Internship Master M2
An Explainability Framework for BDI Normative Agents

Advisors

- Elena YAN (elena.yan@emse.fr)
- Luis Gustavo NARDIN (gnardin@emse.fr)

Location : MINES Saint-Étienne, Institute Henri Fayol, ISI department

Start date : February / March 2026

Application deadline : Open until filled

Duration : up to 6 months

Scholarship : around 650€ / month

Keywords : Explainable AI, Normative Multiagent Systems, Industry of the Future

Description

This internship takes place in the context of the ANR/FAPESP international collaboration project Normative Artificial Intelligence for regulating MANufacturing (NAIMAN).

The digital transformation of manufacturing industries provides a nurturing environment for the adoption of AI technologies that can quickly and flexibly respond to endogenous and exogenous changes, while being transparent and complying with the complex regulatory landscape and requirements, e.g., the EU AI Act, NIS2, the Cyber Resilience Act, GDPR, and various environmental standards. The adoption of AI technologies in these industries imposes significant challenges on companies.

The governance of these solutions to assure their compliance with regulations and standards becomes essential. We focus on automating the governance of systems and processes by explicitly representing regulations and standards, and using autonomous agents and normative multiagent systems (NMAS) to assure their compliance with these regulations and standards.

In NMAS, regulation management systems aim to achieve a balance between the autonomy of the agent and the control of the system. Regulation management systems are composed of regulation representations (i.e., explicit rules guiding the behavior of the entities in a system, e.g., norms, sanctions) and regulation capabilities (i.e., functions, procedures and mechanisms that an entity has to manage the regulation representations, e.g., create, regiment, enforce, adapt) that can be mapped onto several possible regulation architectures. To enable the implementation of NMAS, Yan et al. [1] proposed NPL(s), a normative programming language enriched with the representation of norms and sanctions as first-class abstractions, and a Belief-Desire-Intention (BDI) normative agent architecture embedding an engine for processing this language.

Normative agents generally have the ability to interpret and make decisions influenced by the regulation representations. Although these regulation representations improve upon the transparency of the system's behavior, they do not explain how agents use them to make decisions, which is necessary to establish human trust in the system.

In the context of explainability, Yan et al. [2] proposed a multi-level explainability framework for engineering and understanding BDI agents. The framework allows for generating narratives explaining the agents behavior at different levels of abstractions, targeted to different types of users: *implementation level* (developers); *design level* (designers); and *domain level* (end-users). The framework does not assume anything about specific types of agents beliefs, e.g, regulation representations, in order to generate explanations about the behavior of the agents.

Objective

This internship aims to extend this multi-level explainability framework to incorporate a specific type of belief, i.e., regulation representations as defined in the normative programming language NPL(s), to generate the narratives explaining agents behavior. The proposed framework will be applied in a case study for improving the explainability of agents' decision-making in the Fábrica do Futuro@USP, i.e., an educational skateboard manufacturing assembly line implementing a product-centric production approach.

Responsibilities

- Extend the conceptual explainability framework for BDI normative agents
- Implement a prototype using JaCaMo applied to the sociotechnical system Fábrica do Futuro@USP
- Evaluate the extended explainability model including normative aspects against existing explainability models
- Write reports about the evolution and final outcomes of the developments

Application Procedure

If you are interested in this internship, apply to this position by sending a **Letter of Motivation** highlighting your main strengths and interests for this position, a **CV**, the **Master transcripts**, and any additional information you deem relevant to Elena YAN (elena.yan@emse.fr) and Luis Gustavo NARDIN (gnardin@emse.fr).

References

[1] Yan, E., Nardin, L. G., Hübner, J. F., & Boissier, O. (2025). An agent-centric perspective on norm enforcement and sanctions. In Cranefield, S., Nardin, L. G., Lloyd, N. (Eds.), *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVII* (pp. 79-99). Cham: Springer, COINE 2024. LNCS v.15398. doi: 10.1007/978-3-031-82039-7_6.

[2] Yan, E., Burattini, S., Hübner, J. F., & Ricci, A. (2025). A multi-level explainability framework for engineering and understanding BDI agents. *Autonomous Agents and Multi-Agent Systems*. doi: 10.1007/S10458-025-09689-6.