HOFAI

Evaluation of human oversight for AI and development of an assessment tool

Programm / Ausschreibung	FORPA, Forschungspartnerschaften NATS/Ö-Fonds, InDiss FZOE 2022	**Status**	abgeschlossen
Projektstart	01.02.2023	**Projektende**	31.01.2026
Zeitraum	2023 - 2026	**Projektlaufzeit**	36 Monate
Keywords	human oversight, artificial intelligence, user centered AI, EU AI act		

Projektbeschreibung

In der EU wird es bald verpflichtend risikoreiche künstliche Intelligenz (KI, englisch AI) so zu gestalten, dass ihre NutzerInnen sie ausreichend beaufsichtigen können. Von KI-NutzerInnen mit ausreichender Aufsicht wird erwartet, dass sie KI-Risiken "für die Gesundheit, die Sicherheit oder die Grundrechte" verhindern oder minimieren (EU-KI-Gesetz, Europäische Kommission, 2021) und so gute Resultate der Mensch-KI-Interaktion sicherstellen. Während sich das Gesetz derzeit im europäischen Gesetzgebungsverfahren befindet, gibt es für Firmen die KI-Anwendungen entwickeln noch kein valides Instrument und keine Anleitung um die menschliche Aufsicht zu berücksichtigen. Die einzig verfügbare Methode zur Beurteilung von menschlicher Aufsicht ist eine Checkliste, die aber von den EntwicklerInnen auszufüllen ist (High-Level Expert Group on Artificial Intelligence, 2019). Auch die Forschung hat sich bisher nur wenig mit dem Konzept der menschlichen Aufsicht befasst. Teilaspekte (z.B. mentale Modelle, Situationsbewusstsein) sind jedoch weitgehend erforscht. Empirische Befunde zu den Teilaspekten stützen die Annahme, dass die Möglichkeit zur Aufsicht die Ergebnisse der Mensch-KI-Interaktion verbessert.

Das Dissertationsprojekt wird aufbauend auf der aktuellen Literatur und dem vorgeschlagenen EU-KI-Gesetz die menschliche Aufsicht von KI untersuchen. Ziel ist es, die KI-Industrie mit ausreichenden Erkenntnissen, Anleitungen und Werkzeugen auszustatten, um die menschliche Aufsicht bei der Entwicklung von KI-Anwendungen zu berücksichtigen. Zur Erreichung dieses Ziels wird im Rahmen des Dissertationsprojekts ein Messinstrument für menschliche Aufsicht bei der KI-Nutzung entwickelt. Dieses kann in der KI-Entwicklung und -Forschung verwendet werden. Durch einen modularen Aufbau wird das Messinstrument auch eine Beurteilung verschiedener Teilaspekte von menschlicher Aufsicht ermöglichen. EntwicklerInnen erhalten somit die Möglichkeit zu analysieren, welcher Teilaspekt der Aufsicht in ihrer KI fehlt und noch berücksichtigt werden sollte. Darüber hinaus wird im Rahmen des Dissertationsprojekts ein Modell entwickelt, das erklärt, wie das Ausmaß der Aufsicht die Ergebnisse der Mensch-KI-Interaktion beeinflusst. Auf der Grundlage des Modells der menschlichen Aufsicht und des Messinstruments wird das Dissertationsprojekt untersuchen, wie die verschiedenen Aspekte der Aufsicht die Leistung der Mensch-KI-Interaktion beeinflusst. Die gewonnenen Erkenntnisse zur menschlichen Aufsicht von KI werden helfen, um unrealistische Erwartungen der KI-EntwicklerInnen an KI-NutzerInnen zu verhindern. Basierend auf den Forschungsergebnissen wird das Dissertationsprojekt KI-EntwicklerInnen konkrete Gestaltungsrichtlinien für die Berücksichtigung der menschlichen Aufsicht an die Hand geben.

Abstract

Designing high-risk AI in a way that its users can sufficiently oversee it will soon become mandatory in the EU. AI users with sufficient oversight are expected to prevent or minimize AI risks “to health, safety or fundamental rights” (EU AI act, European Commission, 2021), thus ensuring high human-AI interaction performance. However, while the requirement is currently in the European legislative process, there is insufficient support for the AI industry to consider human oversight. The only available method for assessing human oversight is a checklist, which must be completed by the developers. Similarly, research has only occasionally addressed the concept of human oversight. However, sub-aspects (e.g., mental models, situation awareness) are widely researched, and empirical evidence supports the assumption that oversight improves human-AI interaction performance.
The dissertation project will build upon recent literature and the proposed EU AI act to explore human oversight. The aim is to provide the AI industry with sufficient insights, guidance, and tools to address human oversight in AI development. Therefore, the dissertation project will develop an assessment tool for evaluating human oversight in AI development and research. Through a modular approach, the assessment tool will provide insights into overall oversight and its sub-aspects. Thus, developers will get the opportunity to understand which sub-aspect of oversight is missing in their AI and design for it. Furthermore, the dissertation project will develop a model to explain how oversight influences human-AI interaction performance. Based on the human oversight model and assessment tool, the dissertation project will evaluate how oversight and its sub-aspects influence human-AI interaction performance. Understanding human oversight and its influence on human-AI interaction performance will prevent unrealistic expectations from being placed on AI users. The dissertation project will provide AI developers with concrete design guidelines for human oversight that summarize all results. In the long term, the human oversight assessment tool allows evaluating whether an AI application fulfills the upcoming EU requirement of human oversight.

Endberichtkurzfassung

Das Projekt HOFAI („Evaluation of human oversight for AI and development of an assessment tool“) adressierte eine der zentralen Herausforderungen bei der Entwicklung moderner Künstlicher Intelligenz (KI): die im EU AI Act (Artikel 14) gesetzlich geforderte „menschliche Aufsicht“ (Human Oversight) von einem abstrakten Konzept in eine praxistaugliche und messbare Methodik zu überführen.

Ziel des Projekts war es, Werkzeuge und Modelle zu entwickeln, die Entwickler:innen dabei unterstützen, KI-Systeme so zu gestalten, dass Menschen diese sicher verstehen, ihr Vertrauen korrekt kalibrieren und bei Bedarf effektiv eingreifen können.

Im Rahmen des dreijährigen Projekts wurden folgende Hauptergebnisse erzielt:

Entwicklung eines Human Oversight Modells: Es wurde ein Modell ausgearbeitet, das die regulatorischen Anforderungen des EU AI Acts mit etablierten kognitiven und verhaltensbasierten Konstrukten der Mensch-Maschine-Interaktion (Mentale Modelle, Situationsbewusstsein, Vertrauenskalibrierung) vereint.

Modulares Assessment: Basierend auf umfangreichen experimentellen Fahrsimulatorstudien und Analysen bestehender Datensätze wurden Evaluierungswerkzeuge zusammengestellt. Dieses bietet Forschenden und Entwickler:innen evidenzbasierte, kontextspezifische Metriken, um die Qualität der menschlichen Aufsicht während der Mensch-KI-Interaktion präzise zu messen

Methodik für Human-Centered Design: Um Human Oversight nicht erst retrospektiv zu prüfen, sondern Anforderungen früh zu identifizieren, wurde das präventive Workshop-Format der „Human Oversight Requirement Analysis“ entwickelt. Dieses befähigt interdisziplinäre Teams, die Anforderungen an die menschliche Aufsicht bereits in den frühesten Phasen der Technologieentwicklung proaktiv zu definieren und in das Systemdesign zu integrieren.

Wissenschaftlicher und wirtschaftlicher Impact: Die Ergebnisse des HOFAI-Projekts wurden auf internationalen Fachkonferenzen präsentiert und in wissenschaftlichen Journals veröffentlicht. Darüber hinaus fungiert das Projekt als strategischer Enabler für die Industrie. Es verhindert, dass die Vorgaben zur menschlichen Aufsicht von Unternehmen als reine rechtliche „Checkbox“ missverstanden werden. Stattdessen liefert HOFAI die nötigen Werkzeuge, um vertrauenswürdige, sichere und akzeptierte KI-Lösungen zu entwickeln.

Projektpartner

- Virtual Vehicle Research GmbH