Norms in Language-based Human-AI Interaction (NIHAI)

Project goal: The project aims to identify how communication between artificial intelligence (AI) and humans works, pinpointing potential risks and developing strategies to counteract them.

Human communication is governed by certain rules and norms, such as the expectation of sharing information from trustworthy sources. For example, if one propagates false information, one will be called a liar. Similarly, if one transmits unverified information, one will lose the trust of others, and her words risk being neglected in the future. This project investigates whether similar norms of communication apply when people communicate with AI. Moreover, the project examines whether these communication norms could be exploited to spread misinformation and, if so, how to prevent such misuse. As part of this project, we analyze different discourses, such as the communication between journalists and their audiences. The project is carried out by researchers from 4 countries (Austria, Poland, Romania, Switzerland) and will thus bring a cross-cultural perspective on the project goal by collecting data in multiple countries.

Research description: The project investigates six main questions with interdisciplinary tools. The first question aims to identify general and specific norms in human-AI communication and how they differ from purely human communication. For the second question, the research team will examine possible dark moves in communication with AI, such as lying or misleading. Next, the team will test whether people attribute mental states, such as intentions and beliefs, to artificial agents, and how these attributions and their implications need to be understood, for example, in the case of responsibility attribution. The fourth research question investigates the impact of AI's deviation from established communication norms on people’s trust in the AI. The team will also test whether certain features of the AI, for example, if it looks human-like, influence answers to the previous questions. Finally, the project culminates in formulating of ethical principles for responsible design and use of AI-driven conversational agents.

Our team will tackle these questions with new tools such as extensive online surveys, corpus analysis (the study of large collections of texts), and interviews with users of AI-driven apps.

Reasons for undertaking the research: The current development of AI technology creates many new opportunities. However, at the same time, it also creates new risks, such as the spread of misinformation or manipulation through language-driven AI agents. The project aims to address this concern by unveiling how the communication between AI and humans works, what potential risks arise from it, and how they can be mitigated.

Important expected effects: We will formulate principles for responsible design and use of AI-driven conversational agents so as to mitigate potential risks arising from human-AI communication. We will disseminate the findings to large audiences and publish papers in top tier journals.