Effekt anthropomorpher Gestaltung auf die wahrgenommene Vertrauenswürdigkeit von KI-Assistenzsystem in Leitwarten

Muriel REUTER

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin,
Friedrich-Henkel-Weg 1-25, D-44149 Dortmund

Kurzfassung: Künstliche Intelligenz (KI) ist zu einem allgegenwärtigen Bestandteil des täglichen und zunehmend auch der Arbeitswelt geworden. Dabei spielt die Gestaltung der Mensch-KI Interaktion eine entscheidende Rolle, die in diesem Vorhaben im Hinblick auf den Effekt anthropomorpher Gestaltungselemente auf die wahrgenommene Vertrauenswürdigkeit untersucht werden soll. Positive und negative Konsequenzen sollen dabei in einem Mixed-Methods Ansatz in realitätsnahen Laborsettings erforscht werden. Das Ziel ist die Entwicklung von Richtlinien für die Gestaltung anthropomorpher KI-Systeme im Arbeitskontext.

Schlüsselwörter: Künstliche Intelligenz (KI), Mensch-Computer-Interaktion, Anthropomorphismus, Vertrauen

1. Problemstellung

In diesem Dissertationsvorhaben soll der Effekt anthropomorpher Gestaltung auf die wahrgenommene Vertrauenswürdigkeit von Assistenzsystemen, basierend auf Methoden der Künstlichen Intelligenz (KI), erforscht werden. Als Anwendungskontext wurden komplexe und dynamische Arbeitsfelder ausgewählt.

KI-Systeme haben sich bereits in zahlreichen alltäglichen Szenarien bewährt. Auch im Arbeitskontext gewinnt KI mehr an Bedeutung. Besonders in Umfeldern, in denen umfangreiche Datenmengen verarbeitet und zeitkritische Entscheidungen unter Unsicherheit getroffen werden müssen, ist das Einsatzpotenzial von solchen Systemen gegeben. Ein vielversprechendes Anwendungsgebiet im Arbeitskontext sind Leitwarten, in denen komplexe Tätigkeiten durchgeführt, eine Vielzahl von Informationen verarbeitet und Entscheidungen unter Unsicherheit getroffen werden. Besonders Entscheidungsunterstützende Systeme (decision support systems) bieten sich für diesen Kontext an. In solchen Umgebungen wird die intuitive Bedienbarkeit der Systeme sowie das Vertrauen und die Akzeptanz der Benutzer*innen zu entscheidenden Faktoren für eine erfolgreiche Integration des Systems als Arbeitsmittel.

Mit dem wachsenden Einfluss von KI und den sich daraus eröffnenden Möglichkeiten für die Gestaltung von Systemen gewinnt Anthropomorphismus zunehmend an Relevanz. Im Alltag wird die Tendenz, nicht-menschlichen Agenten menschenähnliche Eigenschaften zuzuweisen (Epley et al. 2007) bereits vielfach als Designmittel für Chatbots und ähnliche Anwendungen genutzt, um eine höhere Akzeptanz und Vertrauen zu erzeugen (Cohen et al. 2021; Jensen et al., 2021). Die Kehrseite des Anthropomorphismus findet sich im u. a. Uncanny-Valley Effekt (Mori et al. 2012) wieder, der durch ein Gefühl von Unwohlsein zu einer ablehnenden Haltung gegenüber dem System führen kann.

Für das Verständnis von Vertrauen soll auf die Definition des interpersonalen Vertrauens von Mayer et al. (1995) zurückgegriffen werden. Diese beschreibt Vertrauen als Bereitschaft einer Partei (Trustor), sich für die Handlungen einer anderen (Trustee) verwundbar zu machen, ohne diese andere Partei zu überwachen oder zu kontrollieren, vor dem Hintergrund von Risiko und Bedeutsamkeit des Ergebnisses. Die Anwendung dieses theoretischen Modells scheint im Hinblick auf das psychologische Phänomen der Vermenschlichung von Technik angebracht, wird jedoch auch durch die adaptierte Version von Lee und See (2004) unterstützt, die die zentralen Bestandteile von Unsicherheit und Verletzbarkeit des Trustors beibehält. Auch die Unterscheidung zwischen tatsächlicher und wahrgenommener Vertrauenswürdigkeit soll Betrachtung finden (Schlicker et al. 2022). Negative Konsequenzen aufgrund von Fehlverhalten durch ein unangemessenes Maß an Vertrauen sollte bei der Gestaltung von Mensch-KI Interaktionen berücksichtigt werden (Parasuraman et al. 1993; Parasuraman & Riley 1997). Sowohl zu niedriges (under-trust) als auch zu hohes (over-trust) Vertrauen kann potenziell schädliche Auswirkungen auf die Sicherheit und das Wohlbefinden der Arbeitnehmenden haben.

Obwohl der Anthropomorphismus in verschiedenen Kontexten, insbesondere im Zusammenhang mit Robotern, intensiv erforscht wurde, gibt es bisher nur begrenzte Erkenntnisse über seine spezifische Auswirkung auf das Vertrauen in KI-Assistenzsysteme, insbesondere im Arbeitsumfeld. Daher stellt die vorliegende Dissertation einen bedeutenden Beitrag zur interdisziplinären Arbeitswissenschaft dar, indem sie diese Forschungslücke adressiert und die Relevanz anthropomorpher Gestaltung für die wahrgenommene Vertrauenswürdigkeit von KI-Assistenzsystemen im Arbeitskontext herausarbeitet.

2. Zielsetzung und Fragestellung

Durch eine vertiefte Analyse der zuvor genannten Aspekte strebt die vorliegende Arbeit an, Erkenntnisse zu generieren, die nicht nur theoretisch fundiert sind, sondern auch praktische Implikationen für die Gestaltung und Implementierung von KI-Assistenzsystemen in realen Arbeitsumgebungen bieten. Dabei sollen folgende Forschungsfragen bearbeitet und beantwortet werden:

1) Wie ist der aktuelle Stand der Forschung über die Beziehung zwischen Anthropomorphismus und Vertrauen?
2) Welcher Zusammenhang besteht zwischen Anthropomorphismus und Vertrauen bei verschiedenen Nutzergruppen?
3) Welche menschlichen/persönlichen Faktoren spielen bei der Wahrnehmung von Anthropomorphismus eine Rolle?
4) Erhöht Stress/höherer cognitive load den Einfluss von Anthropomorphismus auf das Systemvertrauen durch den Rückgriff auf Heuristiken?
5) Kann Anthropomorphismus adaptiv kalibriert werden, abhängig von der Leistung bzw. Angemessenheit des Vertrauens?
6) Welche Gestaltungsprinzipien im Hinblick auf Anthropomorphismus lassen sich für angemessenes Vertrauen ableiten?

Ziel der Dissertation ist es, ein tiefergehendes Verständnis für den Effekt von Anthropomorphismus auf das Vertrauen zu generieren, um einen Leitfaden für die Gestaltung von menschenähnlichen KI-Systemen zu entwickeln. Dabei sollen interindivi-

duelle Faktoren wie Alter, Geschlecht und Persönlichkeitsmerkmale (z. B. die interaktionsbezogene Technikaffinität) berücksichtigt werden. Nicht nur sicherheitsrelevante Aspekte, sondern auch der Erkenntnisgewinn bezüglich langfristiger Auswirkungen auf die Gesundheit und das Wohlbefinden der Beschäftigten durch beispielsweise Technikstress sollen dafür in Betracht gezogen werden. Diese Richtlinien sollen helfen, eine menschgerechte Integration von KI-Systemen am Arbeitsplatz sicherstellen zu können.

3. Überlegungen zum methodischen Vorgehen

Zunächst soll eine systematische Literaturrecherche (September 2023 bis Januar 2024) nach den PRISMA-Richtlinien für Scoping Reviews (Tricco et al. 2018) einen Überblick über den aktuellen Forschungsstand geben. Eingeschlossen wurden originale Studien. Als Ausschlusskriterium galt der Kontext von Robotern, intelligenten Fahrzeugen, Kundenservice oder andere Dienstleistungsformen. Die ersten Ergebnisse deuten darauf hin, dass sich bisher nur wenige Studien systematisch mit dem Effekt von Anthropomorphismus auf das Vertrauen beschäftigt haben. Die Minderzahl der Studien erforschte dies durch eine tatsächliche Interaktion zwischen Probanden und System, nur eine Studie wies einen (militärischen) Arbeitskontext auf. Die Ergebnislage zeigt sich stark heterogen und uneindeutig. Sowohl positive als auch negative Auswirkungen von Anthropomorphismus auf das Vertrauen wurden berichtet.

Das weitere Forschungsvorhaben umfasst Online- und Laborstudien, deren genaue Ausgestaltung noch offen ist. Mit einer online Pilotstudie soll zunächst die Manipulation verschieden anthropomorphisierter KI-Agenten validiert werden. Faktoren wie die wahrgenommene Vertrauenswürdigkeit und den Grad von Anthropomorphismus soll subjektiv bewertet werden.

In den folgenden Laborstudien soll ein realitätsnahes Labor-Setting umgesetzt werden, um eine hohe ökologische Validität zu erreichen. In der ersten Laborstudie soll anhand einer beispielhaften Arbeitstätigkeit der Effekt anthropomorpher Gestaltung auf das Vertrauen untersucht werden. Der anzunehmende moderierende Effekt von Faktoren wie Alter, Geschlecht und interaktionsbezogene Technikaffinität soll dabei miterfasst werden.

Die zweite Studie soll, aufbauend auf dem Setup und den Ergebnissen der ersten Studie, durch eine abgewandelte Tätigkeit den Effekt von Multitasking und erhöhtem cognitive load erheben.

In einer dritten Laborstudie könnte, abhängig von der Ergebnislage der ersten Studien, die adaptive Gestaltung von Anthropomorphismus getestet werden. Je nach gezeigtem Vertrauen könnte durch graduelle Anpassung (höheres, niedrigeres Level von Anthropomorphismus) das Verhalten der Benutzer*innen angepasst werden, um ein angemessenes Level von Vertrauen zu erreichen. Über den Einsatz von Machine Learning Verfahren zu diesem Zweck kann diskutiert werden.

Unterschiede in der Erwartungshaltung gegenüber dem Agenten in einem direkten Vergleich zwischen Mensch-Mensch und Mensch-KI Interaktion könnte ebenfalls Bestandteil der Studien werden. Dies könnte ebenfalls Rückschlüsse auf die Rollenwahrnehmung als Teammitglied oder Arbeitsmittel in Teaming-Prozessen erlauben.

Für die Laborstudien wird ein Mixed-Methods Ansatz verfolgt, der die Verwendung von Fragebögen, wie dem Godspeed-Fragebogen (Bartneck et al. 2009) oder der ATI-Skala (Franke et al. 2018), und (neuro-) physiologischen Methoden und Performanz-

Indikatoren einschließt. Die genaue Operationalisierung der Studien soll auf der Grundlage der systematischen Literaturrecherche aufbauen und wird noch diskutiert. Dies schließt die genaue Gestaltung des KI-Agenten und das Designlevel des Anthropomorphismus ein.

4. Literatur

Bartneck C, Kulić D, Croft E, Zoghbi, S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1, 71–81.

Cohen MC, Demir M, Chiou EK, Cooke NJ (2021) The dynamics of trust and verbal anthropomorphism in human-autonomy teaming. In 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 1–6.

Epley N, Waytz A, Cacioppo JT (2007) On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864–886.

Franke T, Attig C, Wessel D (2018) A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. International Journal of Human–Computer Interaction, 35(6), 456–467.

Jensen T, Khan MMH, Fahim MAA, Albayram Y (2021) Trust and anthropomorphism in tandem: the interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. Designing interactive systems conference 2021, 1470–1480.

Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Academy of management review, 20(3), 709–734.

Mori M, MacDorman KF, Kageki N (2012) The uncanny valley [from the field]. IEEE Robotics & automation magazine, 19(2), 98–100.

Parasuraman R, Molloy R, Singh IL (1993) Performance consequences of automation-induced 'complacency'. The International Journal of Aviation Psychology, 3(1), 1–23.

Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

Schlicker N, Baum K, Uhde A, Sterz S, Hirsch MC, Langer M (2022) A micro and macro perspective on trustworthiness: Theoretical underpinnings of the Trustworthiness Assessment Model (TrAM). PsyArXiv Preprints, PsyArXiv: 10.31234/osf.io/qhwvx.

Tricco AC, Lillie E, Zarin W, O’Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Straus SE (2018) PRISMA Extension for Scoping Reviews (PRISMA-SCR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473.

Gesellschaft für Arbeitswissenschaft e.V.

Arbeitswissenschaft in-the-loop

Mensch-Technologie-Integration und ihre Auswirkung auf Mensch, Arbeit und Arbeitsgestaltung

70. Kongress der
Gesellschaft für Arbeitswissenschaft e.V.

Institut für Arbeitswissenschaft und
Technologiemanagement IAT
Universität Stuttgart

In Zusammenarbeit mit dem Fraunhofer-Institut für
Arbeitswirtschaft und Organisation IAO

06. – 08. März 2024

GfA-Press

Bericht zum 70. Arbeitswissenschaftlichen Kongress vom 06. – 08. März 2024

Institut für Arbeitswissenschaft und Technologiemanagement (IAT), Universität Stuttgart
In Zusammenarbeit mit: Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Sankt Augustin: GfA-Press, 2024
ISBN 978-3-936804-34-8

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle (s. u.) erhältlich.

Geschäftsstelle der GfA
Simone John, Tel.: +49 (0)30 1300-13003, Alte Heerstraße 111, D-53757 Sankt Augustin
info@gesellschaft-fuer-arbeitswissenschaft.de · www.gesellschaft-fuer-arbeitswissenschaft.de

Screen design und Umsetzung
© 2024 fröse multimedia, Frank Fröse,
office@internetkundenservice.de, www.internetkundenservice.de