Multi-agent simulation of AI hybrid ethics and the problem of hidden normativity

The objective of my talk will be to present first results of ongoing research project which pursues to find an answer to the question of how the emerging technology of human assisting embodied robots can be equipped with a system of simulating the attitudes and moral values of its users using contemporary methods of digital humanities.
This aim is pursued using reference to the theory of a hybrid approach to ethics by Wallach, Smit and Allen (Wallach, Smit, Allen 2004). The mentioned concept is based on the claim that an index of trust in social machines, and thus their identifications a group of beings conceived as “Friends” (from Ihde's triad of Friend, Alien, Foe) can be obtained when the personal and local ethical preferences of the technology user are taken into account in process of decision making by the machines.

Therefore we started to build a simple and effective system for identifying of the ethical preferences of users of human assisting social machines – recognising the explicit and implicit normativity influencing their ethical decisions. In the next step, these identified implicit and explicit normativities should be implemented into a system of their digital simulation. The platform for doing so will be multi-agent AI technology, which is regarded as a so-called complex (compound) artificial intelligence system. It has the advantage of providing specialized, automated entities (agents) that, based on the resources of the particular LLM model and additional tools (such as databases, web search or utility calculation engines), can undertake complex, ethical reasoning tasks that go beyond the standard capabilities of the model, thus providing opportunities for the simplified application of Retrieval-Augmented Generation (RAG) procedures.

Therefore, for specific tasks, it is not necessary to undertake a costly tuning of the entire language model, which requires enormous computing power and time, but rather there can be created a specialised network of agents, which in the RAG process will use the resources of the model and enhance them with projected additional skills through the use of local data. A network of agents will form a system, in this case a system of simulation of ethical preferences of the user, which are supposed to be reflected in a human assisting machine.

The implementation of individual preferences of the user in the machine will be controlled through a procedure of setting up supervising ethical agents based on sets of values and higher-level principles, such as the UN Declaration of Human Rights or the EU Declaration of Fundamental Rights. This process of supervision and mitigation should be supported by training and ongoing modification procedures, based on increasingly deeper personalization, in order to constantly adjust the level of adaptation of the assisting robot to the ethical preferences of the user.

As the final result of the presented project we designed a digital system whose goal is to create a special application for ethical personalisation of any human assisting machine. This app might be uploaded in any machine that the user is intending to use and then cancelled if required – as for instance in case of a rented autonomous car.
The general research hypothesis underlying the presented project refers to the claim by Luciano Floridi (Floridi 2002) about the semantic power of information theory, thanks to which any theoretical issue within the scope of philosophy and other humanities and social

sciences can be transformed into the concept of theory of information. According to this opinion, information theory can even be considered nowadays as the new First Philosophy thanks to its ontological primacy in the Aristotelian sense (information is inherent in every object and process being the object of a cognition), as well as its methodological primacy in the Cartesian sense - its language can be a tool for tackling fundamental philosophical problems, including the possibility of simulation of explicit and implicit ethical normativity. The more specific research hypothesis of the project is that it is possible to establish, with the use of multi-agent AI technology, a digital simulation of the system of conscious and unconscious values and goals of human's ethics and the ethical reasoning based on them. This has been proven in preliminary research which will be object of my talk.
The preliminary research I am going to present tested also the usability of the AutoGen multi-agent technology to meet the project objectives, the availability of data needed to define the hidden ethical preferences used in the local, Polish cultural cluster, existing deontological interpretations of the dominant codex system in the local community, the performance of the utility calculation engine available within existing LLM systems, ethical biases of existing LLM systems, available systems for mitigation of clustered decisions based on existing law and value systems.

The project directly relates to the results of the Moral Machine research (Awada et al. 2018) and the problem of using culturally clustered ethics as a response to the problem of trust in autonomous machines.