MASTER THESIS PROJECT

Incorporating user feedback in norm and value estimation

Project description

As autonomous agents (AAs) are becoming ubiquitous, there is a growing need to reduce the risk of mistakes these agents may introduce. Hence, we need to develop methods that take into account the norms and values of those who depend and interact with such agents. One major part of this is gaining feedback of users to see whether the agent is actually right in its estimation of the situation.

In this thesis, you will develop algorithms to learn and estimate the **norms and values** at play in human interaction (e.g. playing a game or dividing up a pie). **User feedback** is essential in value estimation as it provides data that cannot be readily inferred from observables. How can we incorporate this feedback to reduce ambiguity? How and how often should one ask for feedback about norms and values? How does one evaluate user feedback? And what is done when the estimation and feedback are misaligned?

This project will be developed at the Interactive Intelligence group.

Interactive Intelligence Group: The Interactive Intelligence (II) section focuses on socially interactive, intelligent agents. We research the intelligence that underlies and co-evolves during the repeated interactions of human and technology "agents" who cooperate to achieve a joint goal. Our research program aims for synergy and social interaction between humans and technology, to empower humans in their social context.

References and further reading

Siebert, Luciano Cavalcante, et al. "Improving Confidence in the Estimation of Values and Norms." *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*. Springer, Cham, 2017. 98-113.

Dechesne, Francien, et al. "No smoking here: values, norms and culture in multi-agent systems." *Artificial intelligence and law* 21.1 (2013): 79-107.

Gabriel, Iason. "Artificial intelligence, values, and alignment." *Minds and machines* 30.3 (2020): 411-437.

Practical details

Starting date: As soon as possible.

- Location: TU Delft, Interactive Intelligence Group
- Daily supervisor: Luciano Cavalcante Siebert (TU Delft, II)
- Cosupervisor: Sietze Kuilman (TU Delft, II)

If you are interested in this position, please send your CV and a brief motivation to L.CavalcanteSiebert@tudelft.nl