KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Mensch-KI-Interaktion: 03_03Implikation_Vertrauensfaktoren

Vertrauensfaktoren

Erarbeitet von
Dr. Maike Mayer

Lernziele

- Du kannst verschiedene Faktoren benennen, die sich auf unsere Interaktion mit Künstlicher Intelligenz auswirken können
- Du kannst exemplarisch einschätzen, wie sich die Faktoren auswirken können
- Du kannst aufzeigen, wie man mit einigen der möglichen Probleme umgehen kann

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

Inhalt

Einstieg

Genau wie zwischenmenschliche Interaktionen sind auch Interaktionen mit KI-Systemen durchaus komplex. In diesem Video schauen wir uns daher einige Faktoren näher an, die neben der **Zuverlässigkeit** des Systems, dessen **Komplexität** oder der **Fehlerart** unser Vertrauen in und unsere Interaktion mit KI-Systemen beeinflussen können. Dabei werfen wir im Folgenden Schlaglichter auf einige Faktoren, um dich dafür zu sensibilisieren, dass viele Faktoren bei der Interaktion mit technischen Systemen eine Rolle spielen können. Der Versuch eines vollständigen Überblicks würde ein bisschen den Rahmen sprengen.

Einblendung: Icons (Figuren mit Dosentelefon, Figur am Computer, Taschenlampe, Glühbirne); Schlagworte („Zuverlässigkeit“, „Komplexität“, „Fehlerart“)

Und nachdem wir jetzt so transparent über dieses Video und die Zielsetzung gesprochen haben, starten wir direkt mit einem Faktor, der ganz exzellent dazu passt …

Einblendung: Icon (Lupe)

Transparenz

Transparenz ist für unsere Interaktion mit KI-Systemen ein wichtiger Faktor. Ist ein System für uns sehr intransparent gestaltet, wirkt sich das negativ auf unser Vertrauen aus [1]. Wir können nicht nachvollziehen, was passiert oder verstehen nicht, warum ein bestimmtes Ergebnis zustande gekommen ist. Vielleicht bekommen wir von dem System darüber keine Rückmeldung oder erhalten nur für uns unverständliche Hinweise. Das führt dann dazu, dass wir ein System lieber nicht oder nur wenig nutzen. Oder, dass es ein neues System gar nicht oder nur sehr schlecht aus der Forschung in die Praxis schafft [2].

Quelle [1] [2]

Einblendung: Icons (Code mit Lupe, Figur mit Fragezeichen, Laptop, Sprechblase mit Kreuz, Sprechblase mit Fragezeichen); Schlagworte („Transparenz“)

Transparenz zur angemessenen Einschätzung des Systems lässt sich dabei ganz unterschiedlich während unserer Interaktion mit dem System umsetzen [1]. Beispielsweise über graphische Illustrationen, textliche Beschreibungen, oder Ausführungen darüber, wie und welche Fehler möglich sind. Auch Schulungen, die im Vorfeld der Interaktion erläutern, wie das System funktioniert, hängen in gewisser Weise mit dem Aspekt der Transparenz zusammen und können sich positiv auf unser Vertrauen in das jeweilige System auswirken [1].

Quelle [1]

Einblendung: Icons (Figur mit Glühbirne, Figur über Plakat, beschriebenes Papier, Figuren über Flowchart)

Das Problem bei KI-Systemen ist allerdings, dass sie oft sehr komplex sind und eher als Black Box funktionieren [2]. Es ist also nicht klar, wie das System von den Eingaben zu seinem Ergebnis kommt [3]. Hier setzt das Konzept der „Explainable AI"– also der erklärbaren KI [1]. Diese Systeme geben beispielsweise Begründungen für ihre Empfehlungen aus. Man könnte „Explainable AI" also als eine Sammlung von Methoden beschreiben, die es ermöglichen, die Ausgaben (oder Ergebnisse) eines KI-Systems nachzubilden, um dann so ein mögliches Verständnis darüber zu erlangen, wie das KI-System mit den Eingaben arbeitet. Die Herausforderung bei solchen Systemen besteht – sehr vereinfacht ausgedrückt – allerdings darin, dass die Erklärungen auch zu den Prozessen und den Ergebnissen des Systems passen.

Quelle [1] [2] [3]

Einblendung: Flowchart mit Icons („Eingabe", Pfeil, Icon Fragezeichen, Pfeil, „Ergebnis", Icons Sprechblase und Glühbirne); Icons (Werkzeugkoffer, Kopf mit Glühbirne und Zahnrädern, Ausrufezeichen); Schlagwort („BlackBox", „Explainable AI")

Ok, es ist also wichtig, dass wir während unserer Interaktion mit einem KI-System transparent nachvollziehen können, was eigentlich gerade passiert und auch, wie ein Ergebnis vermutlich zustande gekommen ist. Aber der Aspekt der Transparenz geht darüber hinaus. In einem KI-System stecken jede Menge Entscheidungen, die während der Systementwicklung gefallen sind [4]. Das sind beispielsweise Entscheidungen darüber, welche Daten einbezogen werden, wie bestimmte Dinge gemessen werden oder auch welche Methode angewendet wird. Diese Entscheidungen transparent zu machen, ist wichtig, um Systeme verbessern und sie für die passenden Aufgaben einsetzen zu können. Es ist aber vor allem auch wichtig, wenn uns eine KI-gestützte Entscheidung direkt betrifft. Wenn wir mit einer solchen – für uns womöglich sogar negativen – Entscheidung konfrontiert werden, aber weder eine Begründung dafür bekommen, noch Widerspruch einlegen können, kann das zu heftigen Reaktionen führen [4].

Quelle [4]

Einblendung: Icons (Code mit Lupe, Netzwerk, Ausrufezeichen, Glühbirne, ängstliches Gesicht, wütendes Gesicht); Schlagworte („Daten", „Messung", „Methode", „…")

Einsatzbereich

Für unsere Wahrnehmung und Nutzung von Maschinen oder KI-Systemen kann relevant sein, wo oder wofür diese Systeme eingesetzt werden. Stell dir mal vor, du bist seit längerem erkrankt und möchtest endlich eine Diagnose bekommen, um zu wissen, was los ist und was man tun kann. Würdest du lieber eine Diagnose von einem menschlichen Arzt bzw. einer menschlichen Ärztin oder von einem KI-System bekommen? Und welcher Diagnose würdest du eher vertrauen? Na? Laut einer Studie [5] neigen wir dazu, Diagnosen

eines Menschen eher zu vertrauen als den Diagnosen eines KI-Systems. Wir neigen auch eher dazu, den von einem Menschen vorgeschlagenen Behandlungen zu folgen. Und das, obwohl sich Mensch und Maschine für unser Empfinden gleich sicher mit ihrer Diagnose sein müssen, damit wir die Diagnose akzeptieren und den Behandlungsvorschlägen folgen.

Quelle [5]

Einblendung: Icons (Krankenbett, Zettel, Pfeil, Arzt/Ärztin mit Patient/Patientin, Pfeil, Figur vor Computer, Gleichzeichen)

Außerdem fühlen wir uns in Bereichen bzw. bei Entscheidungen, die moralische Aspekte umfassen, mit Maschinen als Entscheidungsinstanzen eher unwohl [6-7]. Hier geht es vor allem um automatisierte Entscheidungen. Sobald Entscheidungen moralische Komponenten haben, bevorzugen wir, dass ein Mensch die Entscheidung trifft [6]. Das scheint unter anderem damit zusammenzuhängen, dass wir Maschinen als unfähig zu denken und zu fühlen wahrnehmen.

Quelle [6] [7]

Einblendung: Icons (Laptop, Daumen runter, Figur, Daumen hoch); Schlagwort („moralische Aspekte”)

Persönliche Eigenschaften

Neben dem KI-System sind wir als Menschen auch Teil der Mensch-KI-Interaktion. Und auch wir bringen Merkmale und Eigenschaften mit, die unsere Interaktion mit KI-Systemen beeinflussen können. Beispielsweise unterscheiden wir Menschen uns in unserer persönlichen Tendenz, Automation, wie zum Beispiel KI-Systemen, zu vertrauen [8]. Diese Tendenz ist unabhängig von dem Kontext, in dem wir uns befinden, oder von dem jeweiligen System, mit dem wir interagieren. Es ist eine stabile persönliche Eigenschaft. Manche Studien weisen sogar darauf hin, dass Menschen mit hoher Bereitschaft, anderen zu vertrauen, auch eher dazu neigen, verlässlichen Systemen zu vertrauen, dafür aber stärker auf Fehler der Systeme reagieren.

Quelle [8]

Einblendung: Icons (Gruppe von Menschen, Lupe mit Gesicht, Kopf mit Gehirn, aufzeigendes Männchen, Laptop mit Daumen hoch, wütendes Gesicht mit Kreuz)

Auch andere persönlichen Eigenschaften könnten einen Einfluss auf unser Vertrauen in Automation haben. Neigen wir beispielsweise eher zu „alles-oder-nichts”-Denkweisen, verringert sich unser Vertrauen nach einem Fehler des Systems stärker als bei Personen, die weniger zu solchen Denkweisen neigen [1]. Außerdem können unser Verständnis, wie die Automation funktioniert, unsere Erwartungen an das System und unsere wahrgenommene Fähigkeit, das System zu benutzen, ebenfalls unsere Interaktion mit dem System beeinflussen [9]. Neben diesen kognitiven Faktoren spielen aber auch emotionale Aspekte

eine Rolle, wie unsere Einstellung gegenüber Automation oder wie wohl wir uns mit Automation fühlen.

Quelle [1] [9]

Einblendung: Icons (nachdenkendes Männchen, Gedankenblase, Häkchen und Kreuz); Schlagworte („Verständnis", „Erwartung", „wahrgenommene Fähigkeit", „emotionale Aspekte")

Eigenschaften der Einsatzsituation

Nachdem wir jetzt sowohl über maschinelle als auch über menschliche Eigenschaften gesprochen haben, fehlt noch ein dritter, großer Bereich: Nämlich die Situation, in der wir mit einem System interagieren. Auch situative Eigenschaften können sich auf unser Vertrauen bzw. unsere Interaktion mit KI-Systemen auswirken [8]. Beispielsweise unsere aktuelle Arbeitsbelastung. Sie beeinflusst, wie viel Zeit wir auf das Überwachen eines Systems verwenden und kann auch unser Vertrauen beeinflussen.

Quelle [8]

Einblendung: Icons (Laptop, Gruppe Menschen, volle Arbeitsmappe, Uhr); Schlagwort („situative Eigenschaften")

Und wo wir gerade schon bei einer Arbeitssituation sind: Die Arbeitskultur und auch die Aufgabe, um die es geht, sind potentiell relevant. So können die Einstellungen unserer Kolleginnen und Kollegen oder unserer Vorgesetzten zu einem bestimmten System auch unsere eigene Einstellung beeinflussen. Dabei spielt unter Umständen auch der Ruf der einzusetzenden Technologie eine Rolle oder gängige Erwartungshaltungen. Aber auch unsere eigenen Erfahrungen mit dem System, unsere individuelle Motivation, unser Stresslevel oder möglicherweise Langeweile wirken sich unter Umständen auf unsere Interaktion mit KI-Systemen aus. Sind wir uns darüber hinaus sicher, dass wir eine bestimmte Aufgabe sehr gut selbst bewältigen können, sind wir auch weniger geneigt, uns auf ein System zu verlassen.

Einblendung: Icons (Welt mit Figürchen drum herum, Figur mit Flowchartabbildung in der Hand, sprechende Menschen, denkende Figur, Gedankenblase mit Laptop, muskulöse Figur, Gruppe an Arbeitsblatt); Schlagworte („Erfahrungen, „Motivation", „Stresslevel", „Langeweile")

Fazit

Unsere Interaktion mit KI-Systemen wird von einer Vielzahl an Faktoren beeinflusst, wie den Eigenschaften der Maschine, Eigenschaften von uns als Nutzende eines Systems und Eigenschaften der Einsatzsituation. Sollen KI-Systeme in der Praxis eingesetzt werden, lohnt sich daher ein umfassender Blick auf die zu gestaltende Mensch-KI-Interaktion. So lassen sich viele Probleme bereits im Vorfeld identifizieren und negative Folgen vermeiden oder

zumindest abfedern. Manche Probleme treten allerdings erst im praktischen Betrieb auf. Aber auch hier lohnt sich ein möglichst breiter Blick, um potentielle Problemquellen zu identifizieren und beheben zu können.

Einblendungen: Icons (Laptop, Gruppe Menschen, Welt mit Figürchen drum herum, Lupe, Figur mit Glühbirne)

Quellen

Quelle [1] 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 [2] Deutsches Forschungszentrum für Künstliche Intelligenz (2023). *DFKI News, 51* (1/2023). https://uk.dfki.de/DFKI_NEWS_ePaper/epaper-DFKI_News_51_d/#1

Quelle [3] High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission. https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf (zuletzt abgerufen am 08.12.2023)

Quelle [4] Zweig, K. (2023). *Die KI war's! Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz.* Heyne.

Quelle [5] Juravle, G., Boudouraki, A., Terziyska, M., & Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. In B. L. Parkin (Ed.), *Progress in Brain Research* (Vol. 253, p. 263-282). Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.006

Quelle [6] Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Quelle [7] Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Quelle [8] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Quelle [9] Schaefer, K. E., Chen, J. Y., Szalma, J. L., &, Hancock, P. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400. https://doi.org/10.1177/0018720816634228

Disclaimer

Transkript zu dem Video „Mensch-KI-Interaktion: Vertrauensfaktoren“, Dr. Maike Mayer.