Enhancing AI Moral Value Representation through Argumentation Frameworks with Gradual Semantics

Internship proposal

Master 2 / Engineer level; 5 months

Keywords

Artificial Intelligence, Ethics, Reinforcement Learning, Argumentation

Scientific Context

With the increasing use of Artificial Intelligence (AI), it is more important than ever to ensure that these systems behave in a way that is aligned with our moral values, to guarantee beneficial use.
This requires several capabilities from the AI system: being able to represent moral values; being able to exhibit a behavior that takes these values into account; and being able to explain the behavior to humans (regulators, users, or even stakeholders) to verify that the moral values are indeed respected.

Traditionally, two different lines of approaches have opposed each other:
- Top-Down (or symbolic) approaches, which define the expected behavior through logic and reasoning;
- Bottom-Up (or learning) approaches, which learn the behavior by trial and error with respect to a metric to optimize.

Both have their advantages; in particular, Top-Down can leverage the existing knowledge from moral philosophy and human expertise, but typically fail to adapt to unknown cases; Bottom-Up can easily adapt, but the metric to get the expected behavior is hard to define, especially when preventing the agent from “hacking” the metric and optimizing for an undesired behavior.

As part of the ECIÉA project (*Explication du Comportement d’un agent IA Éthiquement Aligné* — Explaining the behavior of an ethically-aligned AI agent), we propose to implement these capabilities through a hybrid combination of abstract argumentation and Reinforcement Learning (RL). RL will provide adaptability to agents, but requires a reward function to guide the learning towards the desired behavior. Argumentation is an interesting framework for representing these moral values, as it is fairly close to human reasoning.

Internship Objective

The goal of this internship is to improve the representation of moral values through argumentation graphs, in order to make RL agents able to learn behaviors that respect these values.

A previous work demonstrated the feasibility of using argumentation graphs to define the reward function that guides the learning of RL agents [1]; however, the graphs were very simple and relied on ad-hoc judgment mechanisms to produce rewards from the graphs.
By moving towards other argumentation frameworks, this internship will improve the expressiveness of moral values, facilitate their conception by non-AI experts, and allow to produce rewards directly from the arguments, thus removing ad-hoc components.

What is expected of you

- Review the state of the art of existing argumentation frameworks and semantics, especially gradual (or similar) semantics [2,3].
- Represent moral values in the form of argumentation graphs, using a knowledge engineering method [4].
- Assess the learning of these moral values by an RL agent through experiments on a (simulated) example case of ecological gardening.
- Ideally, write a scientific article to report the interest of the proposed solution.

Two deliverables are expected:
- A research report detailing the methodology, theoretical contributions, and experimental results.
- A Python implementation illustrating the proposed approach.

References

[1] Alcaraz, Benoît, Boissier, Olivier, Chaput, Rémy, Leturc, Christopher. AJAR: An Argumentation-based Judging Agents Framework for Ethical Reinforcement Learning. Proc. of International Conference on Autonomous Agents and Multiagent Systems (AAMAS), p2427-2428, 2023.
[2] Oren, Nir, and Bruno, Yun. Inferring attack relations for gradual semantics. Argument & Computation 14.3, 327-345, 2023.
[3] Amgoud, L., Doder, D., & Vesic, S. (2022). Evaluation of argument strength in attack graphs: Foundations and semantics. *Artificial Intelligence*, *302*, 103607.
[4] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37, 2021.

Who we are looking for

We are looking for a motivated Master's level (Master 2) or final-year engineering student in Computer Science with the following qualifications:
✓ French and English proficiency at B2 level
✓ Strong skills in imperative programming, Python, and the Git version control system
✓ Knowledge in Artificial Intelligence, in particular Reinforcement Learning
✦ Knowledge in abstract argumentation is a plus, but not strictly required
✦ Proficiency in LaTeX is also a plus

How to apply

Submit your application by email to remy.chaput@cpe.fr, guillaume.muller@emse.fr:
- a CV detailing your training experience, as well as personal and/or school projects;

- your grade report for the last 2 years, including temporary grades if the official report is not available;
- a cover letter (non-specific cover letters will not be considered).

Application date: as soon as possible, ideally before end of November 2025

Desired start date: between 2 February 2026 and 2 March 2026

The internship will take place either at the LIRIS laboratory, located in the Nautibus building (22 avenue Pierre de Coubertin, 69622 Villeurbanne, France), or at the LIMOS laboratory (29 rue Ponchardier, F-42023 Saint-Étienne, France).

It will be co-supervised by Dr. Rémy Chaput (LIRIS, CPE), and Dr. Guillaume Muller (LIMOS, ENMSE). Dr. Bruno Yun (LIRIS, UCBL) and Pr. Maxime Morge (LIRIS, UCBL) will also contribute to the supervision.

The student will join the project and have the opportunity to participate in scientific presentations and meetings with other AI researchers, either in the SyCoSMA team in Lyon or the Department of Computer Science and Intelligent Systems in Saint-Étienne.

Gratification: around 640€/month (approximately 3200€ for 5 months)