ICT23-030 - Acquiring and explaining norms for AI systems

Zusammenfassung

Künstliche Intelligenz (KI) ist ein wesentlicher Bestandteil unseres Alltags geworden. Sie beeinflusst Kaufentscheidungen, soziale Kontakte, Berufsentscheidungen und sogar die Betreuung von Kindern und älteren Menschen. Daher müssen KI-Systeme die rechtlichen, sozialen und ethischen Normen der Gesellschaften, in denen sie eingesetzt werden, einhalten. Die Maschinenethik widmet sich dieser Herausforderung, indem sie KI-Systeme entwickelt, die diese Normen verstehen und befolgen können. Ein zentrales offenes Problem ist der Erwerb und die Darstellung normativer Informationen in einer maschinell umsetzbaren Form. Das interdisziplinäre AXAIS-Projekt widmet sich dieser Herausforderung. Die Projektleiter Ciabattoni (Logik), Horty (Philosophie & Legal Reasoning) und Mateis (KI) nutzen ihre Expertisen, um Normen für den Einsatz in KI-Systemen zu erarbeiten und die Erklärbarkeit der daraus resultierenden Entscheidungen zu gewährleisten. Der Ansatz kombiniert Methoden aus der Sprachverarbeitung, der Logik und Legal Reasoning, um ein Rahmenwerk zu schaffen, das umfangreiche Normensammlungen (z.B. Straßenverkehrsordnungen) automatisch in eine für KI-Systeme verständliche und erklärbare Form übersetzt. Das angestrebte Rahmenwerk fördert nachvollziehbare Entscheidungsprozesse und ermöglicht es, komplexe normative Informationen von einfachen Entscheidungen abzuleiten, ähnlich der Fallbasierten Argumentation in juristischen Kontexten.

Wissenschaftliche Disziplinen:
Mathematical logic (35%) | Artificial intelligence (50%) | Legal theory (10%) | Philosophy of law (5%)

Keywords:
Deontic Logic; Large Language Models; Normative Reasoning; Answer Set Programming; Common law; AI and Law

---

Principal Investigator:	Agata Ciabattoni
Institution:	TU Wien
Co-Principal Investigator(s):	John Horty (University of Maryland) Cristinel Mateis (AIT - Austrian Institute of Technology)

Status: Laufend (01.12.2024 - 30.11.2028)
GrantID: 10.47379/ICT23030

Weiterführende Links zu den beteiligten Personen und zum Projekt finden Sie unter
https://www.wwtf.at/funding/programmes/ict/ICT23-030/