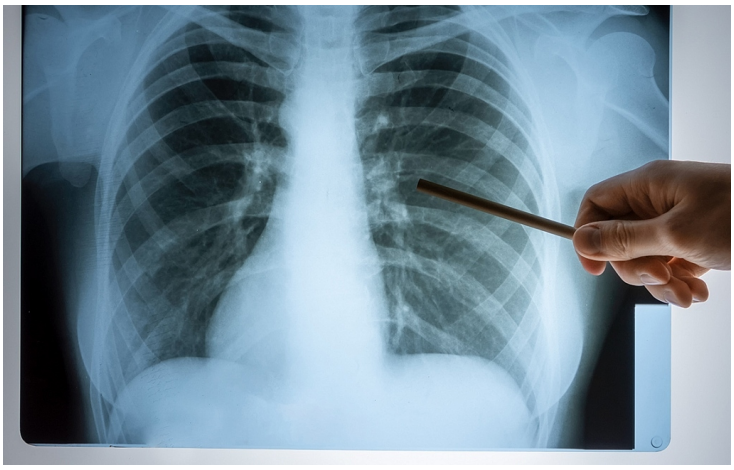
Studie Vertrauen & KI: Expert:innen oft zu skeptisch, Laien oft zu vertrauensselig



Fachkräfte nehmen KI-Empfehlungen oft nicht ernst genug. Privat trauen Menschen der KI jedoch oft zu viel zu - v. a. in Moralfragen. Dies zeigen Studien der Uni Hohenheim

KI-Systeme übertreffen häufig Expert:innen in der Bilderkennung. So können sie beispielsweise Tumore inzwischen oft schneller erkennen als ausgebildetes Fachpersonal. Trotzdem sind Expert:innen in Medizin und anderen Entscheidungskontexten im Gebrauch von KI zögerlich, so die Ergebnisse der Forschung von Wirtschaftsethiker Prof. Dr. Matthias Uhl und seinem Team von der Universität Hohenheim in Stuttgart. Anders gestalte sich die Situation im privaten Kontext: Wenn Nutzer:innen Chatbots als persönliche Berater für Lebensfragen verwenden, entwickeln sie häufig ein übermäßiges Vertrauen in die KI. In beiden Fällen arbeitet Prof. Dr. Uhl an Lösungen, die Abhilfe schaffen.

Trotz Warnung von KI-Expert:innen griffen Nutzer:innen ganz selbstverständlich auf KI zurück, wenn es um moralische oder persönliche Orientierung geht, so eines der Forschungsergebnisse von Prof. Dr. Uhl. "Dabei unterschätzen sie häufig, wie stark die Antworten die eigene Meinungsbildung beeinflussen können. Das überhöhte Vertrauen zeigt sich nicht zuletzt darin, dass Nutzer:innen im Nachhinein glauben, sie wären auch ohne Unterstützung zum gleichen Ergebnis gekommen."

Die Sorge der Expert:innen bestehe daher weniger vor dem Einfluss selbst, sondern mehr darin, dass die Beeinflussung durch die KI unbemerkt geschieht. Prof. Dr. Uhl betont: "Genau diese fehlende Wahrnehmung ist problematisch. Die zentrale Herausforderung besteht darin, Strategien zu finden, mit denen Menschen KI bewusst und reflektiert einsetzen und dabei die Verantwortung für ihre eigene Entscheidung behalten."

Denn: Chatbots hätten selbst kein Wertesystem, nachdem sie ihre Entscheidungen ausrichten. "Je nach Prompt können sich die Moralvorstellungen der KI verschieben."

Wo Laien blind vertrauen neigen Expert:innen zu überhöhter Skepsis

Den gegenteiligen Effekt von blindem Übervertrauen beobachte Prof. Dr. Uhl bei Fachpersonal und deren Umgang mit KI-Tools: "Fachleute wie beispielsweise medizinisches Personal, sind von ihren eigenen Fähigkeiten oft sehr überzeugt und begegnen dem Einsatz von KI im Berufsalltag eher mit Selbstüberschätzung und KI-Skepsis - was ebenfalls nachteilige Folgen haben kann."

Denn diese Skepsis kann riskant sein. "Wenn es etwa darum geht, Röntgenbilder, MRTs oder andere medizinische Bilder zu beurteilen, ist die KI inzwischen meist besser als jeder Mensch. So kann sie Medizinerinnen und Mediziner bei Diagnosen unterstützen." Richtig eingesetzt verbessere die KI deshalb die Qualität medizinischer Entscheidungen.

Das Vertrauensproblem ist nicht auf das Feld der Medizin beschränkt. Prof. Dr. Uhl erklärt: "Je stärker sich jemand als fachlich überlegen einschätzt, desto eher werden Empfehlungen von KI-Modellen ignoriert."

Nur die richtige KI-Gestaltung schafft das richtige Maß an Vertrauen

Der Lösungsansatz für beide Probleme ist im Kern ähnlich. KI muss als Entscheidungshilfe verstanden und eingesetzt werden.

Um sowohl ein zu großes Vertrauen als auch eine zu große Skepsis gegenüber KI-Systemen zu vermeiden, braucht es eine bewusste und reflektierte Nutzung der Technologie. Dabei ist entscheidend, dass die Verantwortung weiterhin der Mensch trägt. Uhl betont: "Die Technik muss dafür so gestaltet sein, dass sie auch genau das leistet, wofür sie entwickelt wurde."

In Lebens- und Moralfragen: Rückfragen statt platter Antworten

Bei Lebens- und Moralfragen wäre die KI hilfreicher, wenn sie statt platter Anweisungen auf den sogenannten sokratischen Ansatz zurückgreifen würde: "Wir haben Chatbots getestet, die bei solchen Fragen Rückfragen an die User stellen, die den Personen helfen, selbst Klarheit zu finden."

Die KI fungiert dabei nicht als moralischer Kompass, sondern als kritisch nachhakende Instanz, die das eigene Wertesystem schärft.

Praxistest zeigt Schärfung des moralischen Wertesystems

Dass diese Schärfung des Wertesystems funktioniert, hat seine Arbeitsgruppe unter anderem an einem Extrembeispiel getestet:

"Für eine Studie in den USA haben wir Republikaner und Demokraten als Testpersonen ausgewählt, die einander in den vergangenen Jahren zunehmend feindselig begegnet sind. Als Testsituation sollten sie Steuern umverteilen und hatten dabei die Möglichkeit, die Menschen der eigenen Partei zu bevorzugen."

Zuvor bekamen die Testteilnehmer:innen noch die Anweisung, sich einige Minuten mit einem Chatbot zu unterhalten. "Was die Probanden nicht wussten: Auch das war Teil der Studie. Die Hälfte der Bots waren klassisch programmiert. Die andere war so konfiguriert, dass sie in Moralfragen die sokratische Methode anwandten."

Letzteres blieb nicht ohne Wirkung: "Probanden, die mit klassischen Chatbots interagiert hatten, verteilten die Steuern eher unmoralisch um, indem sie Parteigenossen bevorzugten." Die andere Gruppe habe meist ethischer gehandelt und Steuergelder eher unabhängig von der Parteizugehörigkeit umverteilt. "Wir konnten sehen, dass Chatbots helfen können, Verständnis für die andere Seite zu schaffen," beschreibt Uhl die Ergebnisse.

Im Umgang mit Expert:innen: KI muss ihre Entscheidungen erklären

Den gleichen Grundsatz aber eine ganz andere Methode verfolgt Prof. Dr. Uhl, im Fall der Expert:innen. Auch hier kommt es bei der KI darauf an, sie zur Zusammenarbeit speziell zu gestalten.

Für die konkrete Umsetzung arbeitet der Wirtschaftsethiker als Teil eines Teams mit Kognitionswissenschaftlerinnen und Informatikerinnen des Bayerischen Forschungsinstitut für Digitale Transformation (bidt) im Projekt "Ethische Implikation hybrider Teams aus Mensch und KI-System" (Ethyde). Derzeit verfolgen die Forscherinnen und Forscher im Projekt noch verschiedene Lösungsansätze.

"Ein Ansatz ist die sogenannte Explainable-AI-Methode. Dabei muss das System möglichst transparent machen, warum es beispielsweise die medizinischen Aufnahmen als unauffällig ansieht oder eine Krankheit diagnostiziert", erklärt Dr. Dr. Sebastian Krügel, der als Postdoc im Projekt arbeitet. Das kann zum Beispiel dadurch geschehen, dass die KI die Bereiche eines Bildes einfärbt, die für ihre Diagnose ausschlaggebend waren.

Eine weitere Möglichkeit ist es, sogenannte Unsicherheitsmaße mitzuliefern, ergänzt Anja Bodenschatz, die als Doktorandin ebenfalls im Projekt mitarbeitet: "Hier liefert die KI eine Einschätzung darüber, wie sicher sie sich bei einer Antwort ist." Beide Ansätze steigern das Vertrauen in die KI-Antwort.

Praxistest soll Prototyp für Medizin-KI entwickeln

Auch in diesem Fall wählte die Arbeitsgruppe einen ungewöhnlichen Weg, um die laufenden Entwicklungen zu testen. "Das Problem, dass sich Personen überschätzen, beschränkt sich schließlich nicht auf Ärzte. Deswegen haben wir einen Test gesucht, den wir bei einem breiten Bevölkerungsspektrum anwenden können", erklärt Dr. Dr. Krügel.

Statt medizinische Bilder bekommen Testpersonen nur Katzenbilder zu sehen. "Die Aufgabe ist, die Bilder in drei Kategorien einzuteilen: Hauskatzen, Wildkatzen und Großkatzen."

Ähnlich wie bei der medizinischen Bilddiagnose komme es auch hier in den Grenzfällen zu den meisten Fehlern. "Kaum eine Testperson widerspricht der KI, wenn sie eine Hauskatze von einem Tiger unterscheidet. Die Meinungsverschiedenheiten treten dann auf, wenn es gilt, zwischen großer Hauskatze und Wildkatze zu unterscheiden oder zwischen Wildkatze und einem jungen Puma."

Sobald das Projekt die KI so angepasst hat, dass Mensch und Maschine harmonieren, will das Team diese Technik mit medizinischen Bildern und Fachpersonal auf Medizinkongressen testen. "Bis Ende 2027 wollen wir die ersten getesteten Prototypen vorstellen können", so Prof. Dr. Uhl.

HINTERGRUND: Ethyde-Projekt

Das Ethyde-Projekt (Ethische Implikation hybrider Teams aus Mensch und KI-System) ist ein Forschungsprojekt des bidt ? Bayerischen Forschungsinstitut für Digitale Transformation, das gemeinsam mit der Universität Hohenheim durchgeführt wird. Ethyde untersucht, wie Menschen und KI-Systeme in sogenannten ?hybriden Teams? zusammenarbeiten und stellt die ethischen Fragen, die dabei entstehen in den Fokus:
https://www.bidt.digital/forschungsprojekt/ethische-implikationen-hybrider-teams-aus-mensch-und-ki-system-ethyde/

Weitere Informationen

Projekt Ethyde
https://www.bidt.digital/forschungsprojekt/ethische-implikationen-hybrider-teams-aus-mensch-und-ki-system-ethyde/
Studie "Justification optional: ChatGPT's advice can still influence human judgments about moral dilemmas"
https://link.springer.com/article/10.1007/s43681-026-01005-6
Studie "ChatGPT's inconsistent moral advice influences users' judgment"
https://www.nature.com/articles/s41598-023-31341-0