Taming the Risks of Digital Technologies - Interdisciplinary Collaboration for a Trustworthy Future. Ruhr-University Bochum, 29 – 31 July 2024.

Overview of Speakers, Keynotes and Talks

Krishna P. Gummadi: Foundations for Foundational Models

Abstract: In this talk, I will present our attempts to investigate two related foundational questions about large language models (LLMs): (a) how can we know what an LLM knows? and (b) how do LLMs memorise and recollect information from training data? The answers to these questions have important implications for the privacy of training data as well as the reliability of generated outputs, including the potential for LLM hallucinations. I will propose experimental frameworks to study these questions: specifically, a framework to reliably estimate latent knowledge about real-world entities that is embedded in LLMs and a framework to study the phenomena of recollecting training data via rote memorisation. I will present some (surprising) preliminary empirical results from experimenting with a number of large open-source language models.

Biography: Krishna Gummadi is a scientific director and head of the Networked Systems research group at the Max Planck Institute for Software Systems in Germany. He also holds a professorship at the University of Saarland. Krishna's research interests are in the measurement, analysis, design, and evaluation of complex Internet-scale systems. His current projects focus on understanding and building social computing systems. Krishna's work on fair machine learning, online social networks and media, Internet access networks, and peer-to-peer systems has been widely cited and his papers have received numerous awards. He received an ERC Advanced Grant in 2017 to investigate "Foundations for Fair Social Computing".

Mark Elliot: Data, Automation and the Human Technology Nexus

Abstract: All technological developments have cultural, social and psychological consequences – some intended, some not. What has changed recently is the degree of interaction between new developments; underpinned in part by the interdisciplinary mixing that 21st century academia has increasingly encouraged and in part by humans' fascination with technology that is seemingly driving us along a path of merging with our artefacts. It seems likely that over the next two decades the manifestation of these socio-technological changes will radically alter the nature of society, from individual lives to how we organise ourselves. In this context, that most human of questions – "what sort of society do we want to live in?" takes on new levels of meaning. This talk will consider some of these emerging technologies, their likely trajectories and impact of humans. I will be particularly focusing on TIPS (Trust, Identity, Privacy and Security).

Biography: Mark Elliot (University of Manchester) has been central to the field of confidentiality and privacy since 1996. He is one of the key international researchers in the field of Statistical Disclosure and collaborates widely with non-academic partners. Since 2012, he has led the UK Anonymisation Network, which provides advice, consultancy and training on anonymisation and works closely with the UKs information commissioner's office. Since 2023 he has led SPRITE+ cross sector community which runs events and funds research on all elements of TIPS (Trust, Identity, Privacy and Security). Aside from Privacy his research interests also include AI and Society and substantive social science topics under the broad heading of Psychological Sociology.

Alice Hutchings: Trusting the Untrustworthy

Abstract: Cybercrime is facilitated by anonymous online environments, yet the degree of specialisation required often means there is a need to trade and collaborate with others. This poses a problem: Why trust those who are inherently untrustworthy? We'll explore issues relating to trust and anonymity as they relate to online marketplaces and forums, including a focus on how offenders overcome the cold start problem.

Biography: Alice Hutchings is Professor of Emergent Harms in the Security Group at the Computer Laboratory, University of Cambridge, and Fellow of King's College. She is also Director of the Cambridge Cybercrime Centre, an interdisciplinary initiative combining expertise from computer science, criminology, and law. Specialising in cybercrime, she bridges the gap between criminology and computer science. Generally, her research interests include understanding cybercrime offenders, cybercrime events, and the prevention and disruption of online crime.

Sarah Spiekermann-Hoff: Value Based Engineering for a Better AI Future

Abstract: This talk is going to give an introduction and overview of the Value-based Engineering Method (short: VBE). It is a method to ensure a value-based and ethical IT design and is based on the world's first standardised ethical model process for system design ISO/IEC/IEEE 24748-7000. The different phases of VBE and underlying philosophies are presented and what this would mean for AI design.

Biography: Since 2009 Sarah Spiekermann is chairing the Institute for Information Systems & Society at Vienna University of Economics and Business (WU Vienna). She published several books, including "Value-based Engineering – a guide to build ethical technology for humanity", "Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert", and "Ethical IT Innovation: A Value-based System Design Approach", as well as over 120 articles in leading academic journals. In 2016 Sarah Spiekermann co-founded the "Sustainability Computing Lab". In the same year she also started vice-chairing IEEE's "P7000" project, leading the development of the first model process for ethical system design published in 2021 (also referenced as ISO/IEC/IEEE 24 748-7000).

Ivan Habernal: It's all solved by ChatGPT now, right? Tales from Legal Natural Language Processing

Abstract: We think that contemporary large language models, such as ChatGPT, are so all-mighty that we might believe there's no task they cannot tackle well. But is NLP (natural language processing) really a "solved problem"? What if we are not interested in boring text generation tasks like writing a fake summer school motivation letter, but want to understand legal argument reasoning instead? What if we want to know which legal arguments matter for the courts to decide? What if we want to answer laymen's question in a language that maybe only few persons on Earth really understand (yes, I'm referring to German "legalese")? In this talk, I'm going to address some of these research questions through empirical research lenses.

Biography: Ivan Habernal is a full professor of Fairness and Transparency at Ruhr University Bochum, Germany, jointly affiliated with the Research Center Trustworthy Data Science and Security. He is leading the Trustworthy Human Technologies group which focuses on various aspects of natural language processing (NLP), including privacy- preserving NLP, legal NLP, and trustworthy models.

Bilal Zafar: On Early Detection of Hallucinations in Factual Questions Answering

Abstract: Hallucinations remain a major impediment to the adoption of LLMs. In this work, we explore if the artifacts associated with model generations can provide hints that a response will contain hallucinations. Our results show changes in entropy in input token attribution and output softmax probability for hallucinated outputs, revealing an “uncertain” behavior during model inference. This uncertain behavior also manifests itself in auxiliary classifiers trained on outputs and internal activations, which we use to create a hallucination detector. We further show that tokens preceding the hallucination can predict subsequent hallucinations even before they occur.

Biography: Bilal is a professor of Computer Science at Ruhr University Bochum and the Research Center for Trustworthy Data Science and Security. Before joining RUB, he was a Senior Scientist at Amazon Web Services where he was building products to support trustworthy use of AI/ML. His research interests are in human-centric Artificial Intelligence (AI) and Machine Learning (ML). His work aims to address challenges that arise when AI/ML models interact with human users. For instance, he develops algorithms for making AI/ML models more fair, explainable and robust.