AJAR: An Argumentation-based Judging Agents Framework for Ethical Reinforcement Learning
Benoît Alcaraz, Olivier Boissier, Rémy Chaput, Christopher Leturc

To cite this version:

HAL Id: hal-04127943
https://hal.science/hal-04127943
Submitted on 14 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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

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

Olivier Boissier
Olivier.Boissier@emse.fr
Mines Saint-Etienne, Univ Clermont Auvergne, CNRS,
UMR 6158 LIMOS, Institut Henri Fayol
F - 42023 Saint-Etienne, 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 on humans. We propose a hybrid multi-agent Reinforcement Learning framework consists of learning agents that learn a task-oriented behavior 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

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 on 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 they should 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], i.e., we have several learning agents that receive partial observations of the real world states.

Free license: This material is free of known copyright restrictions.
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 $AF \equiv \langle AF_0, AF_{\text{Grd}}, AF_{\text{Pros}}, AF_{\text{Con}} \rangle$ where $AF_0$ is a non-empty set of arguments, $AF_{\text{Grd}}$ is a binary relation called attack relation, $AF_{\text{Pros}}$ is a 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, $AF_{\text{Con}}$ 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 $AF_0$, is denoted as $\mathcal{P}(AF) := \{ (\text{Args}', \text{Att}', \text{T}'_\text{c}) : \text{Args}' \subseteq AF_0, \text{Att}' \subseteq \text{Args}'^2 \cap \text{AF}_{\text{Att}}, \text{T}'_\text{c} \subseteq \text{Args}'^2 \cap \text{AF}_{\text{T}'_\text{c}}, \text{T}'_\text{c} \subseteq \text{Args}'^2 \cap \text{AF}_{\text{T}'_\text{c}} \}$. 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 $N_i$, and judging agents $M_{judges}$ 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 $e_j$. Thus, the argumentation graph may differ from one state to another state. They perform judgments through their function $J_j$, 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 $F = \langle M_{\text{judges}}, \{AF_j\}_{j \in M_{\text{judges}}}, N_i, S, \{\text{Act}_i\}_{i \in N_i}, T, \{\text{R}_i\}_{i \in N_i}, \{\text{O}_i\}_{i \in N_i}, \text{g}_{agr}, \text{g}_{agg} \rangle$ where $(N_i, S, \{\text{Act}_i\}_{i \in N_i}, T, \{\text{R}_i\}_{i \in N_i}, \{\text{O}_i\}_{i \in N_i}, \text{g}_{agr}, \text{g}_{agg})$ is a DecPOMDP, $\forall j \in M_{\text{judges}}, e_j : N_i \times S \rightarrow \mathcal{P}(AF_j)$ is a function that from a RL-agent $i \in N_i$ and a current state $s_i \in S$, assigns the sub-AFJD that the judging agent $j$ uses to judge the agent $i, \forall j \in M_{\text{judges}}, J_j : \mathcal{P}(AF_j) \rightarrow \mathbb{R}$ is the judgment function, $g_{agr} : \mathbb{R}^{M_{\text{judges}}} \rightarrow \mathbb{R}$ is an aggregation function, for all RL-agents $i \in N_i$ and, for all states $s_i \in S, R_i$ is s.t.:

$$R_i(s_i) = g_{agr}\left( \bigotimes_{j \in M_{\text{judges}}} J_j(e_j(i, s_i)) \right)$$

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 $J_j(e_j(i, s_i))$ is given below. For a learning agent $i$ (being judged), and for a current state $s_i$, with $Grd(e_j(i, s_i))$ the computed grounded extension from $e_j(i, s_i)$, we define: $\text{pros} = |Grd(e_j(i, s_i)) \cap e_j(i, s_i)|, \text{cons} = |Grd(e_j(i, s_i)) \cap e_j(i, s_i)^c|, \text{and } J_j(e_j(i, s_i)) = \frac{\text{pros}}{\text{pros} + \text{cons}}$ if $\text{pros} + \text{cons} \neq 0$, 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