Learning to identify and settle dilemmas through contextual user preferences

Rémy Chaput

remy.chaput@univ-lyon1.fr

UCBL, LIRIS UMR5205

Laetitia Matignon

UCBL, LIRIS UMR5205

Mathieu Guillermin

CONFLUENCE (EA 1598), Lyon Catholic University

6-8 November 2023 – ICTAI 2023

https://rchaput.github.io/publication/ictai2023/

Context

- More and more AI systems “leaving the lab” to be deployed into our society[1]
- Significant impact over human lives
- Need to align with humans’ (moral) values
- Humans have various and contextual preferences over values

1. Luccioni, Alexandra, and Yoshua Bengio. 2019. “On the Morality of Artificial Intelligence.” https://arxiv.org/abs/1912.11945

Objectives

Objectives

- Learn ethically-aligned behaviours

Objectives

- Learn ethically-aligned behaviours
- Integrate contextual human preferences over multiple moral values

Objectives

- Learn ethically-aligned behaviours
- Integrate contextual human preferences over multiple moral values
- Manageable preferences for (non-expert) humans

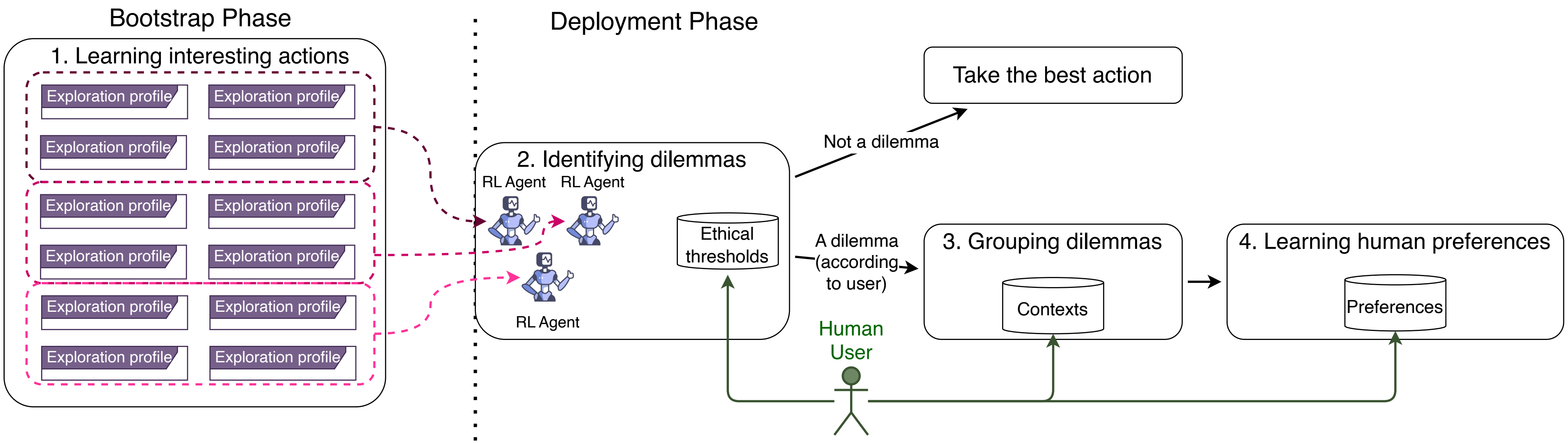
Objectives

- Learn ethically-aligned behaviours
- Integrate contextual human preferences over multiple moral values
- Manageable preferences for (non-expert) humans
- Explicitly identify dilemmas and ask users when we do not know how to solve them

Objectives

- Learn ethically-aligned behaviours
- Integrate contextual human preferences over multiple moral values
- Manageable preferences for (non-expert) humans
- Explicitly identify dilemmas and ask users when we do not know how to solve them
- Learn users’ preferences so we can automate the dilemmas that are already known

Architecture



- Block-based architecture ; Multi-Objective Reinforcement Learning
- We leverage the QSOM[1] learning algorithm

1. Chaput, Rémy, Olivier Boissier, and Mathieu Guillermin. 2023. "Adaptive Reinforcement Learning of Multi-Agent Ethically-Aligned Behaviours: The QSOM and QDSOM Algorithms." https://arxiv.org/abs/2307.00552

The bootstrap phase

Learning interesting actions

- Goal: find actions (parameters and Q-Values) that can be proposed during dilemmas
- Should offer different trade-offs ⇒ we cannot focus only on, e.g., averaging multiple objectives

⇒ We introduce exploration profiles

The bootstrap phase

Exploration profiles

Exploration profile p

- State-SOM: Self-Organizing Map[1] that maps continuous observations to discrete states
- Action-SOM: SOM that maps action identifiers to continuous action parameters
- Q-Table Q_p: multi-objective interests of actions in states
- Vector of exploration weights ϱ

Learns a subset of the action space, directed by ϱ

1. Kohonen, Teuvo. 1990. "The Self-Organizing Map." Proceedings of the IEEE 78 (9): 1464–80. https://doi.org/10.1109/5.58325

The bootstrap phase

Determining if an action is interesting

The action space is learned by:

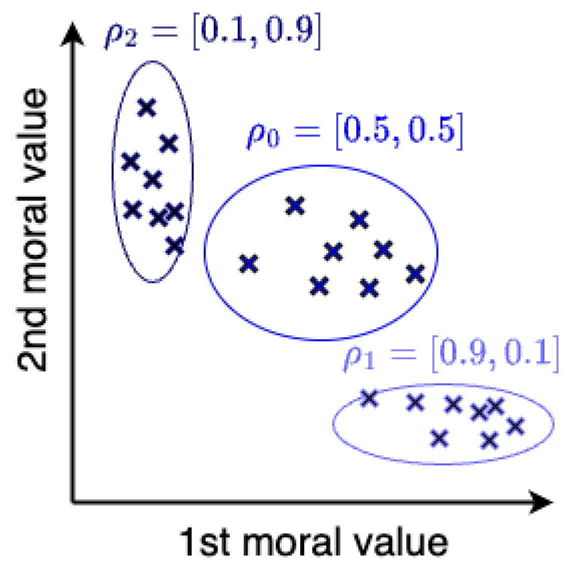
- Selecting an action
- Randomly noising it to explore
- Determining whether the noised action is better than the learned one:

$$\vec{\varrho} \cdot \underbrace{\vec{r_t}}_{\text{Reward}} + \gamma \underbrace{\underset{j'}{\operatorname{argmax}}\left(\vec{\varrho} \cdot Q_p(s_{t+1}, j') \right)}_{\text{Prediction of next action}} \overset{?}{>} \vec{\varrho} \cdot \underbrace{Q_p(s_t, j)}_{\text{Learned action}}$$

The bootstrap phase

Exploration weights

- $m + 1$ exploration profiles (m = number of moral values)
- generalist profile
  - → $\varrho_0 = [\frac{1}{m}, \cdots, \frac{1}{m}]$
- m specialized profiles
  - → $\varrho_1 = [0.9, \frac{0.1}{m-1}, \cdots, \frac{0.1}{m-1}]$
  - → $\varrho_2 = [\frac{0.1}{m-1}, 0.9, \cdots, \frac{0.1}{m-1}]$
  - → $\cdots$
  - → $\varrho_m = [\frac{0.1}{m-1}, \cdots, \frac{0.1}{m-1}, 0.9]$



The deployment phase

Learning users' preferences

- Goal: learn to execute actions corresponding to users' preferences in dilemmas
- Need to identify dilemmas
- Reduce cognitive load: do not ask each time, but re-apply same actions in similar situations

The deployment phase

Theoretical interests

Theoretical interests

Q^{theo} of same shape as Q (3D Q-Table)

Learned by assuming the action obtained the maximal reward

Represent interests an action *would have* if it had perfect impact

Using the ratio $\frac{Q(s,a)}{Q^{theo}(s,a)}$ gives an idea of how well the action performs

The deployment phase

Ethical thresholds

Ethical thresholds

Set by users

Represent expectations over permissible actions

Constraints relative to *interests* and *theoretical interests*

ζ = set (of any size) of vectors (of size m), *or* relationships over *and*

For example, $(0.6 \wedge 0.6) \vee (0.8 \wedge 0.5)$

The deployment phase: Acceptable actions and Dilemmas

Acceptable action

Action that is deemed permissible by user, based on ethical thresholds ζ

$$\mathtt{acceptable}(\vec{o}, p, a, \zeta) \Leftrightarrow \exists i\ \forall k \in [[1, m]]\ \frac{Q_p(\mathtt{States}_p(\vec{o}), a, k)}{Q_p^{theo}(\mathtt{States}_p(\vec{o}), a, k)} \geq \zeta_{i,k}$$

For example, $[\frac{9}{10}, \frac{5}{10}]$ is acceptable w.r.t.
$(0.6 \wedge 0.6) \vee (0.8 \wedge 0.5)$

The deployment phase: Acceptable actions and Dilemmas

Acceptable action

Action that is deemed permissible by user, based on ethical thresholds ζ

$$\text{acceptable}(\vec{o}, p, a, \zeta) \Leftrightarrow \exists i\ \forall k \in [[1, m]]\ \frac{Q_p(\text{States}_p(\vec{o}), a, k)}{Q_p^{theo}(\text{States}_p(\vec{o}), a, k)} \geq \zeta_{i,k}$$

For example, $[\frac{9}{10}, \frac{5}{10}]$ is acceptable w.r.t. $(0.6 \wedge 0.6) \vee (0.8 \wedge 0.5)$

Dilemmas

Situations in which no action is permissible

$$\text{dilemma}(\vec{o}, \zeta) \Leftrightarrow \nexists(p, a) : \text{acceptable}(\vec{o}, p, a, \zeta)$$

The deployment phase

Contexts

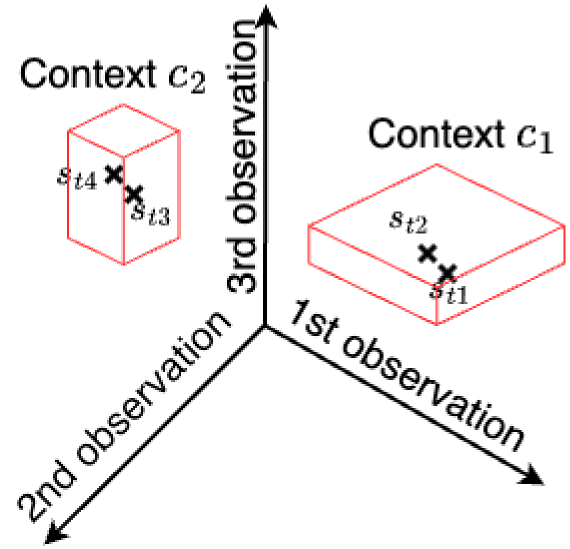
Context

Allows to group similar dilemmas together

Defined by users based on situations

Bounds over the observations

$c = \langle (b_1, B_1), \cdots, (b_g, B_g) \rangle$ for g dimensions



- System memorizes chosen action when a context is created
- The same action is automatically re-applied when the same context is identified

Experiments and results

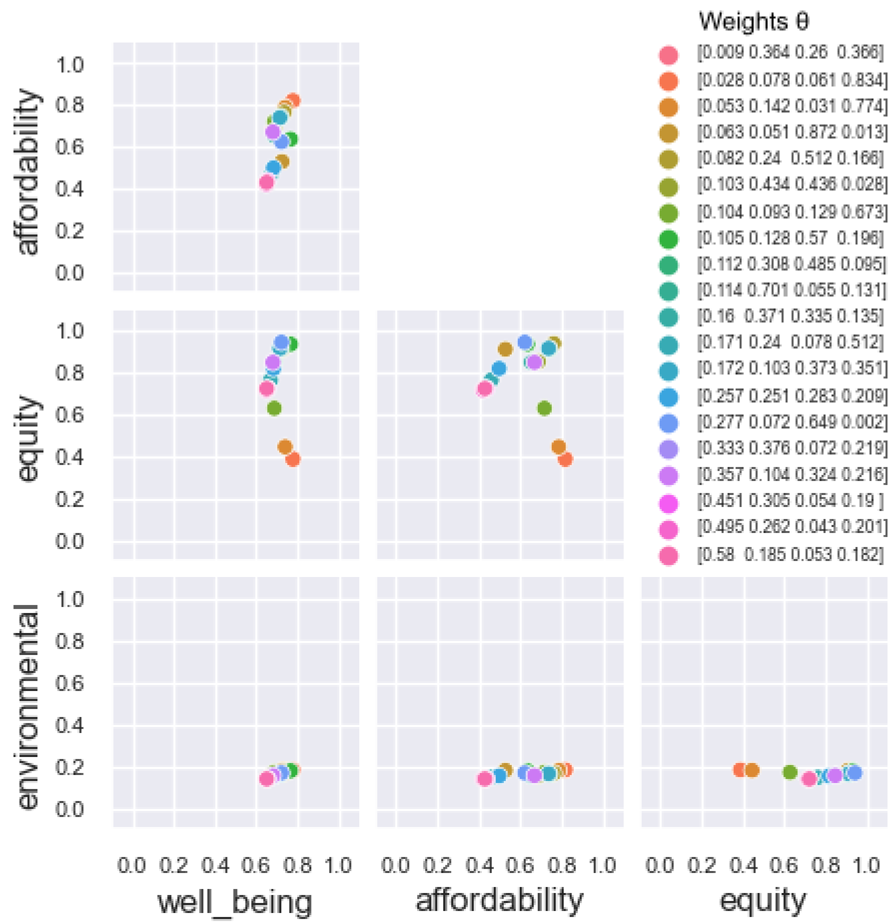
- Case study: energy distribution within a small simulated Smart Grid
  - → 4 moral values, handcrafted [1]

Two experiments:

- Checking that agents learn various actions
- Checking that dilemmas are manageable (cognitive load)

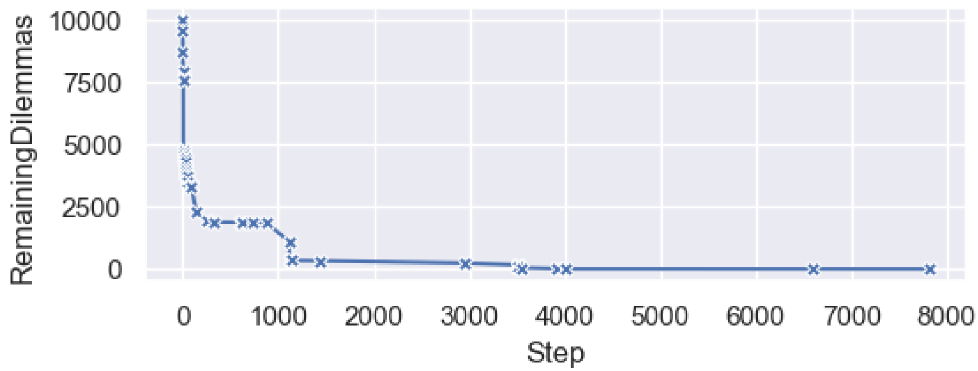
1. Alcaraz, Benoît, Olivier Boissier, Rémy Chaput, and Christopher Leturc. 2023. "AJAR: An Argumentation-Based Judging Agents Framework for Ethical Reinforcement Learning." In AAMAS '23. https://dl.acm.org/doi/abs/10.5555/3545946.3598956

Experiments: Agents learn various actions



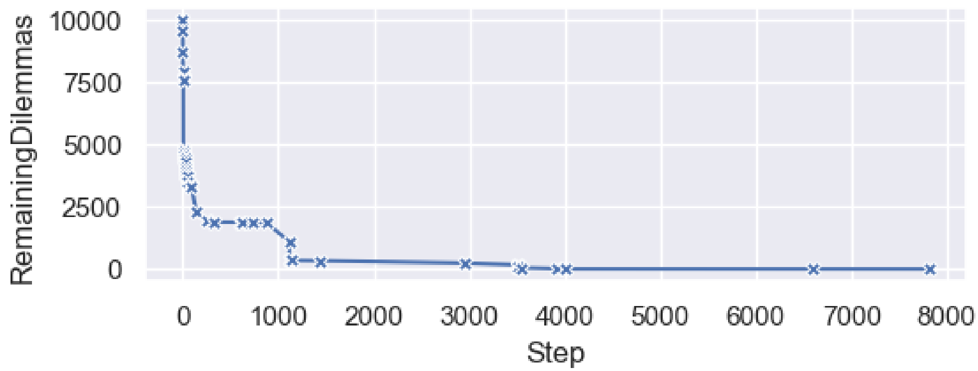
- Automatic policies
- $\pi_\theta(s) = \mathrm{argmax}_a\ \theta \cdot$
- Policies' scores plotted moral values 2-by-2
- Various trade-offs iden
- But exploration could b better (especially for *Environmental*)

Experiments: Manageable dilemmas

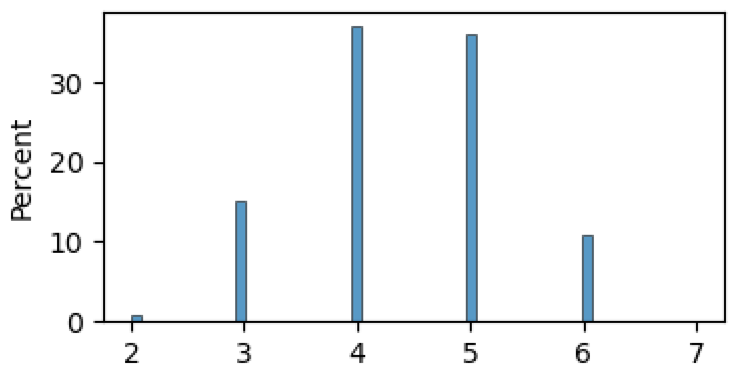


- Number of dilemmas diminishes very quickly

Experiments: Manageable dilemmas



- Number of dilemmas diminishes very quickly



- In average, 4.4 actions proposed per dilemma

Conclusion

- A novel approach for Multi-Objective RL
- Learning ethically-aligned behaviours
- Focuses on explicitly identifying dilemmas
- Algorithm learns various trade-offs, but exploration could have been better
- The block-based architecture allows improvements (e.g., curiosity-based exploration)

Thank you for your attention

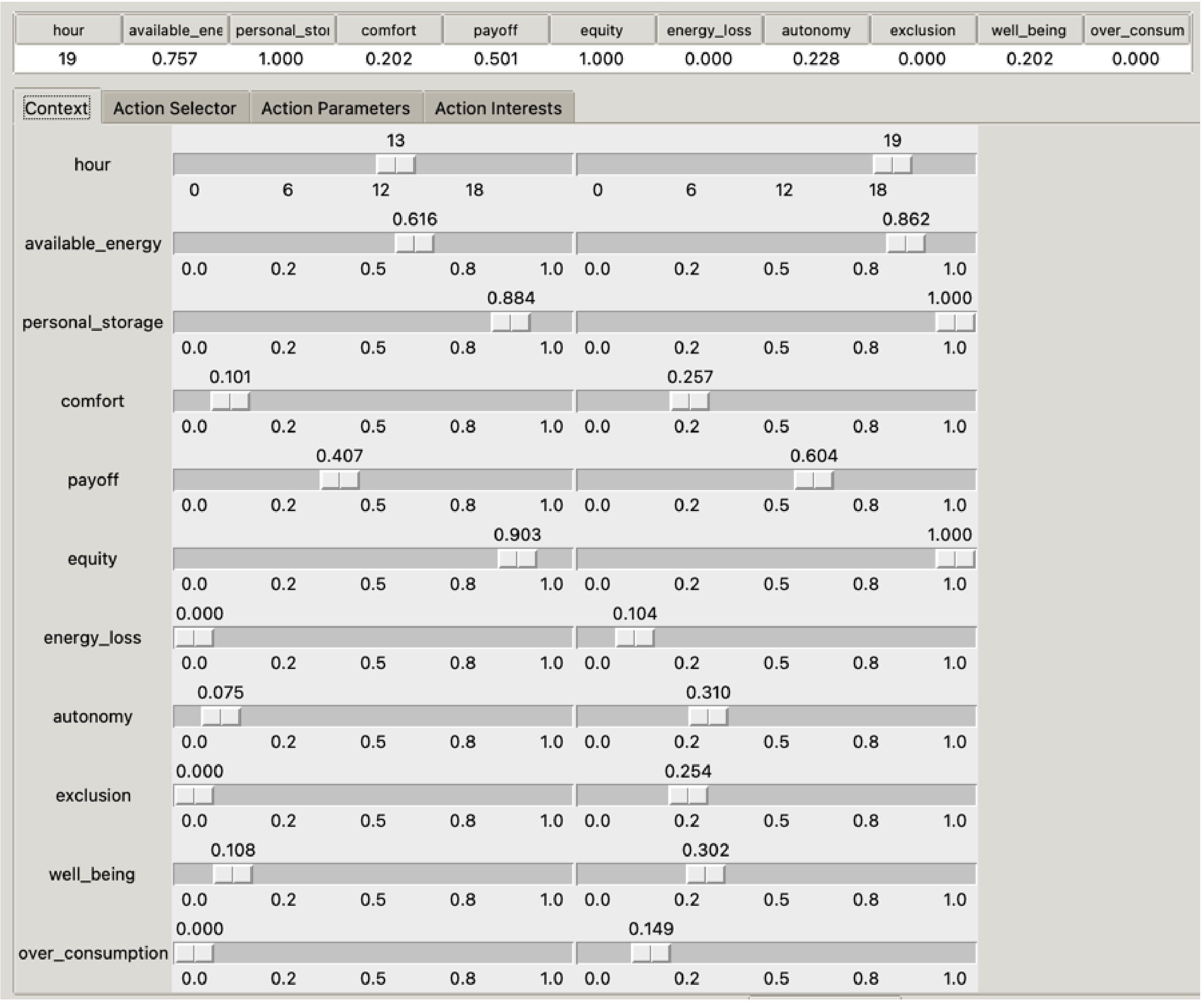
Any questions?

The bootstrap phase: updated Bellman equation

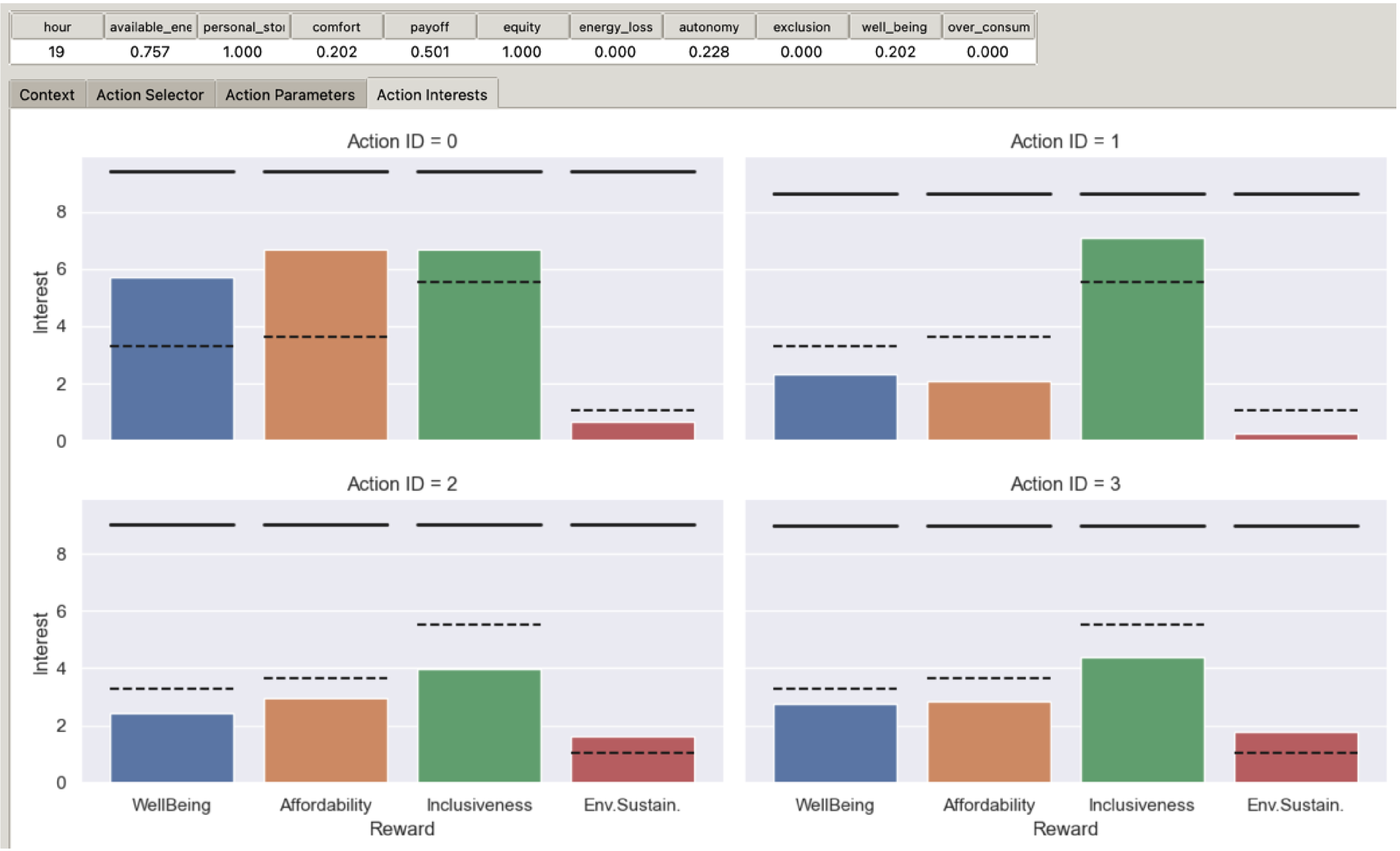
We add a 3rd dimension (the moral value)

$$\forall k \in [\![1, m]\!] : Q_p^{t+1}(s_t, a_t, k) \leftarrow \alpha \left[r_{t,k} + \gamma \max_{a'} \varrho \cdot Q_p^t(s_{t+1}, a')_k \right] + (1-\alpha) Q_p^t(s_t, a_t, k)$$

Graphical User Interface: choosing a context



Graphical User Interface: comparing actions' interests



Graphical User Interface: comparing actions' parameters