Themenblock Mensch-KI-Interaktion: 03_03Implikation_Biaswirkung

Fallbeispiele aus der Praxis

Erarbeitet von
Dr. Maike Mayer

Lernziele

- Du kannst Beispiele einem Phänomen zuordnen
- Du kannst einschätzen, wie sich die vorgestellten Phänomene auswirken können

Projekt ai4all „KI für Alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Wir schreiben das Jahr 2023. In den USA muss sich ein Rechtsanwalt einer Gerichtsanhörung unterziehen, weil er in einem Rechtsstreit auf nicht existierende Fälle verwiesen hat [1]. Man könnte jetzt meinen, hier hat einfach jemand versucht, mit erfundenen Fällen einen Rechtsstreit zu gewinnen – also schlicht zu betrügen. Aber der Fall liegt etwas anders, denn hier war Künstliche Intelligenz im Spiel. Genauer gesagt: ChatGPT. Und auch Vertrauen spielt hier eine Rolle.

Quelle [1]

Einblendung: Icons (Waagschalen, durchgestrichene Dokumente, Teufel); Schlagworte („Künstliche Intelligenz", „ChatGPT")

Anhand von diesem und einem weiteren Fallbeispiel schauen wir uns einmal näher an, wie sich unser Vertrauen in KI auf unsere Nutzung der KI auswirken kann. Dabei werden wir auch einige Faktoren identifizieren, die unser Vertrauen beeinflussen oder die dabei helfen können, dass wir angemessen mit KI umgehen.

Einblendung: Icon (Pfeil, Glühbirne); Schlagworte („Vertrauen", „Nutzung")

Jetzt aber erstmal zurück in den Gerichtssaal …

Fallbeispiel 1: ChatGPT im Gerichtssaal

Steven A. Schwarz hat im Sommer 2023 Schlagzeilen gemacht [1]. Aber was genau war eigentlich passiert? In einem Rechtsstreit zwischen einem Mann und einer Airline stellten sich angeführte Rechtsfälle als frei erfunden heraus. Das besondere hieran: Das Schreiben, in dem diese Fälle aufgeführt wurden, hatte Herr Schwarz von ChatGPT schreiben lassen. Und genau hier entstand das Problem. ChatGPT sagt bei all seiner Raffinesse – verkürzt und vereinfacht ausgedrückt – im Prinzip nämlich einfach das wahrscheinlichste nächste Wort in einem Satz vorher. Und nach diesem Schema entstehen dann fiktive Referenzen oder – wie hier – fiktive Gerichtsfälle. Herr Schwarz hatte sich darauf verlassen, dass ChatGPT echte Fälle kennt und diese einbeziehen würde. Wie eine große Suchmaschine quasi [1].

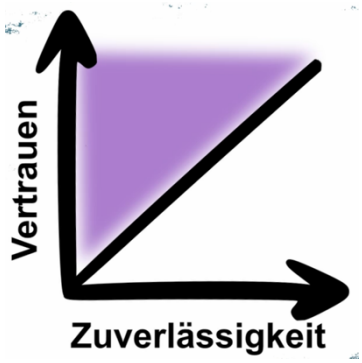
Quelle [1]

Einblendungen: Icons (Waage, Mann, Flugzeug, durchgestrichene Dokumente, schreibender Stift, Text mit Lücke und Fragezeichen, Mappe mit Haken, Lupe); Schlagwort („ChatGPT")

Na, welche Vertrauensphänomene könnten hier eine Rolle spielen? Du kannst das Video kurz pausieren und darüber nachdenken.

Einblendung: Icons (Fragezeichen, Pausenknopf)

Die Auflösung folgt jetzt: Wir befinden uns hier im Bereich von „zu viel Vertrauen". Was hier passiert ist, kann man sehr gut mit „Complacency" bzw. „Automation Bias" beschreiben. Herr Schwarz hat ChatGPT scheinbar so sehr vertraut, dass er sich nicht die Mühe gemacht hat, die Empfehlung der KI – hier in Form eines Textes und einer Auswahl an Fällen – zu überprüfen. So fiel die fehlerhafte Empfehlung der KI erst später auf – unglücklicherweise für Herrn Schwarz den gegnerischen Anwälten.



Einblendung: Icons (pfeifendes Männchen, Dokumente, Figur mit erhobener Hand); Grafik mit markiertem Bereich „zu viel Vertrauen"

Aber woher kam dieses übermäßige Vertrauen? Im Rahmen der Anhörung hat Herr Schwarz gesagt, er habe nicht verstanden, dass ChatGPT Fälle erfinden könnte [1]. Herrn Schwarz war nicht klar, wie ChatGPT funktioniert. Also, was es kann und was es nicht kann. Das hat zum einen dazu geführt, dass er dem System zu viel vertraut hat. Zum anderen hat es dazu geführt, dass er es für etwas eingesetzt hat, wofür das System nicht gedacht war. In diesem Fall Rechtsgutachten sachlich und fachlich korrekt zu schreiben.

Quelle [1]

Einblendungen: Icons (Figur mit Fragezeichen, Daumen hoch, Daumen runter, Figur mit Idee)

Haben wir – wie Herr Schwarz – eine fehlerhafte Vorstellung davon, wie ein System funktioniert, können wir es nicht richtig einschätzen. Daraus kann dann auch eine zu hohe Erwartung an das resultieren, was die Maschine kann. In solchen Fällen hilft es, aufzuklären [2]. Mögliche Varianten könnten Informationen sein, Warnhinweise oder auch Schulungen. Verstehen wir, wie eine Maschine funktioniert, hilft es uns dabei, unser Vertrauen anzupassen und die Maschine angemessen einzusetzen.

Quelle [2]

Einblendung: Icons (denkende Figur, Gedankenblase, Laptop, Kreuz, Glühbirne, Zettel, Hinweisschild, Figuren an Flowchart, Daumen hoch)

Fallbeispiel 2: KI bei Notrufen

Das zweite Fallbeispiel ist nicht ganz so spektakulär wie das erste. Gegenstand ist der Einsatz von KI bei der Annahme von Notrufen. Genauer gesagt geht es um ein dänisches KI-System, das bei Notrufen „mithört" und nach Anzeichen für einen Herzstillstand sucht [3-4].

Entdeckt das System entsprechende Anzeichen wie beispielsweise fehlende Atmung, zeigt es einen Alarm auf dem Bildschirm der Notrufmitarbeitenden an, die dann Folgefragen stellen, um die Diagnose abzuklären. Tatsächlich attestiert eine Studie dem System, Herzstillstände mit höherer Sensitivität zu entdecken – also bei einem tatsächlichen Herzstillstand diesen auch zu erkennen – und das auch schneller als die Notrufmitarbeitenden [4]. Es identifiziert aber durchaus auch häufiger mal fälschlicherweise einen Herzstillstand. Soweit, so gut. Warum ist das jetzt ein interessantes Fallbeispiel?

Quelle [3] [4]

Einblendung: Icons (Medizin, telefonierende Figur, Herz, Alarm, Lupe, Uhr, Fragezeichen)

Interessant ist es, weil der Einsatz des KI-Systems nicht zu einer höheren Erkennungsrate von Herzstillständen durch die Notrufmitarbeitenden geführt hat [5]. Schneller wurden die Mitarbeitenden auch nicht. Und das, obwohl das System eigentlich Herzstillstände eher und schneller erkennt als ein Mensch. Woran könnte das liegen? Und was könnte das mit Vertrauen zu tun haben? Du kannst das Video wieder kurz pausieren und darüber nachdenken.

Quelle [5]

Einblendung: Icons (Glühbirne, durchgestrichene Lupe, durchgestrichene Uhr, Pause-Knopf)

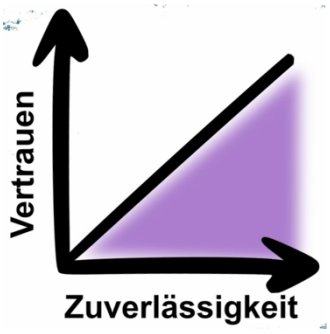
Die Auflösung folgt jetzt: Ein Hinweis auf ein mögliches Problem findet sich tatsächlich in der Leistungsfähigkeit des Systems. Das System erkennt zwar Herzstillstände eher als ein Mensch, aber das System zeigt leider trotzdem in einigen Fällen einen Alarm an, obwohl gar kein Herzstillstand vorliegt [5]. Hier könnte also das „Cry Wolf"-Phänomen eine Rolle spielen, dass auch als „Alert Fatigue" oder Alarmermüdung bezeichnet wird. Das wäre vor allem dann der Fall, wenn die zusätzlich durch das System entdeckten Herzstillstände die zusätzlichen Falschdiagnosen des Systems überwiegen würden, es den Nutzenden aber nicht so vorkommt. Das System wäre also eigentlich trotzdem nützlich, die Nutzenden sehen aber nur die Fehler. Es wäre bei diesem Fallbeispiel also durchaus denkbar, dass die Menschen dem System möglicherweise nicht oder nur sehr wenig vertrauen [6], weil es oft warnt, obwohl eigentlich gar nichts ist.

Quelle [5] [6]

Einblendung: Icons (Figur mit Idee, Kreuz, Wolf); Schlagworte („Entdeckungen", „Fehler", „>"); Grafik mit markiertem Bereich „zu wenig Vertrauen"

Wie geht man damit um? Auch hier könnten Schulungen und Aufklärung helfen [3]. Wie funktioniert das System und wie gut erkennt es Herzstillstände tatsächlich? Das hilft mir, einzuschätzen, wie sehr ich mich auf das System verlassen kann. Ein weiterer Ansatz könnte auch mehr Transparenz sein [4]. Wenn die Mitarbeitenden mehr Informationen darüber

hätten, wieso das System einen Alarm ausgegeben hat, könnten sie die Diagnose besser einschätzen.



Quelle [3] [4]

Einblendung: Icons (Figuren an Flowchart, denkende Figur, Lupe)

Ausblick & Fazit

Du hast zwei Fallbeispiele kennengelernt, bei denen möglicherweise zu viel oder zu wenig Vertrauen in der Realität den Einsatz bzw. die Nutzung eines KI-Systems beeinflusst haben. In beiden Beispielen ist ein größerer Schaden glücklicherweise ausgeblieben. Für Herrn Schwarz ist der Imageschaden vermutlich enorm, aber es wurde niemand verletzt, benachteiligt oder ungerecht behandelt. Der Einsatz des Notruf-KI-Systems hat ebenfalls keinen Schaden verursacht, aber leider auch nicht zu einer Verbesserung der Ausgangslage geführt. Allerdings muss das nicht unbedingt immer der Fall sein. Wenn wir KI-Systemen in anderen Szenarien zu viel oder zu wenig vertrauen, kann das sehr ernste Konsequenzen haben. Wir könnten beispielsweise fälschlicherweise Bewerbende ablehnen, Menschen fälschlicherweise eines Verbrechens beschuldigen und so weiter.

Einblendung: Icons (Waage, Medizin, Ausrufezeichen, erschrecktes Gesicht); Schlagworte („zu viel/wenig Vertrauen”)

Daher ist es wichtig, unsere Interaktion mit KI-Systemen zu überprüfen und zu reflektieren. Verlasse ich mich zu sehr auf das System? Vertraue ich dem System zu wenig, obwohl es mir helfen könnte? Setze ich das System angemessen ein? Bei der Beantwortung dieser Fragen könnten unter anderem Informationen über die Funktionsweise des Systems helfen, Trainings und Schulungen über den Umgang mit dem System oder auch eine transparente Systemgestaltung. Darüber wird es für mich nämlich leichter zu prüfen, ob ich das System angemessen einsetze. Voraussetzung dafür ist natürlich, dass mir die entsprechenden Informationen über die Zuverlässigkeit und die Funktionsweise des Systems auch zur Verfügung stehen.

Einblendung: Icons (Grübelnde Figur, Zettel, Figuren an Flowchart, Lupe, Glühbirne)

Quellen

Quelle [1] Weiser, B., & Schweber, N. (2023, June 8). The ChatGPT lawyer explains himself. The New York Times. Verfügbar unter:

https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html (zuletzt abgerufen am 12.12.2023)

Quelle [2] Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2022). *Engineering psychology and human performance* (5th ed.). Routledge. [Chapter 13: Human-Automation Interaction, p. 516-551]. https://doi.org/10.4324/9781003177616

Quelle [3] Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., & Madai, V., I. (2021). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health, 1(*2), Article e0000016. https://doi.org/10.1371/journal.pdig.0000016

Quelle [4] Blomberg, S. N., Folke, F., Ersbøll, A. K., Christensen, H. C., Torp-Petersen, C., Sayre, M. R., Counts, C. R., & Lippert, F. K. (2019). Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation, 138*, 322-329. https://doi.org/10.1016/j.resuscitation.2019.01.015

Quelle [5] Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., & Folke, F. (2021). Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services. A randomized clinical trial. *JAMA Network Open, 4*(1), Article e2032320. https://doi.org/10.1001/jamanetworkopen.2020.32320

Quelle [6] Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., Holm, S., Kühne, U., Madai V. I., Osika, W., Spezzatti, A., Schnebel, E., Tithi, J. J., Vetter, D., Westerlund, M., … Kararigas, G. (2021). On assessing trustworthy AI in Healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics, 3*, Article 673104. https://doi.org/10.3389/fhumd.2021.673104

Disclaimer

Transkript zu dem Video „Mensch-KI-Interaktion: Fallbeispiele aus der Praxis“, Dr. Maike Mayer.