IACAI

Interfaces of Agent-Centric AI

Programm / Ausschreibung	KS 24/26, KS 24/26, COMET-Module 2024	**Status**	laufend
Projektstart	01.01.2026	**Projektende**	31.12.2029
Zeitraum	2026 - 2029	**Projektlaufzeit**	48 Monate
Keywords	multi-agent-ai-systems, energy efficiency, trustworthy AI, human-ai agent support, LLMs		

Projektbeschreibung

Kurz nach der Veröffentlichung von ChatGPT durch Open AI im November 2022 wurden kritische Stimmen laut. Viele hochkarätige Forscher aus der KI-Gemeinschaft forderten ein Memorandum und damit einen Entwicklungsstopp der generativen KI. Nobelpreisträger Geoffrey Hinton trat Ende Mai 2023 von seinem Job bei Google zurück, und nannte langfristige Risiken der KI als Grund. Trotzdem beschleunigte sich die Entwicklung der KI ungebremst weiter. Wir erwarten, dass zukünftige KI in der Lage sein wird, aktiv und autonom zu handeln. Die KI wird dazu Fähigkeiten entwickeln zur Planung und Nutzung von Werkzeugen und auch von anderen KI-Systemen. Dadurch entwickeln sich AI-Modelle hin zu AI-Agenten. Dies erhöht zwar die Nützlichkeit der KI, allerdings erhöhen sich auch die damit verbundenen Risiken. Wir antizipieren diesen Entwicklungsprozess und schlagen vor, sich über State-of-the-art-Ansätze in der Forschung hinauszubewegen und konkret sich die Schnittstellen und Verbindungen zwischen KI-Agenten und anderen Akteuren, wie Menschen, IoT-Geräten und Wissensbasen, zu untersuchen. Daher nennen wir unser Modul: Interfaces of Agent-Centric Artificial Intelligence (IACAI).

IACAI wird folgende Aspekte abdecken: (i) Algorithmen und Berechnungsverfahren, (ii) Mensch-Computer-Interaktion und (iii) soziale und ethische Überlegungen bezüglich einer zukünftigen, agenten-zentrierten KI. Um die Hauptbarrieren für die Kollaboration von Agenten in gemeinsamen Szenarien zu überwinden, wird unsere Forschung auf folgende Aspekte ausgerichtet: Grundlagen der effizienten Datenübertragung und -verwaltung zwischen Akteuren, Verbesserung der Berechnungseffizienz von KI-Modellen sowie die Steigerung der logischen Schließens von KI-Modellen. Wir adressieren Fragen zur Produktivität, wenn Menschen die KI in komplexen Aufgaben verwenden, und dazu werden wir untersuchen, wie man die synergetische Leistung von Menschen und KI verbessern kann, beispielsweise durch interaktive Designs und explizites Domänenwissen.

Um soziale Aspekte und ethische Technologieentwicklung besser zu verstehen, werden wir KI-Agenten, sowie Interaktionen zwischen den KI-Agenten modellieren, um auf theoretischer und empirischer Ebene zu analysieren, wie Interaktionen und Schnittstellen in agenten-zentrierter KI die Konzepte wie Verantwortung und Vertrauen beeinflussen. Wir werden Algorithmen entwickeln, um Fairness und Diversität bei Themen der Ressourcenzuweisung und Entscheidungsfindung in zukünftigen agenten-zentrierten AI-Systemen zu gewährleisten.

Unsere Forschung ist primär der Disziplin Informatik zugeordnet, wird aber von Theorien anderer Disziplinen wie Psychologie, Sozialwissenschaften und Philosophie inspiriert. Die Forschungsarbeiten innerhalb des IACAI-Moduls werden daher Erkenntnisse aus der Forschung in den Bereichen Daten-Management von großen Datensätzen, maschinelles Lernen, Natural Language Processing, Wissensmodellierung, Mensch-Computer-Interaktion, Informationsvisualisierung, KI-Explainability, sowie Wissenschafts- und Technologiestudien und Fairness in KI mit einbeziehen.

Unsere Forschung wird einen wesentlichen Beitrag liefern, um agenten-zentrierte KI zu einem korrekteren, zuverlässigeren, effizienteren und effektiveren Instrument für Menschen zu machen, welches selbst bei komplexen Aufgaben eine Unterstützung bietet. Die Forschungen innerhalb des IACAI-Moduls wird dabei von einem Code-of-Conduct angeleitet, welches eines der ersten Ergebnisse des Moduls sein wird. Auswirkungen auf die Themen Umwelt- und soziale Auswirkungen werden kontinuierlich bewertet und dokumentiert. Die Ergebnisse von IACAI stellt dann eine Reihe von Prototypen und Leitlinien für die Implementierung agenten-zentrierter KI dar. IACAI wird agenten-zentrierte KI entwickeln, die technische Anforderungen erfüllt und ethische Standards entspricht, die den Leitlinien der Gesellschaft entspricht, und die dabei ein hohes Maß and Vertrauenswürdigkeit aufweist.

Abstract

Soon after ChatGPT was released by Open AI in November 2022, many voices raised concerns. Key researchers of the AI community urged for a memorandum of AI development. Nobel prize winner Geoffrey Hinton resigned from his job at Google in May 2023 citing long-term risks of AI. But AI development accelerated even further. We expect that future AI will increasingly gain the ability to actively and autonomously act, based on capabilities to plan and to utilise tools and other AI systems. Thus, AI models will evolve into AI agents. This increases the usefulness of AI, but crucially also increases the associated risks. Anticipating this development, we propose to go beyond the majority of state-of-the-art research by investigating the interfaces and connections between AI agents and other actants, such as humans, as well as more passive technological components such as IoT devices and knowledge bases. Hence, we coin our proposed module: Interfaces of Agent-Centric Artificial Intelligence (IACAI).
IACAI will address (i) algorithmic and computing, (ii) human-computer interaction, as well as (iii) social and ethical considerations regarding a future, agent-centric AI. To overcome major algorithmic, computational and implementation impediments for (generative) AI in collaborative agent scenarios our research will include work on foundations of efficient data transfer and management across actors, improving the computational efficiency of AI models, and boosting the reasoning capabilities of AI models. To overcome issues of productivity when humans use AI in complex tasks, partially due to the unresolved question of how humans can understand decisions in multi-stakeholder networks, we will investigate how to improve human-AI synergetic task performance via interaction designs that use explicit domain knowledge and will develop algorithms and interfaces for traceability of AI decisions in multi-agent environments. To inform socially responsible and ethical technology design, we will model actors, interaction and agency to analyse, on theoretical and empirical levels, how interactions and interfaces in agent-centric AI shape the notions of responsibility and trust, and will develop algorithms to ensure fairness and diversity in resource distribution and decision-making in the future agent-centric AI that we are envisioning.
Our research is centrally rooted in computer science but is inspired and grounded in theories of other disciplines, including psychology, social sciences and philosophy. Research conducted in this module will therefore draw from and make contributions to research on large-scale data management, machine learning, natural language processing, knowledge modelling, human-computer interaction, information visualisation, explainability, science and technology studies, and bias

and fairness in AI. We will contribute to make agent-centric AI more correct, more reliable, more resource-efficient, and more efficient in supporting humans conduct complex tasks.
Research within IACAI will be executed guided by a code of conduct, which will be one of the initial outcomes of the module. Environmental and societal impact will be continuously assessed and documented. The outcomes of IACAI will be a series of prototypes and guidelines on how to implement agent-centric AI. IACAI will establish agent-centric AI that fulfils technical requirements, as well as conforms to ethical standards as imposed by society and achieves trustworthiness expected by its users.

Projektkoordinator

- Know Center Research GmbH

Projektpartner

- Technische Universität Berlin
- Technische Universität Graz
- Atos Technologies Austria GmbH
- Selmo Technology GmbH
- DAPHOS GmbH
- Universität Graz
- SIATLAB GmbH
- FH OÖ Forschungs & Entwicklungs GmbH
- AVL List GmbH
- Dynatrace Austria GmbH