KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Mensch-KI-Interaktion: 03_02Begriffe_Systemvertrauen

Systemvertrauen

Erarbeitet von
Dr. Maike Mayer

Lernziele

- Du kannst wichtige Begriffe („Vertrauen in ein System“, „Complacency“, „Automation Bias“, „Cry Wolf“) erklären
- Du kannst erläutern, welche Probleme bei der Interaktion mit KI auftreten (können)
- Du kannst Einflussfaktoren auf unser Vertrauen in ein KI-System benennen

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

Inhalt

Einstieg

Stell dir vor, du arbeitest in einer Personalabteilung. Auf die letzte Stellenausschreibung sind über 180 Bewerbungen eingegangen. Deine Aufgabe ist es jetzt, diese Bewerbungen durchzusehen und zu entscheiden, wer zu einem Gespräch eingeladen wird und wer nicht. Puh – das wird eine ganze Weile dauern. Es sei denn, ein KI-System hilft dir dabei und sortiert die Bewerbungen schonmal vor, in – sagen wir – vielversprechend, mittel und eher nicht. Das dürfte eine Menge Arbeit sparen. Aber es ist auch eine verantwortungsvolle Aufgabe. Was nun? Nehmen wir an, es gibt keine Vorschriften deiner Vorgesetzten. Wonach entscheidest du, ob du dem System die Aufgabe geben möchtest?

Einblendung: Icons (Papierstapel, Daumen hoch, Daumen runter, Person, die erschreckt auf Uhr schaut, Glühbirne, lachender Smiley, neutraler Smiley, trauriger Smiley, Uhr, denkende Person); Schlagwort („> 180")

Wenn du die Aufgabe an eine Kollegin oder einen Kollegen weitergeben müsstest, statt an ein KI-gestütztes System, hättest du vielleicht gesagt: Naja, wenn ich dem Kollegen bzw. der Kollegin vertraue, dann gebe ich die Aufgabe weiter. Gilt das auch für ein KI-System?

Einblendung: Icons (denkende Person, Personen, Handschlag, Fragezeichen)

Eine Vertrauenssache?

Ja, tatsächlich spielt Vertrauen auch bei unserer Interaktion mit Maschinen oder eben KI-Systemen eine Rolle! Und zwar eine ziemlich wichtige [1]. Aber was versteht man eigentlich unter dem Vertrauen in ein System?

Quelle [1]

Einblendung: Icons (Person mit Glühbirne, Fragezeichen)

Wissenschaftlerinnen und Wissenschaftler haben sich noch nicht auf *die* Definition von Vertrauen in diesem Kontext geeinigt [2], aber recht häufig begegnet einem die folgende Definition [3]: Vertrauen ist die Einstellung bzw. Erwartung eines Vertrauenden, dass ein Agent dabei hilft, die Ziele des Vertrauenden in einer von Unsicherheit und Verletzlichkeit bestimmten Situation zu erreichen. Das klingt etwas abstrakt, oder? Daher nochmal etwas einfacher gesagt. Wenn wir einem KI-System vertrauen, dann gehen wir davon aus, dass dieses System uns dabei hilft, unser Ziel zu erreichen – also beispielsweise die Bewerbungen angemessen zu sortieren. Und dieses Vertrauen entsteht in einer Situation, in der wir eben nicht wissen, ob das System das ordentlich macht. Wir müssen uns bei Einsatz des Systems ein Stück weit darauf verlassen. Wichtig ist hierbei noch, dass unser Vertrauen in ein System und die tatsächliche Nutzung eines Systems zwei verschiedene Konzepte sind, die aber oft

miteinander zusammenhängen [1]. Je mehr ich einem System vertraue, desto mehr nutze ich es auch. Allerdings kann es auch zu Situationen kommen, in dem ich einem System zwar vertraue, es aber nicht einsetze oder es einsetze, obwohl ich ihm nicht vertraue. Das kann beispielsweise an betrieblichen Vorgaben liegen oder auch an persönlichen Präferenzen. Beispielsweise könnte die Nutzung eines Systems verpflichtend vorgeschrieben sein oder ich mache eine bestimmte Aufgabe einfach sehr gerne selbst.

Quelle [2] [3] [1]

Einblendung: Icons (Bücherei, Glühbirne, Laptop, Pfeil, Zielflagge, 3 Papierstapel mit je einem fröhlichen, einem neutralen und einem traurigen Smiley, Pfeile, Vertrag, schreibende Person vor dem Computer); Schlagworte (komplette Definition, „Vertrauen“, „Nutzung“, „≠“)

Aber zurück zu dem Konzept „Vertrauen”. Idealerweise vertrauen wir einem System in dem Ausmaß, in dem es eine Aufgabe auch erfüllen kann. Einer Maschine, die perfekt funktioniert, können wir auch komplett vertrauen. Einer Maschine, die eine Aufgabe überhaupt nicht kann, der vertrauen wir besser gar nicht. Unser Vertrauen sollte sich also im Idealfall an der Zuverlässigkeit – oder auch Reliabilität – der Maschine orientieren. Das ähnelt vielleicht auch unserem Vorgehen bei Kolleginnen und Kollegen. Wenn wir wissen, dass sie die Aufgabe gut machen werden – also eine hohe Zuverlässigkeit haben – vertrauen wir ihnen eher und geben die Aufgabe eher ab. Die Zuverlässigkeit des Systems definiert sich dabei als der Anteil der richtig gelösten Aufgaben an allen durchgeführten Aufgaben [1].

Quelle [1]

Einblendung: Icons (Daumen hoch, Daumen runter, Handschlag); Schlagwort („Zuverlässigkeit”); Formel ((Aufgaben - Fehler)/Aufgaben)

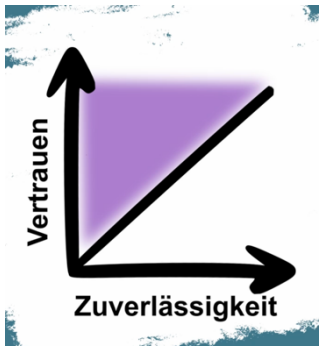
Schauen wir uns den Zusammenhang zwischen unserem Vertrauen in ein System und der Zuverlässigkeit des Systems mal graphisch an [1]. Auf der x-Achse tragen wir die Zuverlässigkeit des Systems ab und auf der y-Achse unser Vertrauen in das System. Wäre unser Vertrauen perfekt an die Zuverlässigkeit des Systems angepasst, würden wir dem System genau in dem Umfang vertrauen, in dem es verlässlich ist. Dieser Zusammenhang lässt sich durch eine Gerade visualisieren. Leider gelingt es uns nicht immer, uns perfekt an die Zuverlässigkeit des Systems anzupassen. Tatsächlich schwanken wir, basierend auf unseren Erfahrungen mit dem System, um diese Gerade herum und vertrauen dem System mal mehr und mal weniger, als wir idealerweise sollten [1]. Und das kann potentiell problematisch werden.

Quelle [1]

Einblendung: Icons (Glühbirne); Grafik (erst x-Achse, dann y-Achse [inkl. Beschriftung], Gerade, exemplarische Punkte im Koordinatensystem); Schlagworte („mehr“, „weniger“)

Ein Problem: Zu viel Vertrauen

Wenn wir einem System mehr vertrauen, als wir eigentlich sollten, liegt das Problem auf der Hand. Wir verlassen uns übermäßig darauf und das kann zu Fehlern führen. In unserer Grafik entspricht übermäßiges Vertrauen diesem Bereich hier. Vor allem, wenn ein System sehr gut und zuverlässig arbeitet, aber eben noch nicht perfekt ist, neigen wir dazu, das System nicht genauer zu überwachen [1, 4]. Das bezeichnet man auch als „**Complaceny**”. Vor allem bei Systemen, die uns bei Entscheidungen unterstützen, spricht man in diesem Zusammenhang auch oft von **„Automation Bias**” [1].



Quelle [1] [4]

Einblendung: Icons (Figur mit Idee); Grafik (Bereich oberhalb der Geraden markieren); Schlagworte („Complacency”, „AutomationBias”)

Bei einem entscheidungsunterstützenden KI-System könnte sich das beispielsweise darin äußern, dass wir davon ausgehen, dass die Empfehlungen korrekt sind. Wir prüfen also beispielsweise die von der KI ausgegebene Diagnose nicht mehr ordentlich und übernehmen sie einfach. Das führt in der Folge auch dazu, dass Fehler des Systems ggf. schwerer entdeckt werden können. Wir prüfen die Empfehlungen ja nicht! Die fehlerhaften Empfehlungen des Systems fallen so unter Umständen erst später auf, beispielsweise bei weiteren Behandlungsschritten oder durch externe Rückmeldungen.

Einblendung: Icons (Daumen hoch, durchgestrichene Lupe, pfeifendes Männchen, erschrecktes Gesicht, Ausrufezeichen)

Aber nicht nur übermäßiges Vertrauen in ein System ist potentiell problematisch …

Auch ein Problem: Zu wenig Vertrauen

Vertrauen wir einem System zu wenig, bedeutet das, dass wir ihm weniger vertrauen als es bei der Leistung des Systems angemessen wäre. In der Grafik entspräche das diesem Bereich. Das System könnte uns helfen, denn es ist besser als wir glauben, aber wir vertrauen ihm nicht.

Einblendung: Grafik (Bereich unterhalb der Geraden markieren)

Ein Phänomen in diesem Bereich ist der sogenannte **„Cry Wolf"-Effekt** [1]. Kennt ihr vielleicht das Sprichwort „Wer einmal lügt, dem glaubt man nicht"? Oder die Geschichte von dem Hirtenjungen, der zu oft „Wolf" rief? Wie in der Geschichte, in der die Dorfbewohnerinnen und Dorfbewohner dem Hirtenjungen nicht mehr glauben, wenn er „Wolf!" ruft, auch wenn der Wolf dann tatsächlich kommt, neigen wir dazu, ein System zu ignorieren, wenn es oft warnt, ohne dass in der Realität tatsächlich ein Problem vorliegt. Schließlich warnt es ständig, aber nie ist etwas. Das kann dann dazu führen, dass wir tatsächlich korrekte Warnungen oder richtige Empfehlungen verpassen bzw. aus Prinzip ablehnen.

Quelle [1]

Einblendung: Icons (Wolf, wütende Gesichter, Junge, durchgestrichener Computer, wütendes Gesicht)

Wie bisher schon ein bisschen mitgeschwungen ist: Unser Vertrauen hängt unter anderem auch von unserer Erfahrung mit dem jeweiligen System ab [1]. Wenn wir in der Vergangenheit festgestellt haben, dass das System zwar oft warnt oder empfiehlt, aber häufig daneben liegt, vertrauen wir ihm weniger. Es gibt aber noch weitere Faktoren, die unser Vertrauen in ein System beeinflussen können.

Quelle [1]

Einblendung: Icons (Figur mit Idee, Glühbirne); Schlagwort („Erfahrung")

Was beeinflusst unser Vertrauen?

Über zwei Einflussfaktoren haben wir bereits gesprochen: Nämlich die Zuverlässigkeit des Systems und unsere Erfahrungen mit dem System. Bei der **Zuverlässigkeit** handelt es sich um einen sehr maßgeblichen Einflussfaktor [1], der teilweise auch als der wichtigste Einflussfaktor bezeichnet wird. Je zuverlässiger ein System ist, desto mehr vertrauen wir ihm. Hier spielt dann aber auch unsere **Erfahrungen** mit dem System eine Rolle, denn wir müssen ja erst ein Gefühl für die tatsächlilche Zuverlässigkeit des Systems bekommen. Wenn das System dann lange fehlerfrei arbeitet, neigen wir irgendwann dazu, uns darauf zu verlassen. Hier könnte es dann zu „Complacency" bzw. zum „Automation Bias" kommen. Passiert dann der erste Fehler und wir bemerken diesen, korrigieren wir unser Vertrauen in das System – allerdings oft zu stark, sodass es dann zu *zu* wenig Vertrauen kommt. Dass wir auf den ersten Fehler oft sehr heftig reagieren wird auch als **„First Failure" Effekt** bezeichnet [1]. Aber nach und nach bekommen wir dann ein Gefühl für die tatsächliche Zuverlässigkeit des Systems. Neben dem ersten Fehler spielt übrigens auch die **Art der Fehler** eine Rolle. Interessanterweise beeinträchtigen Fehler, die für uns offensichtlich sind, unser Vertrauen stärker als Fehler, die uns auch hätten passieren können.

Quelle [1]

Einblendung: Icons (Ausrufezeichen, pfeifendes Männchen, Wolf, Glühbirne); Schlagworte („Zuverlässigkeit", „Erfahrungen", „First Failure" Effekt, „Fehlerart")

Ein weiterer Einflussfaktor ist die **Komplexität** eines Systems oder eines Algorithmus [1]. Je komplizierter und undurchdringlicher so eine Maschine uns erscheint, desto geringer ist unser Vertrauen in sie. Wir sind skeptisch. Dieser Punkt hängt auch mit **Transparenz** zusammen [1]. Ist für uns nicht ersichtlich, wie das System funktioniert, oder bekommen wir zu wenig oder gar kein Feedback über die Prozesse, die ablaufen, wirkt sich das ebenfalls negativ auf unser Vertrauen aus. Sowohl die Komplexität als auch die Transparenz ist eine Herausforderung für KI-Systeme, die oft auf komplizierten neuronalen Netzen oder Algorithmen beruhen. Teilweise ist für uns gar nicht genau nachvollziehbar, wie sie konkret funktionieren. Ein Stichwort in diesem Zusammenhang ist „**Black Box KI**". Bei einer „Black Box KI" kennt man nur den Input und das Ergebnis. Wie das System auf dieses Ergebnis gekommen ist, bleibt für uns unklar [5].

Quelle [1] [5]

Einblendung: Icons (Netzwerk, Code mit Lupe, Fragezeichen, Pfeil); Schlagworte („Komplexität", „Transparenz", „Black Box KI", „Input", „Ergebnis")

Fazit

Fassen wir noch einmal zusammen. Für unsere Interaktion mit KI-gestützten Systemen spielt unser Vertrauen in sie eine wichtige Rolle. Systeme, denen wir vertrauen, nutzen wir in der Regel auch stärker als Systeme, denen wir nur wenig oder gar nicht vertrauen. Dabei handelt es sich bei dem Vertrauen in ein System um eine persönliche Einstellung oder Erwartung.

Einblendung: Icon (Kopf mit Hirn); Schlagwort („Vertrauen")

Obwohl wir uns idealerweise bei unserem Vertrauensausmaß an der Zuverlässigkeit des Systems orientieren sollten, vertrauen wir Systemen teilweise zu viel und übernehmen Empfehlungen oder Diagnosen des Systems unkritisch. Stichwort „Complacency" und „Automation Bias". Wir können aber auch zu wenig Vertrauen in ein System haben und dadurch wichtige Empfehlungen ignorieren. Neben der Zuverlässigkeit des jeweiligen Systems beeinflussen noch weitere Faktoren unser Vertrauen wie beispielsweise unsere Erfahrungen mit dem System, die Komplexität des Systems oder die Art der Fehler, die das System macht.

Einblendung: Icons (Glühbirne); Schlagworte („Complacency", „Automation Bias", „Zuverlässigkeit", „Erfahrungen", „Art der Fehler")

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] 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 [3] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Quelle [4] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30(3), 286-297. https://doi.org/10.1109/3468.844354

Quelle [5] 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)

Weiterführendes Material

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Madhavan, P. & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. https://doi.org/10.1080/14639220500337708

Disclaimer