AJAR: An Argumentation-based Judging Agents Framework for Ethical Reinforcement Learning

Extended Abstract

Benoît Alcaraz
benoit.alcaraz@uni.lu
University of Luxembourg
Esch-sur-Alzette, Luxembourg

Olivier Boissier
Olivier.Boissier@emse.fr
Mines Saint-Etienne, Univ Clermont Auvergne, CNRS,
UMR 6158 LIMOS, Institut Henri Fayol
F - 42023 Saint-Etienne, France

Rémy Chaput
remy.chaput@univ-lyon1.fr
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205
F-69622 Villeurbanne, France

Christopher Leturc
christopher.leturc@inria.fr
Inria, Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

ABSTRACT

An increasing number of socio-technical systems embedding Artificial Intelligence (AI) technologies are deployed, and questions arise about the possible impact of such systems onto humans. We propose a hybrid multi-agent Reinforcement Learning framework consists of *learning agents* that learn a task-oriented behaviour defined by a set of *symbolic moral judging agents* to ensure they respect moral values. This framework is applied on the problem of responsible energy distribution for smart grids.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**.

KEYWORDS

Argumentation; Reinforcement Learning; Machine Ethics; Hybrid Neural-Symbolic Learning; Ethical Judgment; Artificial Moral Agent

ACM Reference Format:
Benoît Alcaraz, Olivier Boissier, Rémy Chaput, and Christopher Leturc. 2023. AJAR: An Argumentation-based Judging Agents Framework for Ethical Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

With the deployment of more and more Artificial Intelligence (AI) systems in real-life contexts, questions arise about the ability of such systems to not only achieve their assigned goal, but to have a beneficial impact onto humans (daily life, society as a whole, etc.). These impacts involve several ethical considerations, making such agents what Moor calls "ethical impact agents" [13]. Scholars propose the integration of various concerns into these systems, such as: transparency, justice, fairness and privacy [10]. However, most of them are no mere technical requirements, and it is unclear how they should be implemented and conflicts resolved. Many works in the Machine Ethics field [3] proposed to implement "ethical principles", inspired from moral philosophy, by introducing, e.g., moral values into the system, represented as multiple criteria that agents must respect. Proposed implementations are mostly Top-Down formalization of existing principles, or Bottom-Up learning of new principles [1], both having advantages and drawbacks. Yet, there is no consensus on which ethical principle(s) should be used, nor how should they be implemented [14]. We represent those ethical considerations as a set of moral values, which are identified by designers during the conception and made available to agents by a reward signal. Argumentation can be successfully leveraged to represent complex judging and decisions structures, in opposition to numerical functions or simple logical rules. We propose to use argumentation for judging about the respect of moral values and separate a reinforcement learning task (agents learn to act ethically) from an argumentation-based judging task (agents are judged on their acts). Learning agents need to adapt their behaviour according to a context and a set of moral values to respect, represented as a set of multiple criteria to maximize. Thus, our contribution offers the following advantages: (1) Combines advantages of both Top-Down and Bottom-Up. (2) Argumentation gives a richer structure by allowing explicit conflicts. (3) It is easier for the designers to structure the rules. (4) Argumentation graphs are more readable to external users or regulators (non-developers) [11]. (5) Arguments themselves, and their activation context, can be leveraged by explanation methods to understand the learnt behaviour. (6) Reward hacking can be detected and corrected. We apply it to the use-case of energy distribution within a Smart Grid, adapted from [7] and pictured in Figure 1. The agents' goal is to consume energy to improve the inhabitants' comfort, whereas the system's goal is to make agents learn to respect different moral values adapted from the literature [5, 8, 12], namely: *Security of Supply*, *Affordability*, *Inclusiveness*, and *Environmental Sustainability*.

2 THE AJAR FRAMEWORK

We model our system as a Decentralized Partially Observable Markov Decision Process (DecPOMDP) [4][1], i.e., we have several learning agents that receive partial *observations* of the real world states.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[1] An implementation of AJAR can be found at https://doi.org/10.5281/zenodo.7628903

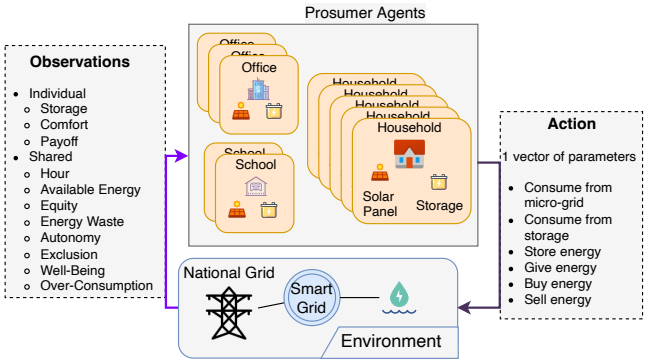
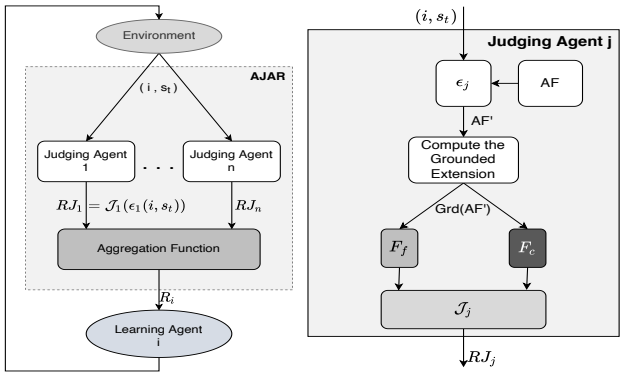


Figure 1: The smart grid use-case, adapted from [7].



(a) Overview of AJAR. **(b) Detailed judgment.**

Figure 2: Representation of the AJAR framework.

They take *actions* that are vectors of (continuous) parameters, and receive a scalar *reward* to learn the best action in each situation. Note that the reward function is the same for each learning agent, yet they may receive different rewards. Each reward represents the individual judgment of one RL-agent, based on the aggregation of the compliance with several criteria, such as moral values, given by *judging agents* from the processing of their argumentation graphs.

Each judging agent judges the behaviour of RL-agents, w.r.t. their own specific moral value. Indeed, separating them allows more flexibility, i.e., it facilitates the addition or removal of moral values into the system. Judging agents require multiple types of arguments to express whether the RL-agent's last decision is moral, immoral, or neutral, w.r.t. its own moral value. This corresponds to a notion of pros- and cons-arguments. We adapt the model of Amgoud and Prade [2] and call *An Argumentation Framework for Judging a Decision* (AFJD) a tuple $AF = \left(AF_{[\mathcal{A}rgs]}, AF_{[\mathcal{A}tt]}, AF_{[\mathcal{F}_f]}, AF_{[\mathcal{F}_c]}\right)$ where $\mathcal{A}rgs$ is a non-empty set of arguments, $\mathcal{A}tt$ is a binary relation called *attack relation*, $\mathcal{F}_f \in 2^{\mathcal{A}rgs}$ is the set of arguments which indicates that the RL-agent's last decision is moral, w.r.t. the moral value considered by the judging agent, $\mathcal{F}_c \in 2^{\mathcal{A}rgs}$ is the set of arguments which indicates that the RL-agent's last decision is immoral. The set of all possible sub-AFJD for AF, i.e., all AFJD which arguments are a subset of $AF_{[\mathcal{A}rgs]}$, is denoted as: $\mathcal{P}(AF) := \{(\mathcal{A}rgs', \mathcal{A}tt', \mathcal{F}_f', \mathcal{F}_c') : \mathcal{A}rgs' \subseteq AF_{[\mathcal{A}rgs]}, \mathcal{A}tt' \subseteq \mathcal{A}rgs'^2 \cap AF_{[\mathcal{A}tt]}, \mathcal{F}_f' \subseteq \mathcal{A}rgs' \cap AF_{[\mathcal{F}_f]}, \mathcal{F}_c' \subseteq \mathcal{A}rgs' \cap AF_{[\mathcal{F}_c]}\}$. A judging agent is defined as an agent which reasons about decisions using an AFJD. This argumentation model, embedded into a judging agent, reflects a pre-reasoning by the software designers who may have applied some ethical principles to design these arguments. Designers may introduce deontological or consequentialist arguments. It is possible for another argument to attack the previous argument, making it unacceptable. To describe this "acceptability", scholars consider the notion of *admissibility*. It characterizes which arguments are relevant to the judging agent, to compute the compliance with a moral value. Formally, it corresponds to all sets of arguments which are conflict-free and acceptable [9]. Admissible sets of arguments are also called extensions. We use the grounded extension, since a very efficient algorithm has been proposed in [15] and its uniqueness [6]. Finally, the judging agent gives its judgment as a reward value between $[0, 1]$.

We define *A Judging Agents-based RL-framework* (AJAR) as a RL-framework where we consider learning agents $\mathcal{N}_l$, and judging agents $\mathcal{M}_{judges}$ with their embedded AFJD, noted as AF. Each judging agent j builds its own argumentation graph AF_j according to the observations it gets about the judged RL-agent i, using a function ϵ_j. Thus, the argumentation graph may differ from one state to another state. They perform judgments through their function $\mathcal{J}_j$, which takes an argumentation graph and returns a real number that corresponds to the compliance of the RL-agent behaviour with the moral value. The reward given to a RL-agent comes from the aggregation function g_{agr} applied on all judgments.

Definition 2.1. A Judging Agents-based RL-framework (AJAR) is a tuple $\mathcal{F} = \langle \mathcal{M}_{judges}, \{AF_j\}_{j \in \mathcal{M}_{judges}}, \mathcal{N}_l, \mathcal{S}, \{Act_i\}_{i \in \mathcal{N}_l}, \mathcal{T}, \{\mathcal{R}_i\}_{i \in \mathcal{N}_l}, \{\Omega_i\}_{i \in \mathcal{N}_l}, \{O_i\}_{i \in \mathcal{N}_l}, \gamma, \{\epsilon_j\}_{j \in \mathcal{M}_{judges}}, \{\mathcal{J}_j\}_{j \in \mathcal{M}_{judges}}, g_{agr} \rangle$ where $\langle \mathcal{N}_l, \mathcal{S}, \{Act_i\}_{i \in \mathcal{N}_l}, \mathcal{T}, \{\mathcal{R}_i\}_{i \in \mathcal{N}_l}, \{\Omega_i\}_{i \in \mathcal{N}_l}, \{O_i\}_{i \in \mathcal{N}_l}, \gamma \rangle$ is a DecPOMDP, $\forall j \in \mathcal{M}_{judges}, \epsilon_j : \mathcal{N}_l \times \mathcal{S} \rightarrow \mathcal{P}(AF_j)$ is a function that from a RL-agent $i \in \mathcal{N}_l$, and a current state $s_t \in \mathcal{S}$, assigns the sub-AFJD that the judging agent j uses to judge the agent i, $\forall j \in \mathcal{M}_{judges}, \mathcal{J}_j : \mathcal{P}(AF_j) \rightarrow \mathbb{R}$ is the *judgment function*, $g_{agr} : \mathbb{R}^{|\mathcal{M}_{judges}|} \rightarrow \mathbb{R}$ is an aggregation function, for all RL-agents $i \in \mathcal{N}_l$ and, for all states $s_t \in \mathcal{S}$, $\mathcal{R}_i$ is s.t.:

$$\mathcal{R}_i(s_t) = g_{agr}\Big(\bigtimes_{j \in \mathcal{M}_{judges}} \mathcal{J}_j\big(\epsilon_j(i, s_t)\big)\Big)$$

Figure 2a shows how a RL-agent i is judged, and Figure 2b describes how a judging agent builds a reward. An example of a judgment function $\mathcal{J}_j\left(\epsilon_j(i, s_t)\right)$ is given below. For a learning agent i (being judged), and for a current state s_t, with $\mathcal{G}rd(\epsilon_j(i, s_t))$ the computed grounded extension from $\epsilon_j(i, s_t)$, we define: $pros = \left|\mathcal{G}rd(\epsilon_j(i, s_t)) \cap \epsilon_j(i, s_t)_{[\mathcal{F}_f]}\right|$, $cons = \left|\mathcal{G}rd(\epsilon_j(i, s_t)) \cap \epsilon_j(i, s_t)_{[\mathcal{F}_c]}\right|$, and $\mathcal{J}_j\left(\epsilon_j(i, s_t)\right) = \frac{pros}{pros+cons}$, if $pros + cons \neq 0$, otherwise $\frac{1}{2}$.

3 CONCLUSION

We proposed a framework which allows defining reward functions based on formal argumentation to judge RL-agents on their ethical behaviours and justify the rewards given to RL-agents.

ACKNOWLEDGMENTS

This work is partially funded by the French Région Auvergne Rhônes-Alpes (AURA), as part of the Ethics.AI project (Pack Ambition Recherche), and by the Luxembourg National Research Fund under Grant No. IS/14717072. We gratefully acknowledge support from the CNRS/IN2P3 Computing Center (Lyon — France) for providing computing and data-processing resources.

REFERENCES

[1] Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology* 7, 3 (2005), 149–155.

[2] Leila Amgoud and Henri Prade. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* 173, 3-4 (2009), 413–436.

[3] Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics*. Cambridge University Press.

[4] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[5] Anne R. Boijmans. 2019. *The Acceptability of Decentralized Energy Systems: Identifying Value Conflicts Through Simulations Of Decentralized Energy Systems For City Districts*. Master's thesis. Delft University of Technology.

[6] Martin Caminada. 2007. Comparing two unique extension semantics for formal argumentation: ideal and eager. In *19th Belgian-Dutch Conference on Artificial Intelligence*. 81–87.

[7] Rémy Chaput, Jérémy Duval, Olivier Boissier, Mathieu Guillermin, and Salima Hassas. 2021. A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society (AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society)*. ACM, Virtual Event USA, United States, 13–23. https://doi.org/10.1145/3461702.3462515

[8] Tristan E. De Wildt, Émile J.L. Chappin, Geerten van de Kaa, Paulien M. Herder, and Ibo van de Poel. 2019. Conflicting values in the smart electricity grid a comprehensive overview. *Renewable and Sustainable Energy Reviews* 111 (2019), 184–196.

[9] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2 (1995), 321–357.

[10] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[11] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences* 34, 2 (2011), 57–74.

[12] Christine Milchram, Geerten Van de Kaa, Neelke Doorn, and Rolf Künneke. 2018. Moral values as factors for social acceptance of smart grid technologies. *Sustainability* 10, 8 (2018), 2703.

[13] James H. Moor. 2009. Four kinds of ethical robots. *Philosophy Now* 72 (2009), 12–14.

[14] Vivek Nallur. 2020. Landscape of machine implemented ethics. *Science and engineering ethics* 26, 5 (2020), 2381–2399.

[15] Samer Nofal, Katie Atkinson, and Paul E Dunne. 2021. Computing grounded extensions of abstract argumentation frameworks. *Comput. J.* 64, 1 (2021), 54–63.