ICT22-023 - Training and Guiding AI Agents with Ethical Rules

Zusammenfassung

Ethische Normen, die auf Konzepten wie Verpflichtung und Verbot fußen, sind in einer Reihe von Bereichen von großer Wichtigkeit. Sie haben zudem eine Schlüsselrolle, um das Verhalten von mit künstlicher Intelligenz (KI) ausgestatteten autonomen Agenten, die vermehrt in unser Umfeld integriert werden, in geordneten Bahnen zu halten. Solche autonomen Agenten werden in der Regel mittels maschinellem Lernen erstellt, wobei das gewünschte Verhalten aus einer (oft extrem großen) Menge an Beispielen automatisch gelernt wird, ganz im Sinne eines Vorschlags des Informatikpioneers Alan Turing anfangs der 1950er Jahre. Es ist jedoch ein offenes Problem, wie lernende Agenten realisiert werden sollen, die sensitiv mit Normen und Regulierungen umgehen können. Das Projekt TAIGER stellt sich dieser Herausforderung und strebt an, maschinelle Lernverfahren mit logik-basierten Techniken zu erweitern und zu integrieren. Dazu wird TAIGER effektive Rahmenwerke einführen, mittels deren Verwendung autonome Agenten Aktionen legal, norm-sensitiv und sozial akzeptiert ausführen können.

Dabei steht eine transparente Rechtfertigung von gezogenen Schlüssen, Modularität und die Fähigkeit zum Umgang mit Widersprüchen in Normen und mit Situationen im Vordergrund, in denen kein norm-konformes Verhalten möglich ist. Das Verhalten des Agenten soll dazu mit möglichst wenig Korrekturen angepasst werden. Die Resultate von TAIGER sind ein Beitrag zu einer vertrauenswürdigeren und somit menschenfreundlicheren KI.

Wissenschaftliche Disziplinen:
102001 (55%) | 101013 (25%) | 102034 (20%)

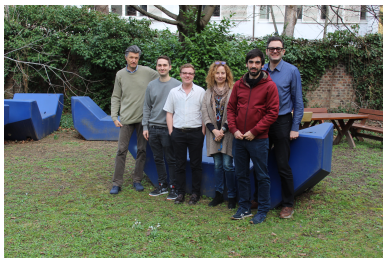
Schlagworte:
Knowledge representation, Deontic Logic, Answer set programming, reinforcement learning, cyber-physical systems

---

Principal Investigator: Agata Ciabattoni

Institution: TU Wien

Co-Principal Investigator(s): Thomas Eiter (TU Wien)
Ezio Bartocci (TU Wien)



Status: Laufend (01.05.2023 – 31.01.2028)

Weiterführende Links zu den beteiligten Personen und zum Projekt finden Sie unter
https://wwtf.at/funding/programmes/ict/ICT22-023/