Preferences and Value Alignment in AI

PhD advisor: Paolo Viappiani

Description

The development of advanced AI systems introduces significant risks, especially when autonomous agents adopt unintended behaviors by optimizing metrics that do not perfectly represent what we really care about. This is known as the "value alignment problem": ensuring AI systems act in line with human values and ethics. Manually specifying ethical rules is often impractical, as values are complex, context-dependent, and vary across cultures and time; enumerating all ethical behaviors in advance is in practice impossible.
In this work we will explore models and methods for representing and learning ethical behaviours, focusing on adaptive methods, enabling AI agents to learn and align with users' values through interaction.
As highlighted by S. Russell in his influential book « Human Compatible » dealing with preferences is a key issue in AI, as autonomous systems should learn and adapt to what humans really want. The departing point of the thesis will be focused in modeling both "preferences" and "values", and in seeing how methods for preference elicitation can be adapted or extended to address the problem of value alignement.

Some references

Andrés Holgado-Sánchez, Sascha Ossowski, Holger Billhardt and Sara Degli-Esposti. Learning the value systems of societies from preferences, ECAI 2025

Jana Schaich Borg, Walter Sinnott-Armstrong, Vincent Conitzer. Moral AI: And How We Get There. Pelican 2024

Adrien Ecoffet, Joel Lehman. Reinforcement Learning Under Moral Uncertainty. ICML 2021

Stuart Russell *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking - Penguin Random House, 2019