Embodied AI

Design und Evaluation vertrauenswürdiger und erklärbarer Embodied-AI-Systeme

Programm / Ausschreibung	FORPA, Dissertaionen 2024, Industrienahe Dissertationen 2025	**Status**	laufend
Projektstart	01.03.2026	**Projektende**	28.02.2029
Zeitraum	2026 - 2029	**Projektlaufzeit**	36 Monate
Keywords	AI, Embodied AI, Challenging Contexts		

Projektbeschreibung

Die Dissertation befasst sich mit der Gestaltung vertrauenswürdiger und erklärbarer Embodied-AI Systeme, die als physisch präsente KI-Agenten zunehmend in sicherheits- und verantwortungskritischen Kontexten eingesetzt werden. Während technische Fortschritte in Robotik und KI bereits eine hohe Leistungsfähigkeit ermöglichen, fehlt bislang ein konsistentes Rahmenwerk, das unterschiedliche Grade von Autonomie, menschlicher Aufsicht und Entscheidungsverantwortung integriert.

Ziel der Arbeit ist es, ein wissenschaftlich fundiertes Gestaltungs- und Evaluationsmodell für Interaktion mit Embodied AI zu entwickeln, das Vertrauen, Transparenz und Akzeptanz maximiert. Methodisch kombiniert die Dissertation Literaturanalysen, Prototyping, partizipatives Design und empirische Studien mit Multi-Level-Messung. Drei Use Cases bilden dabei den Rahmen für die Erforschung dieser Ansätze: Interaktion zwischen Menschen und Roboter, Qualitätskontrolle und Mobilität. Die Ergebnisse werden in ein Referenzmodell überführt, das übertragbar auf weitere Domänen ist.

Diese Arbeit entwickelt und validiert daher ein Framework für Embodied AI Interfaces. Untersucht wird, wie unterschiedliche Grade menschlicher Beteiligung (Human-in-the-Loop, Human-on-the-Loop und Human-in-Command) mit adaptiver Autonomie kombiniert werden können, um Verlässlichkeit, Transparenz und Nutzerakzeptanz in kritischen Kontexten zu gewährleisten. Mithilfe eines Multi-Level-Messansatzes integriert die Dissertation Systemleistung, kognitive Belastung, Vertrauen und Teamdynamiken, indem physiologische Indikatoren, Verhaltensmetriken und subjektive Einschätzungen zu einer ganzheitlichen Evaluationsstrategie verbunden werden.

Wissenschaftlich leistet die Dissertation einen Beitrag, indem sie Embodied AI, Human Factors und Interaktionsdesign in einem kohärenten Rahmen vereint. Praktisch liefert sie Gestaltungsstrategien für sichere, akzeptierte und effiziente Embodied AI. Die Ergebnisse bereichern die breitere Debatte um vertrauenswürdige KI und bieten konkrete Orientierung für die Umsetzung von Prinzipien wie Transparenz, Erklärbarkeit und menschlicher Aufsicht.

Abstract

This dissertation investigates the design of trustworthy and explainable Embodied AI systems, increasingly deployed as physically present agents in safety- and responsibility-critical contexts. While advancements in robotics and AI enable high performance, a consistent framework integrating varying degrees of autonomy, human oversight, and decision-making responsibility remains lacking.

This work aims to develop a scientifically grounded design and evaluation model for interaction with Embodied AI that maximizes trust, transparency, and acceptance. The methodology combines literature reviews, prototyping, participatory design, and empirical studies utilizing multi-level measurement. Three use cases – human-robot interaction, quality control, and mobility – provide the context for exploring these approaches. The findings are translated into a reference model transferable to other domains.

This work develops and validates a framework for Embodied AI interfaces, investigating how different levels of human involvement (Human-in-the-Loop, Human-on-the-Loop, and Human-in-Command) can be combined with adaptive autonomy to ensure reliability, transparency, and user acceptance in critical contexts. Utilizing a multi-level measurement approach, the dissertation integrates system performance, cognitive load, trust, and team dynamics by combining physiological indicators, behavioral metrics, and subjective assessments into a holistic evaluation strategy.

Scientifically, this dissertation contributes by uniting Embodied AI, Human Factors, and Interaction Design within a coherent framework. Practically, it provides design strategies for safe, accepted, and efficient Embodied AI, enriching the broader debate on trustworthy AI and offering concrete guidance for implementing principles such as transparency, explainability, and human oversight.

Projektpartner

- AIT Austrian Institute of Technology GmbH