Trust in Artificial Intelligence

Artificial intelligence is not just a new technology. It is a powerful force that is transforming our daily practices, personal and professional interactions, and the social environment we live in. It presents us with unparalleled opportunities but also entails novel, sometimes unknown challenges. The major challenge related to the development of artificial intelligence is establishing harmonious human-artificial intelligence relations necessary for harnessing artificial intelligence's potential to use its power for good. To address this challenge, the current project focuses on trust—the critical building block of any society. It aims at discerning the differences between trust in humans and in artificial intelligence agents, and identifying the key psychological factors that determine the development of trust towards artificial intelligence. To date, very little is known about trust or trustworthiness in human-artificial intelligence interactions, and almost nothing about involved psychological mechanisms. This project will contribute to a better understanding of how trust between humans and artificial intelligence agents is established, what makes artificial intelligence agents trustworthy, and how their morality is judged.

The aim of this project is to answer the following research questions: What are the similarities and differences between interpersonal trust and trust towards artificial intelligence? What psychological factors affect people's willingness to trust artificial intelligence? What is the role of morality and competence when assessing the trustworthiness of artificial intelligence? How important is what artificial intelligence did versus what were the consequences of its actions when making moral judgments about artificial intelligence?

The project consists of seven studies, which are divided into three separate lines of research. The first line of research is related to psychological factors affecting trust in artificial intelligence, such as optimism, anxiety or resistance to change. It will explore the psychological features that make people willing to trust artificial intelligence, or reluctant to do so. The second line of research will focus on the two universal dimensions on which people evaluate others – one related to morality, and another related to competence. It will investigate how information about an artificial intelligence agent's morality and competence affects people's willingness to trust them. The third line of research addresses the issue of making moral judgments about artificial intelligence. It will verify if people give more importance to artificial intelligence's actions being morally right (e.g. telling the truth, not stealing, following the rules) versus the consequences of their actions, since a good deed may bring about a catastrophe, while a wrongful act may be done in the name of a greater good. The studies will be conducted both in a laboratory as well as online, and will employ both existing measures and materials, as well as ones designed specifically for the purpose of this research.