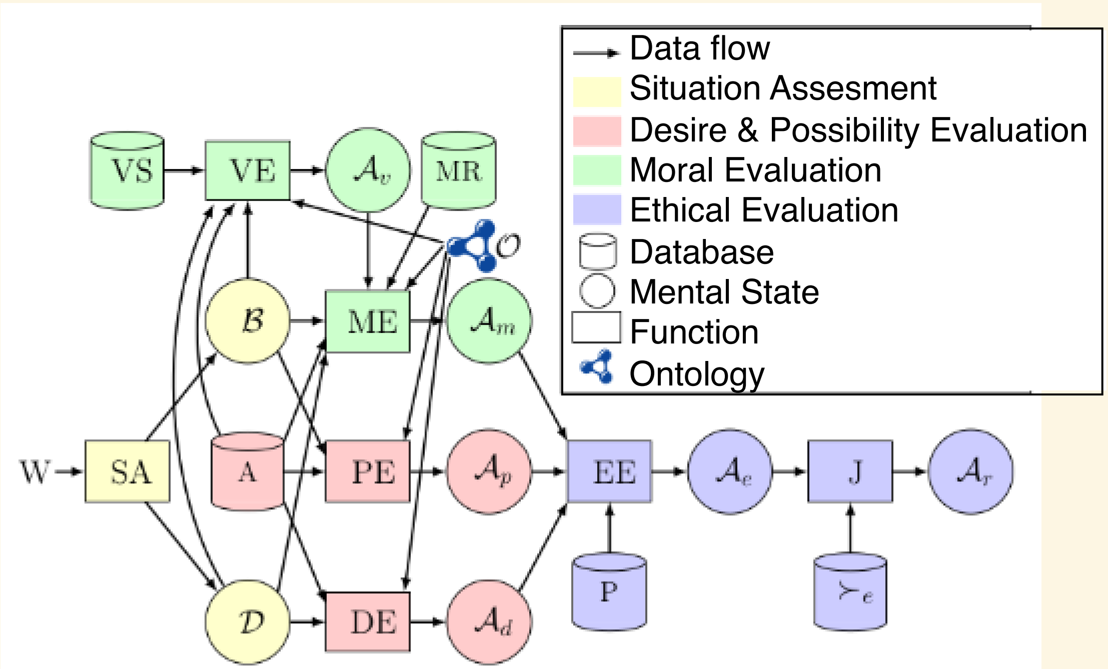
EXAMPLE: ETHICAL LAYER



Bremner, Paul, et al. “On proactive, transparent, and verifiable ethical reasoning for

EXAMPLE: ETHICAA

Principles priority



Cointe, Nicolas, Grégory Bonnet, and Olivier Boissier. "Ethical Judgment of Agents'

ARGUMENTATION FOR JUDGMENT

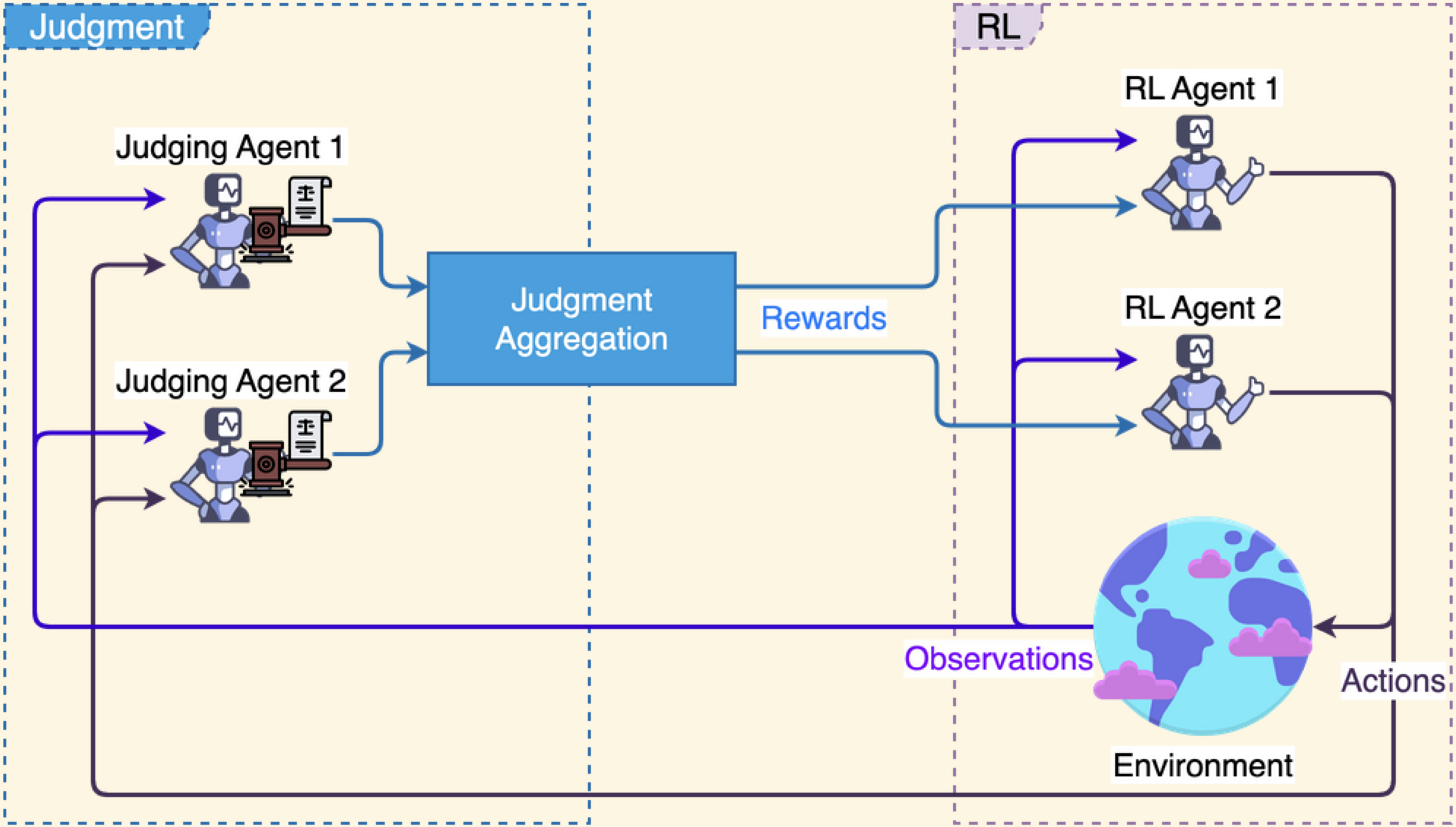
The AJAR framework

OUR IDEA

- We do not know the correct action, but we can judge an action
- RL is great for learning behaviours based on a reward signal
- Argumentation is great to specify what we want

⇒ Why not combining them?

CONCEPTUAL ARCHITECTURE



ARGUMENTATION FRAMEWORK FOR JUDGING A DECISION

We define an AFJD as a graph AF containing:

- Arguments $\mathrm{AF}_{[\mathrm{Args}]}$ (nodes)
- Attack relationship $\mathrm{AF}_{[\mathrm{Att}]}$ between arguments (edges)
- Set of *pro*-arguments $\mathrm{AF}_{[\mathrm{F_p}]}$
- Set of *con*-arguments $\mathrm{AF}_{[\mathrm{F_c}]}$

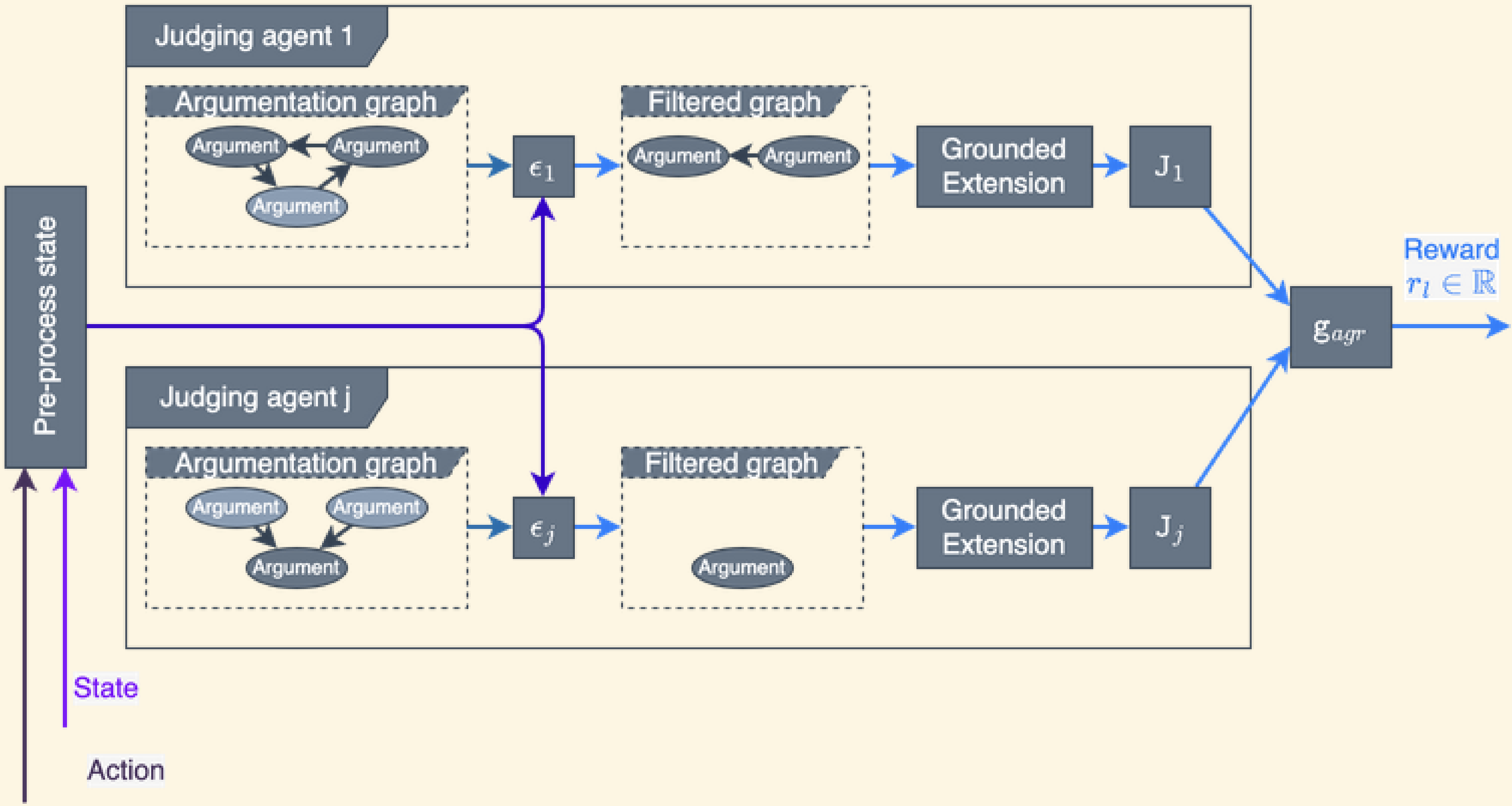
JUDGING AGENTS

We define a judging agent as a tuple:

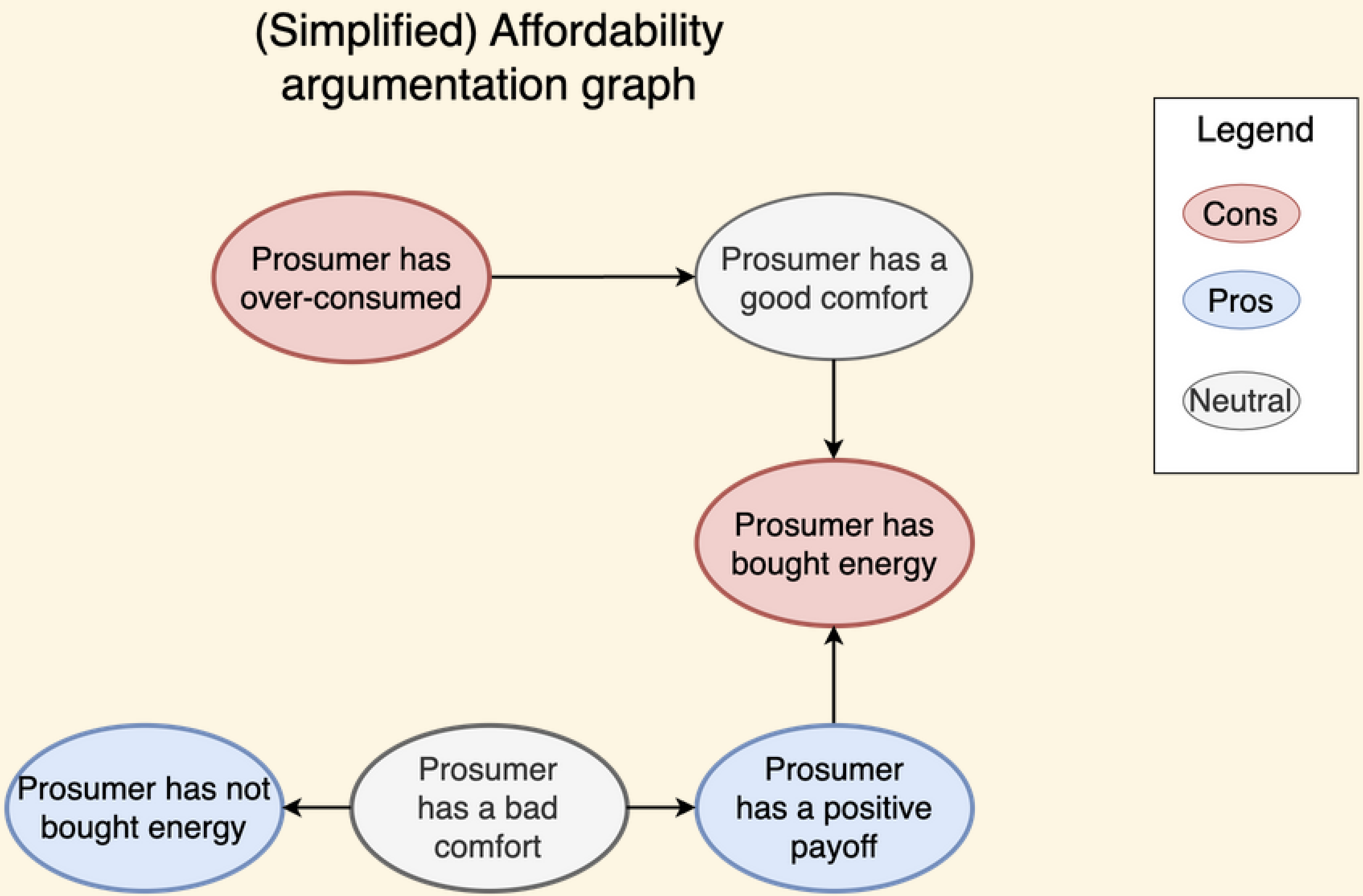
- A moral value
- An AFJD (graph with pros and cons)
- A filtering function ϵ
- A $\mathtt{grd}$ function to compute the *grounded* extension
- A judgment function $\mathrm{J} : \mathrm{AFJD} \to \mathbb{R}$, e.g.,

$$\mathrm{J}(\mathrm{AF}) = \frac{|\mathrm{pros} \in \mathtt{grd}(\mathrm{AF}_{[\mathrm{Args}]})|}{|\mathrm{pros} \in \mathtt{grd}(\mathrm{AF}_{[\mathrm{Args}]})| + |\mathrm{cons} \in \mathtt{grd}(\mathrm{AF}_{[\mathrm{Args}]})|}$$

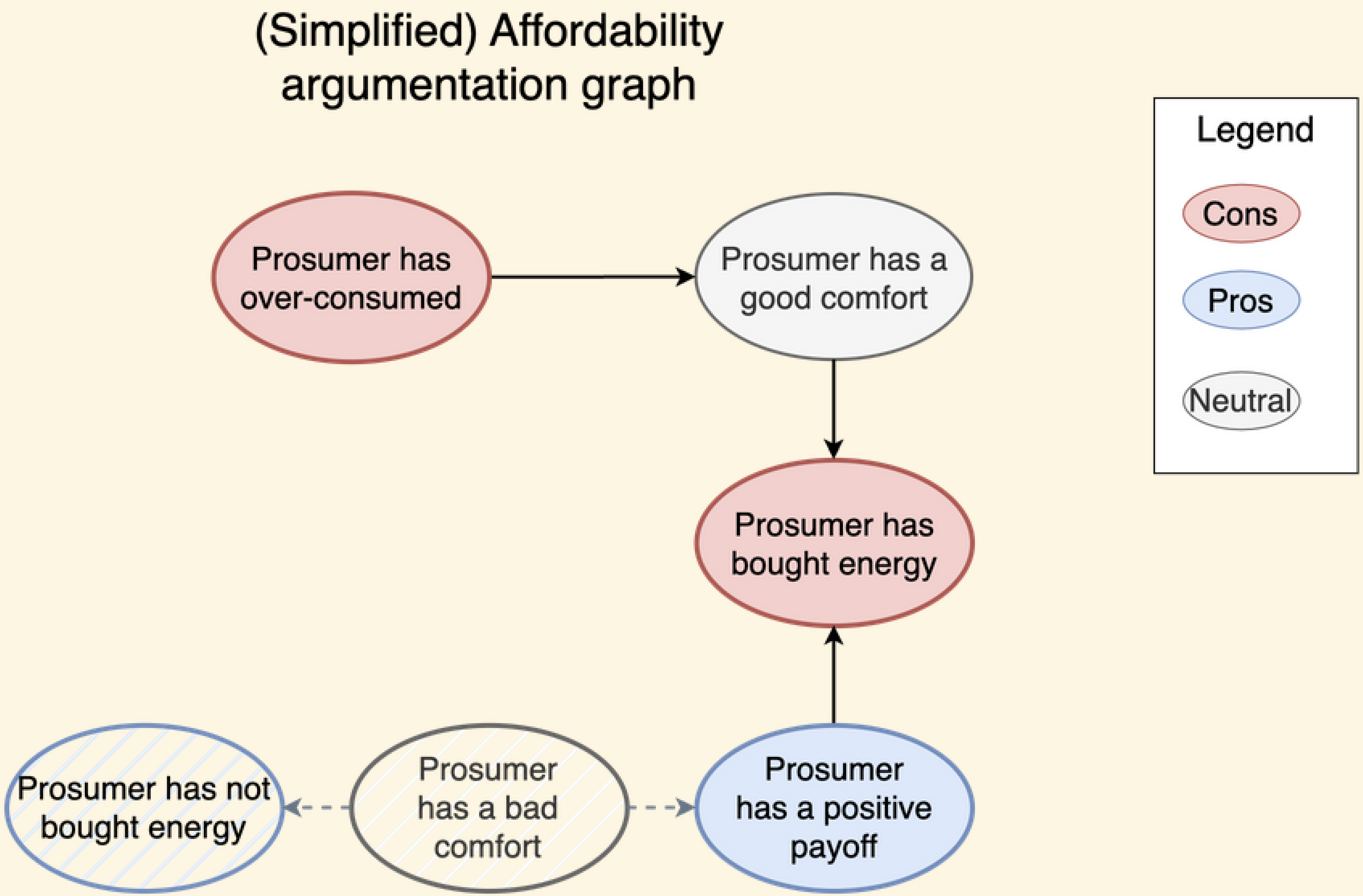
FINAL ARCHITECTURE



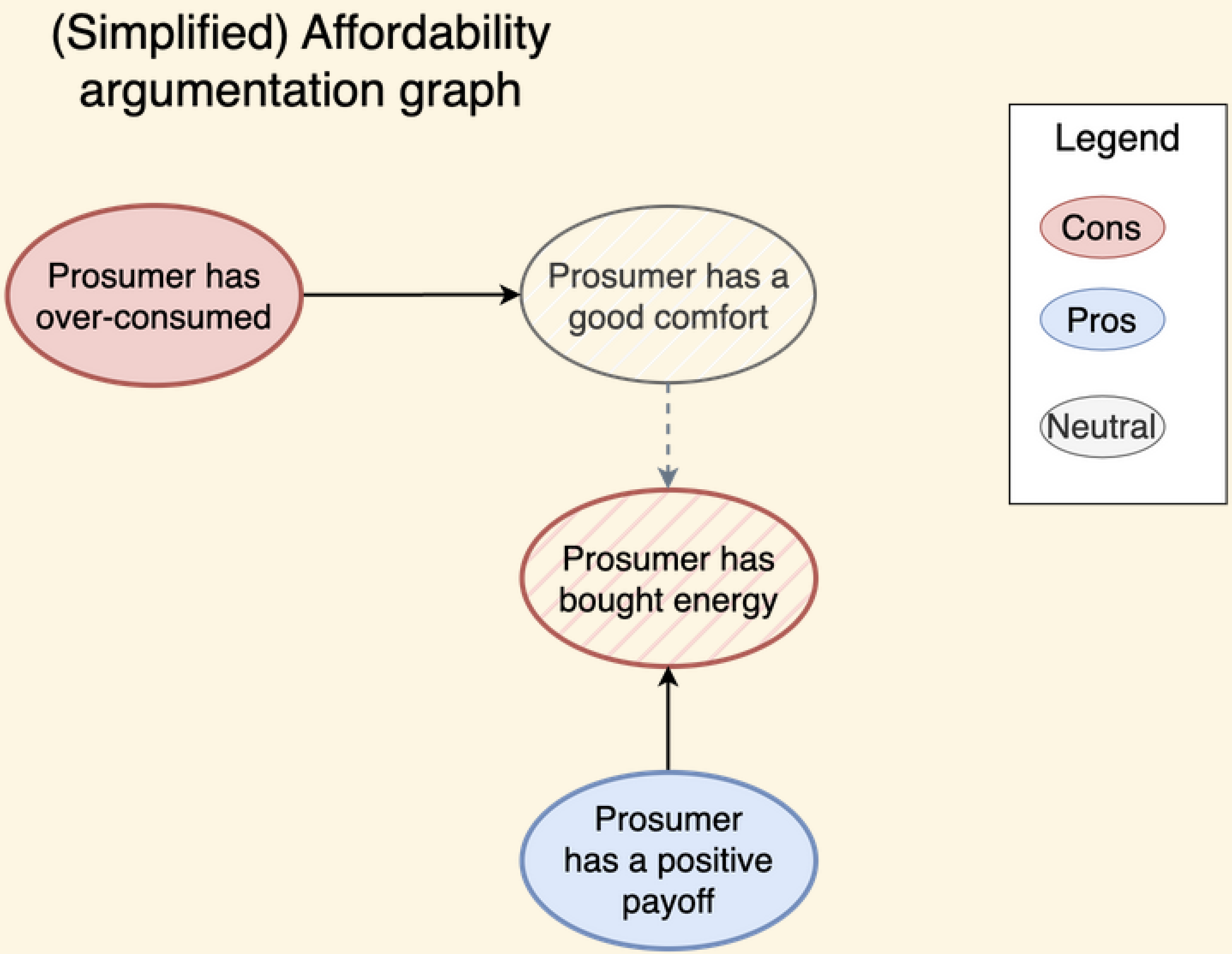
EXAMPLE OF JUDGMENT



EXAMPLE OF JUDGMENT

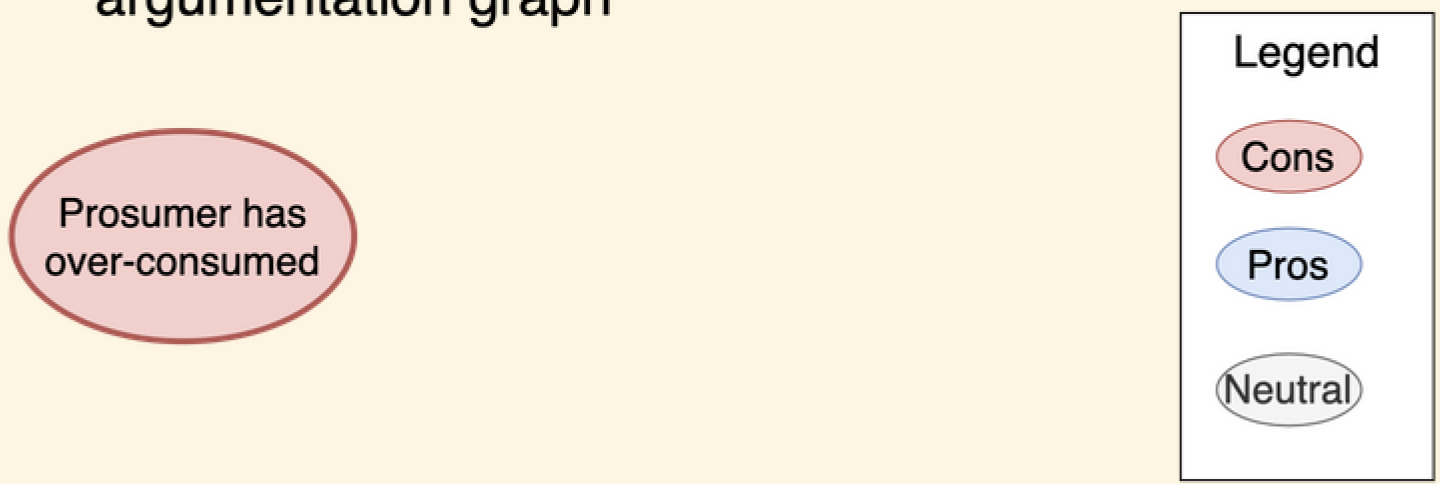


EXAMPLE OF JUDGMENT

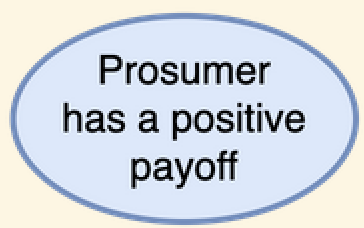


EXAMPLE OF JUDGMENT

(Simplified) Affordability argumentation graph



$$\frac{\#Pros}{\#Pros+\#Cons} = \frac{1}{2}$$



ADVANTAGES

- Explicit multiple moral values
- Easier to communicate with non-AI experts (regulators, domain experts, users, …)
- Possibility to justify/explain why a reward was given
- Paving the way for co-construction loop

LIMITATIONS

- Same aggregation method used for all learning agents
- Aggregation $\Rightarrow$ reducing information, hiding dilemmas

TOWARD USER IN THE LOOP

Multi-Objective Reinforcement Learning and human preferences

THE IDEA

- Providing separate rewards (for each moral value)
- ⇒ Capability to compare rewards, detect situations of conflicts (dilemmas)
- ⇒ Raise dilemmas to human users (better explainability)
- ⇒ Ask them for their preferences (better alignment)
- Focus on *contextualized* preferences
  - Different human users ⇒ different preferences
  - Different situations ⇒ different preferences

IDENTIFYING DILEMMAS

- Using multiple rewards ⇒ manipulating multiple interests for each action
- ⇒ Difficult to compare!
- Examples:
  - $Q(a_1) = [3, 4, 3.5, 3]$
  - $Q(a_2) = [1, 2, 3.5, 3]$
  - $Q(a_3) = [5, 3, 2.5, 3]$
- a_2 is Pareto-dominated by a_1 ; what about a_3?
- ⇒ Provide a “theoretical max” as a reference point, and ask users what they find acceptable

ETHICAL THRESHOLDS

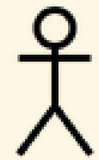
- Intuitively represent which trade-offs between moral values an user would accept
- A vector of thresholds (between 0% and 100%) for each moral value
- E.g., $\zeta_1 = [50\%, 75\%, 50\%, 60\%]$

DIFFERENT USERS RECOGNIZE DILEMMAS DIFFERENTLY

Action	Interests $Q(a_i)$	Theoreticals $Q^{th}(a_i)$	Ratio $\frac{Q(a_i)}{Q^{th}(a_i)}$
a_1	[3, 4, 3.5, 3]	[5, 5, 5, 5]	$\left[\frac{3}{5}, \frac{4}{5}, \frac{3.5}{5}, \frac{3}{5}\right]$
a_3	[5, 3, 2.5, 3]	[6, 6, 6, 6]	$\left[\frac{5}{6}, \frac{3}{6}, \frac{2.5}{6}, \frac{3}{6}\right]$

DIFFERENT USERS RECOGNIZE DILEMMAS DIFFERENTLY

Action	Interests $Q(a_i)$	Theoreticals $Q^{th}(a_i)$	Ratio $\frac{Q(a_i)}{Q^{th}(a_i)}$
a_1	[3, 4, 3.5, 3]	[5, 5, 5, 5]	$[60\%, 80\%, 70\%, 60\%]$
a_3	[5, 3, 2.5, 3]	[6, 6, 6, 6]	$[83\%, 50\%, 42\%, 50\%]$



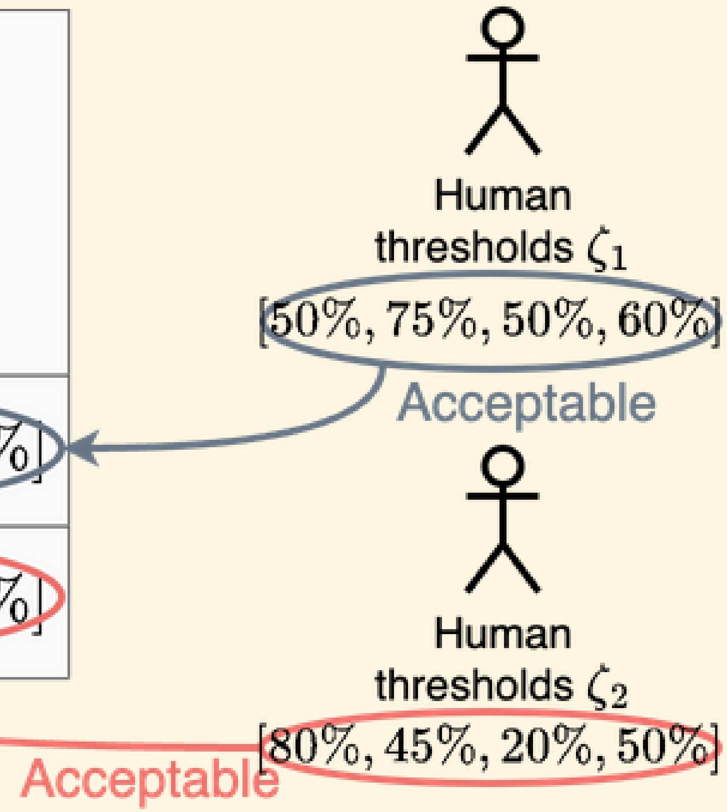
Human thresholds ζ_1

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

Acceptable

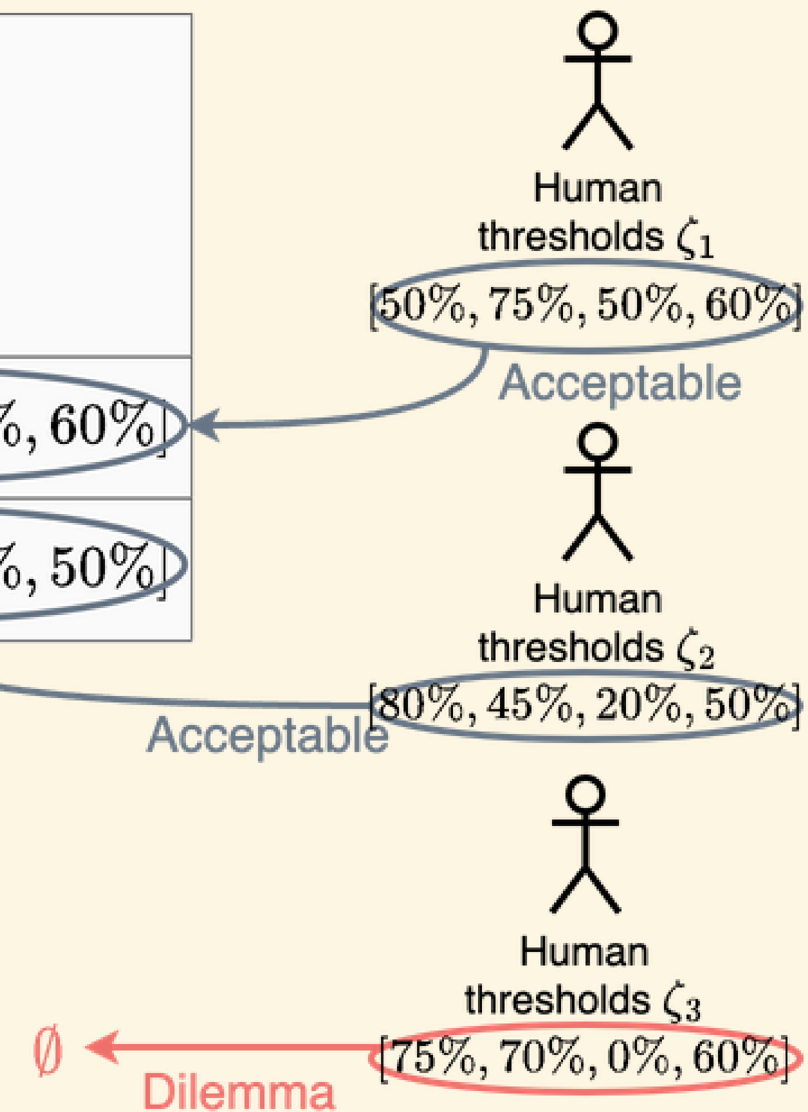
DIFFERENT USERS RECOGNIZE DILEMMAS DIFFERENTLY

Action	Interests $Q(a_i)$	Theoreticals $Q^{th}(a_i)$	Ratio $\frac{Q(a_i)}{Q^{th}(a_i)}$
a_1	[3, 4, 3.5, 3]	[5, 5, 5, 5]	$[60\%, 80\%, 70\%, 60\%]$
a_3	[5, 3, 2.5, 3]	[6, 6, 6, 6]	$[83\%, 50\%, 42\%, 50\%]$



DIFFERENT USERS RECOGNIZE DILEMMAS DIFFERENTLY

Action	Interests $Q(a_i)$	Theoreticals $Q^{th}(a_i)$	Ratio $\frac{Q(a_i)}{Q^{th}(a_i)}$
a_1	[3, 4, 3.5, 3]	[5, 5, 5, 5]	$[60\%, 80\%, 70\%, 60\%]$
a_3	[5, 3, 2.5, 3]	[6, 6, 6, 6]	$[83\%, 50\%, 42\%, 50\%]$



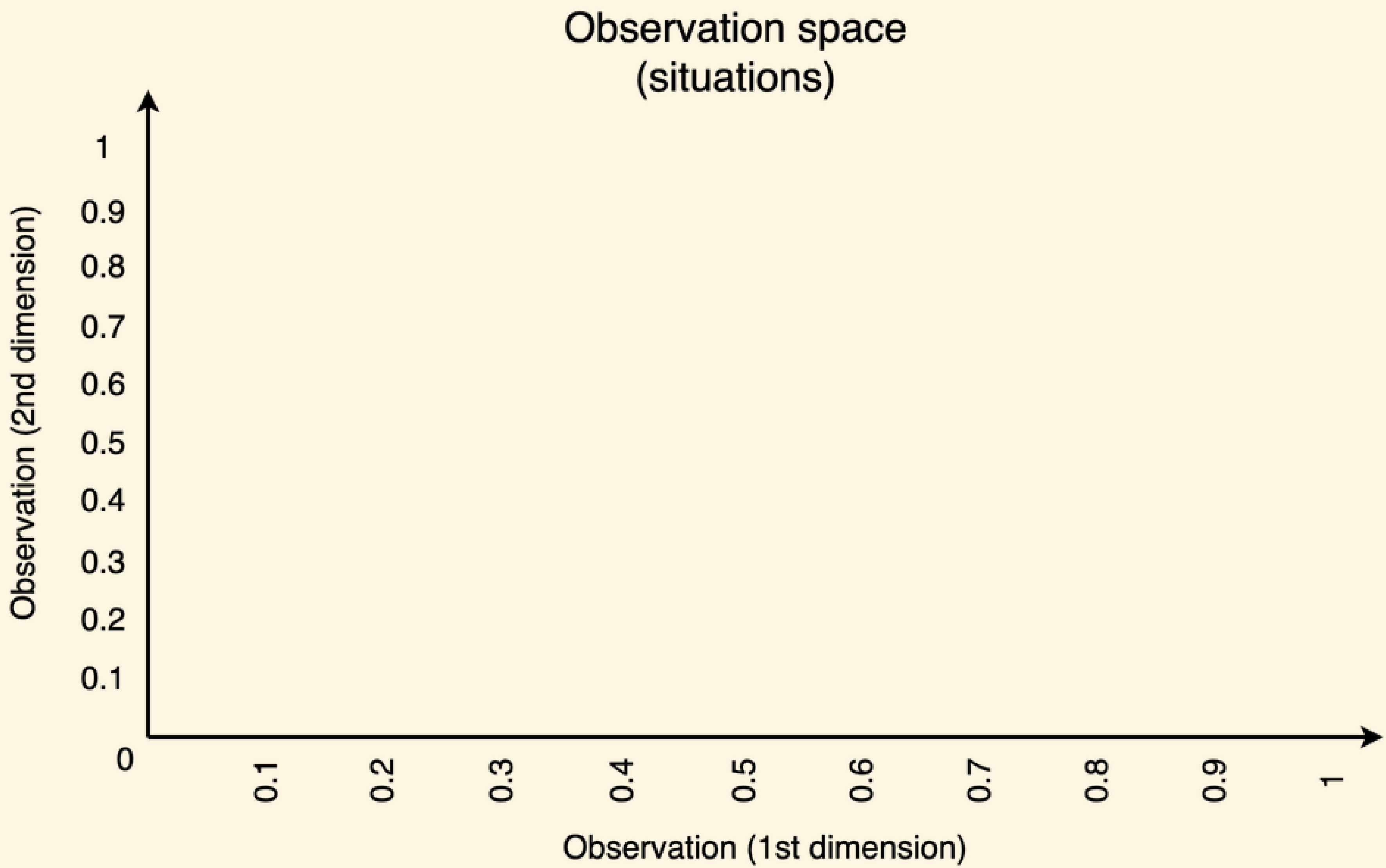
SETTLING DILEMMAS THROUGH USER PREFERENCES

- When a dilemma is identified, the agent cannot settle it autonomously
- ⇒ We ask the user what trade-off they would prefer
- Simple technique: directly select an action among the proposed ones
- Problem: the system would risk being too overwhelming if we ask each time there is a dilemma!
- ⇒ Some dilemmas might be similar, maybe we can group them

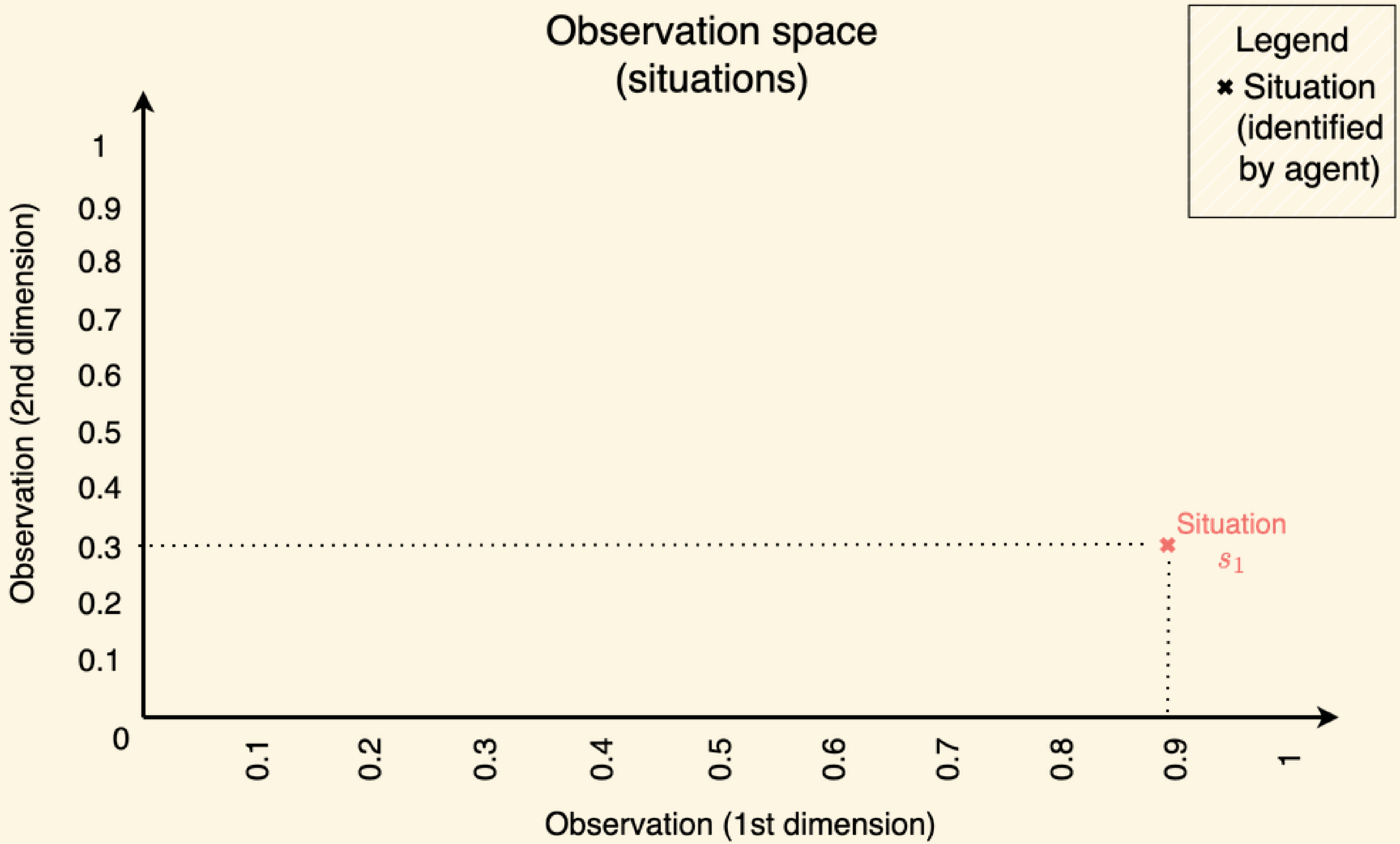
LEARNING PREFERENCES

- Dilemmas happen in *situations*
- A situation = a set of *observations* $\in \mathbb{R}$
- E.g., hour = 8, available energy = $4,000$, etc.
- We define a *context* as a set of bounds (min, max) for each observation
- E.g., $c_1 = \{\{6,9\},\{2000,5000\}\}$

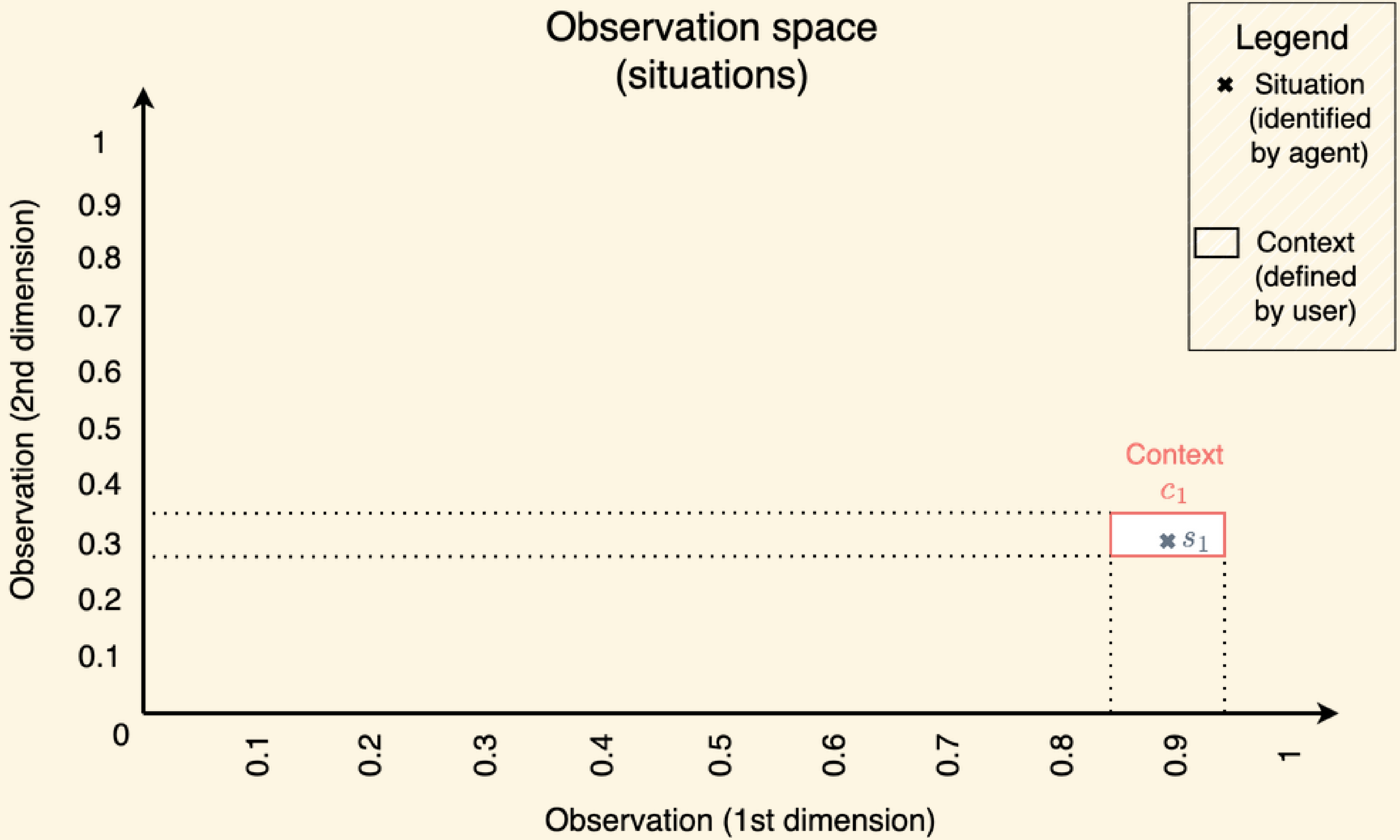
EXAMPLES OF CONTEXTS



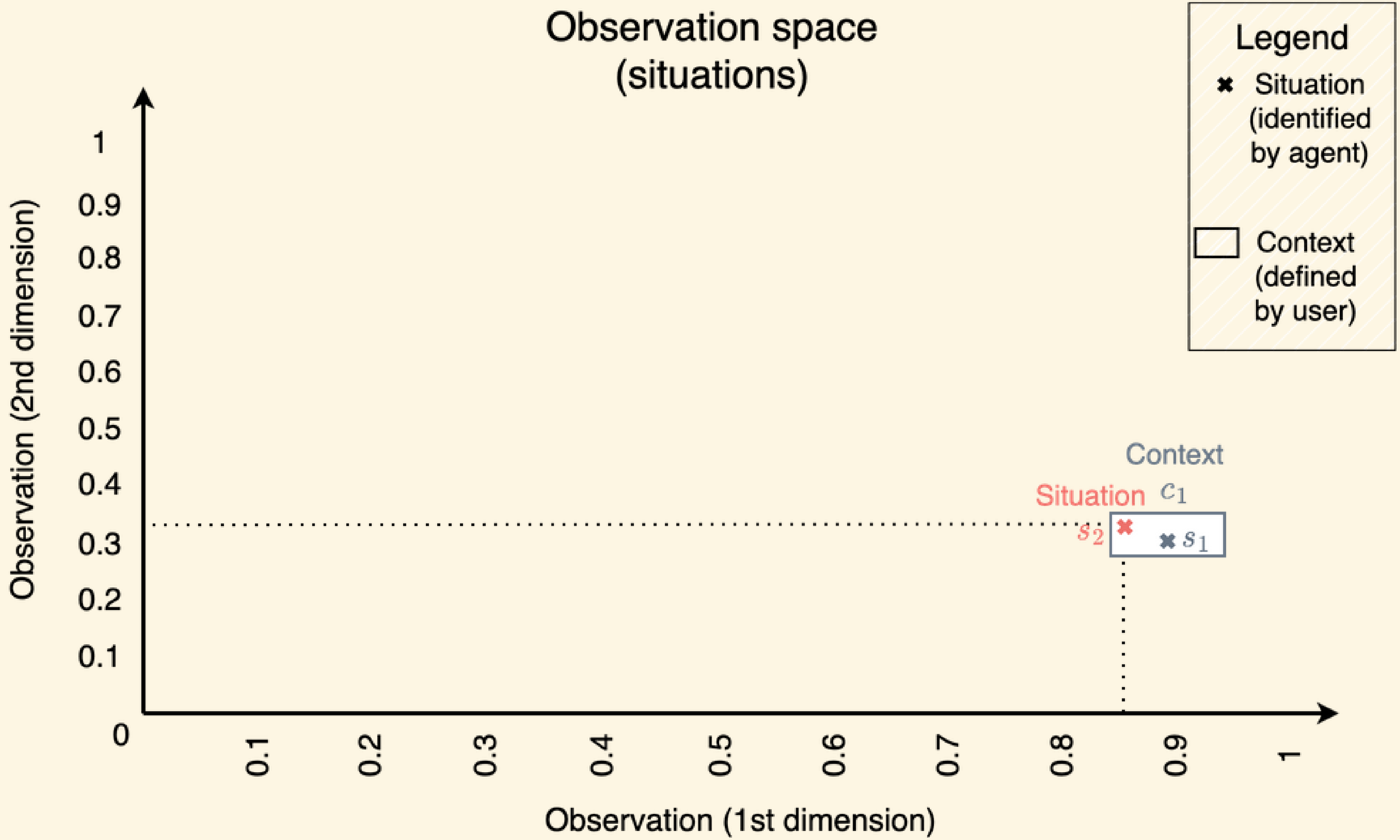
EXAMPLES OF CONTEXTS



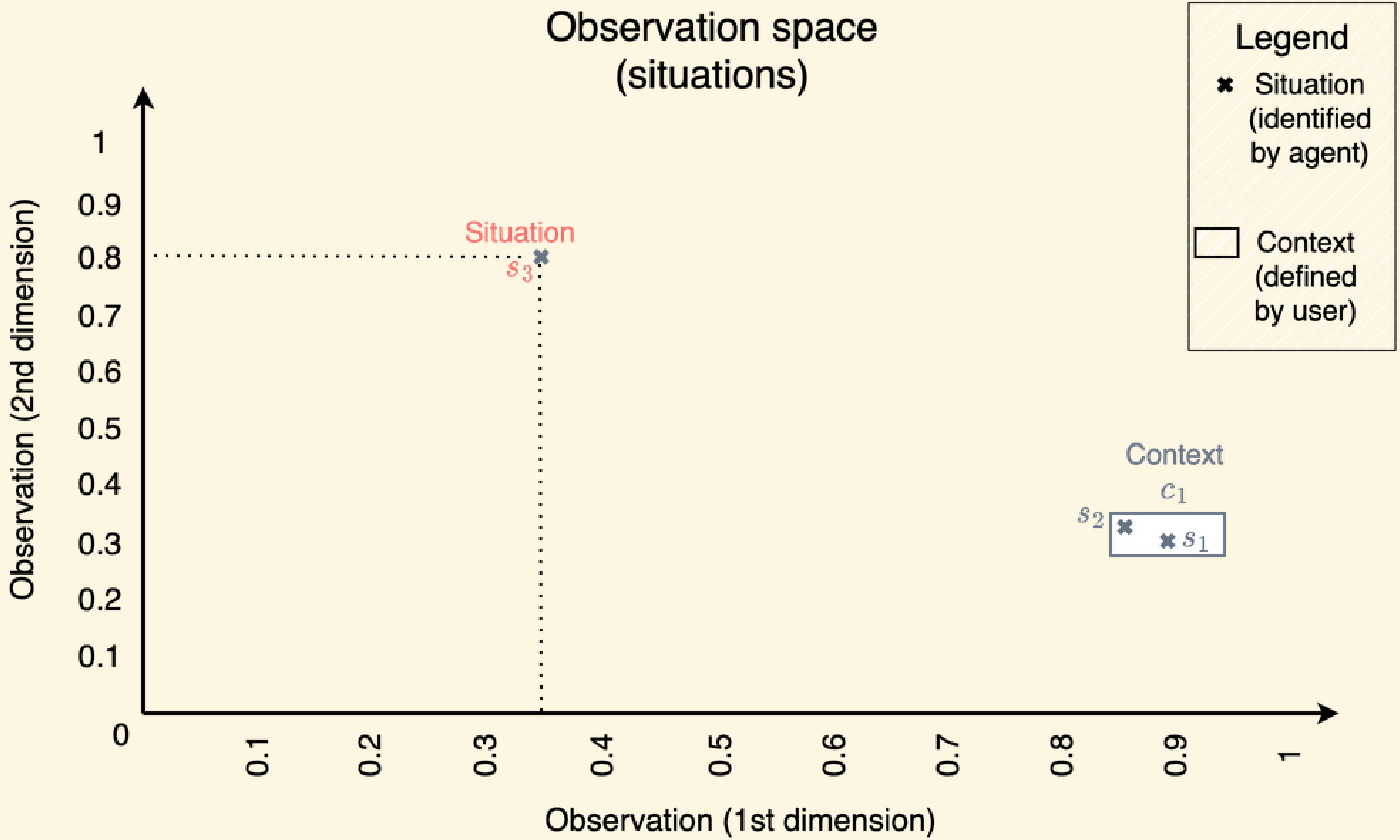
EXAMPLES OF CONTEXTS



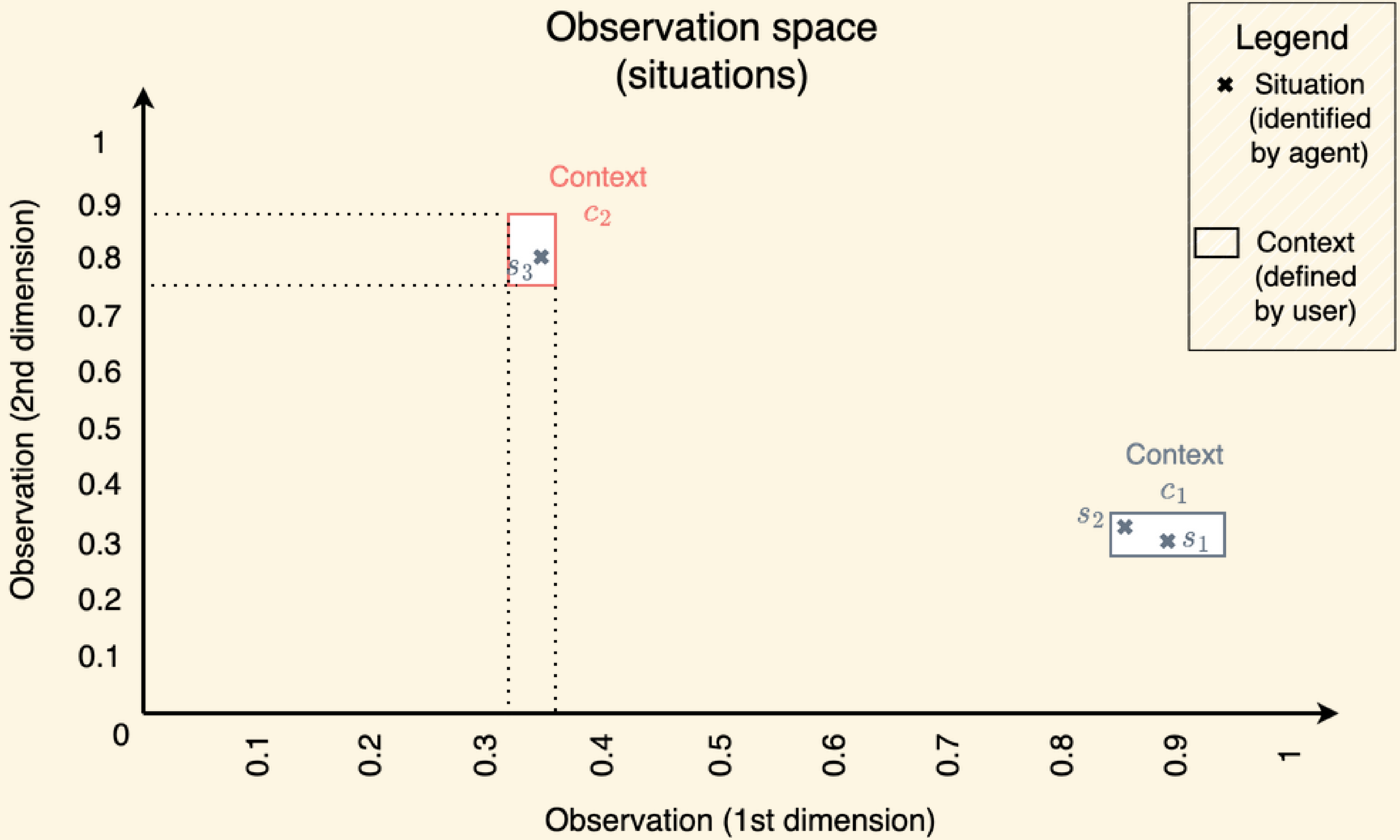
EXAMPLES OF CONTEXTS



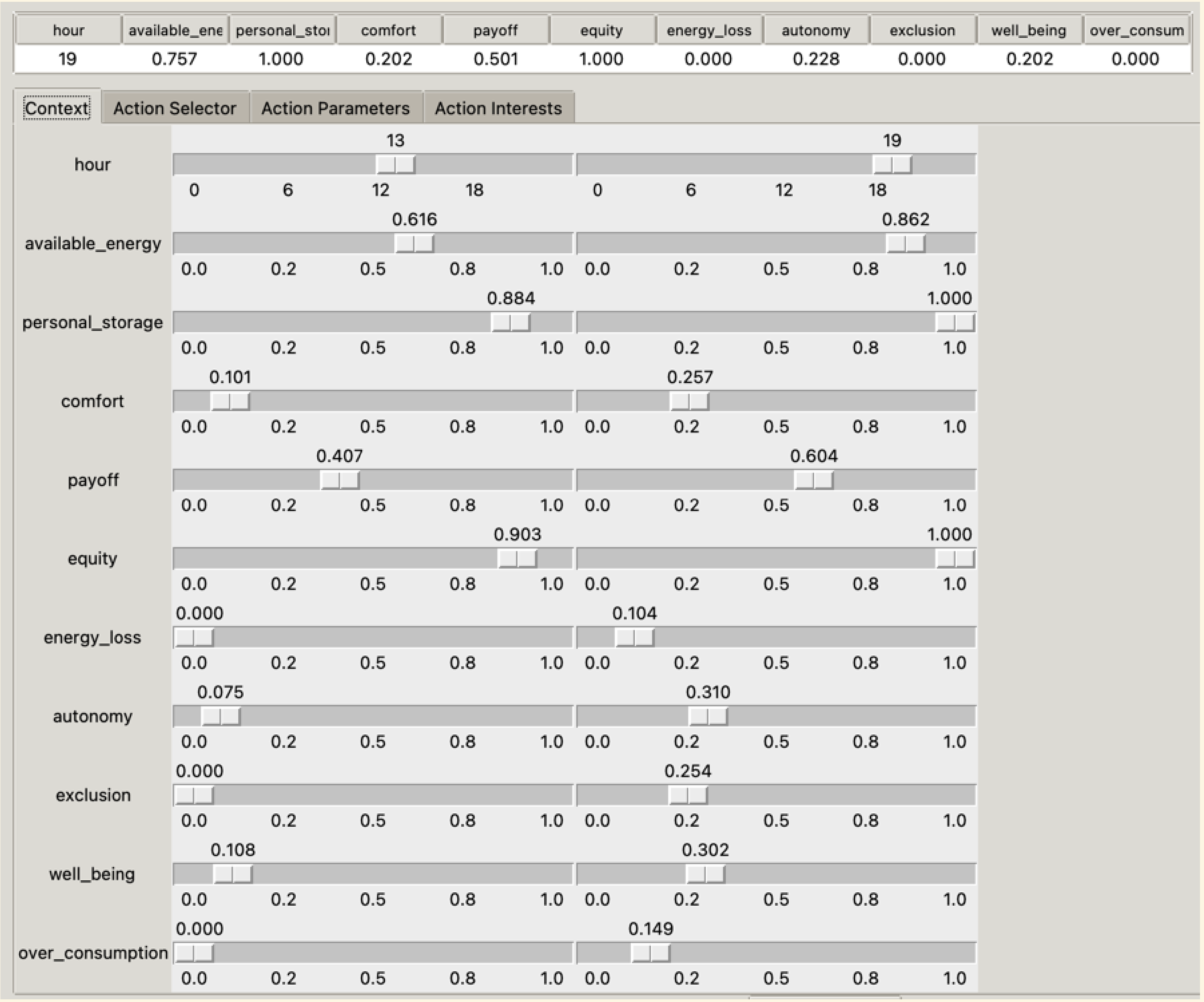
EXAMPLES OF CONTEXTS



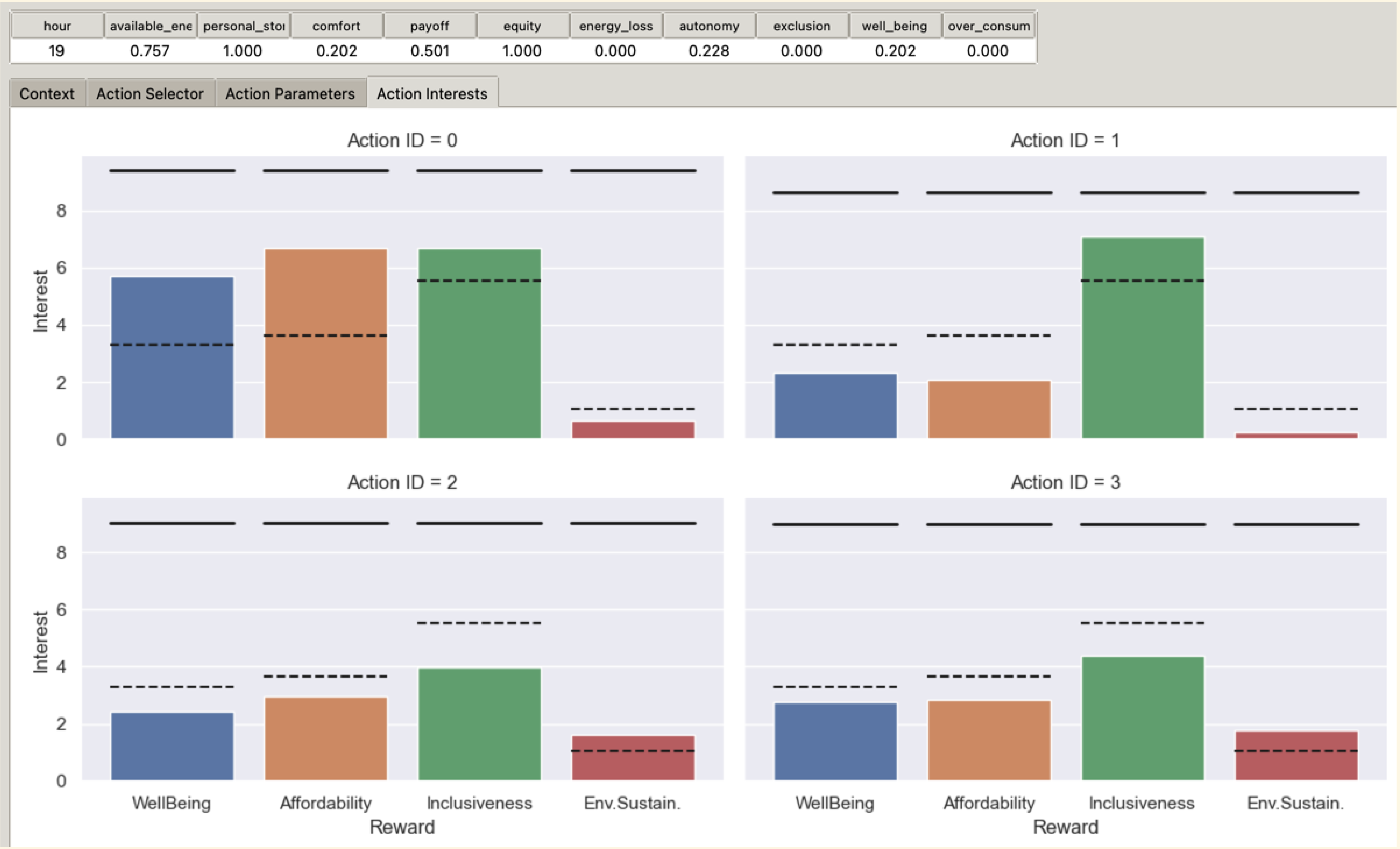
EXAMPLES OF CONTEXTS



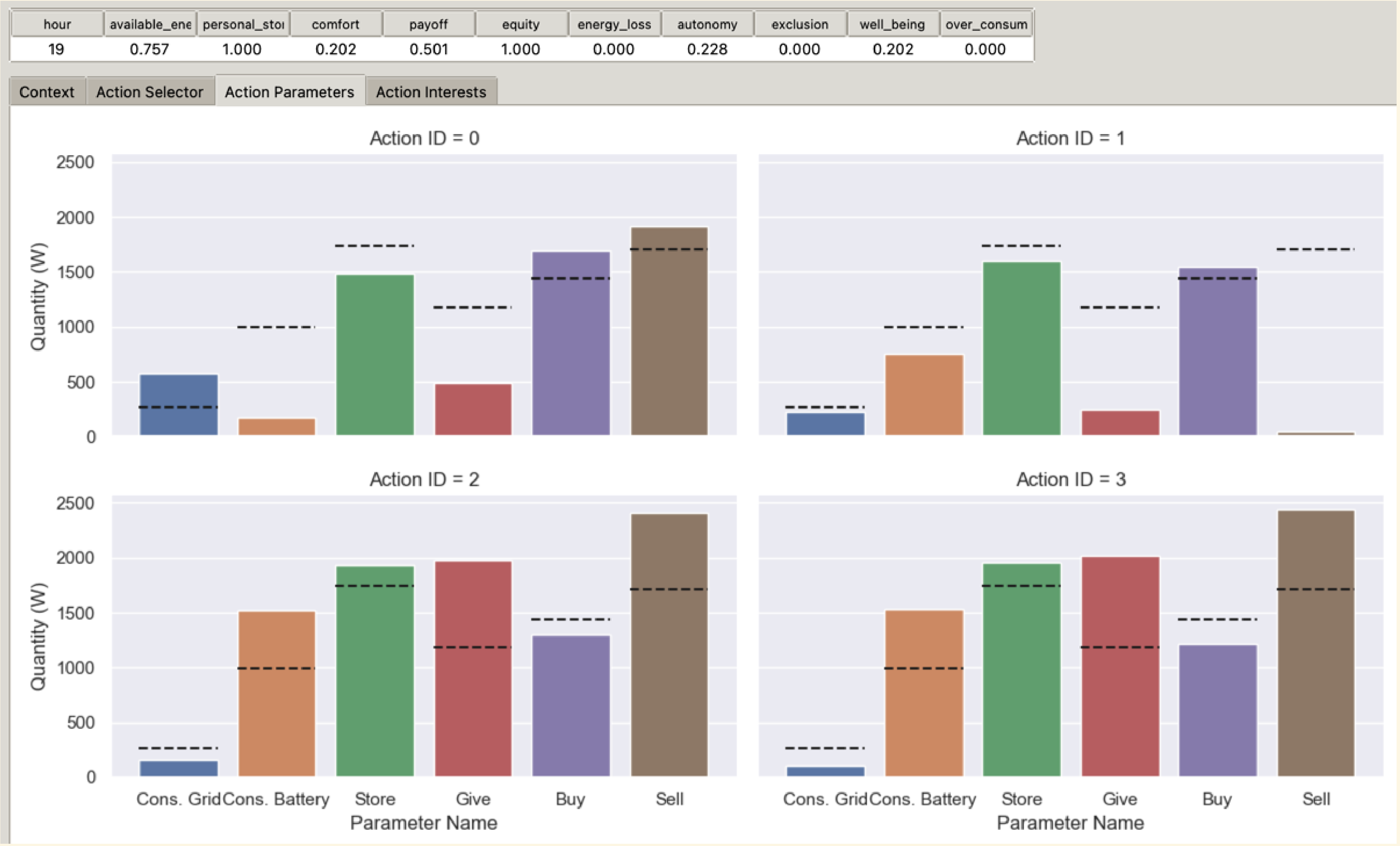
PROTOTYPE GUI



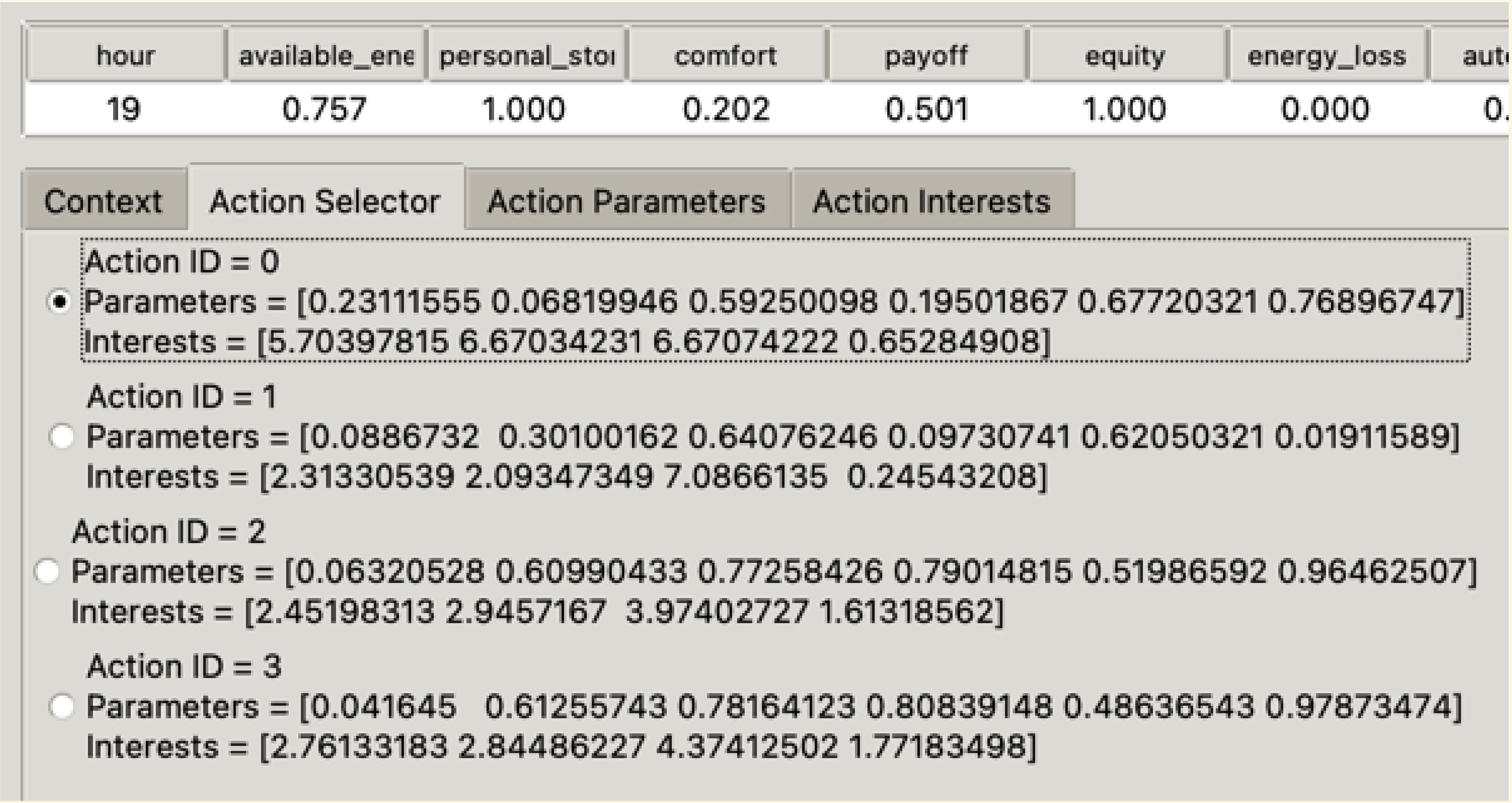
PROTOTYPE GUI



PROTOTYPE GUI



PROTOTYPE GUI



CONCLUSION

OUR PROPOSITION

- Combining RL and normative systems (e.g., argumentation)
- Learning a behaviour with a judgment-based reward signal
- Putting user in the loop with dilemmas and preferences

THANK YOU FOR YOUR ATTENTION

SMARTGRID USE-CASE