Themenblock Mensch-KI-Interaktion: 03_05Transfer_VertrauenswürdigeKI

Vertrauenswürdige KI

Erarbeitet von
Dr. Maike Mayer

Lernziele 1
Inhalt 2
Einstieg 2
Was versteht man unter „vertrauenswürdiger KI"? 2
Ethische Grundsätze für vertrauenswürdige KI 3
Anforderungen an vertrauenswürdige KI 5
Diskussion & Fazit 5
Quellen 7
Disclaimer 7

Lernziele

- Du kannst vertrauenswürdige KI gemäß der Hochrangigen Expertengruppe für Künstliche Intelligenz der Europäischen Union definieren
- Du kannst Verbindungen und Parallelen zu bisherigen Inhalten des Themenblocks aufzeigen
- Du kannst einordnen, inwiefern Vertrauen bzw. Vertrauenswürdigkeit nicht nur bei der eigenen Nutzung von KI-Systemen eine Rolle spielt

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

Inhalt

Einstieg

Stell dir vor, du wartest darauf, dass du etwas von deinen Steuern zurückbekommst. Schließlich erhältst du die Rückmeldung, dass du nichts zurückbekommst und sogar noch etwas nachzahlen musst. Nehmen wir weiter an, an dieser Entscheidung war eine KI beteiligt. Was würdest du dir von dieser KI wünschen, damit du die Entscheidung akzeptierst und dem Prozess vertraust? Was müsste das KI-System ganz grundsätzlich mitbringen, um aus deiner Sicht als vertrauenswürdig zu gelten? Und was bedeutet das überhaupt – vertrauenswürdige KI?

Einblendung: Icons (Sparschwein, Sparschwein durchstreichen, entsetztes Gesicht, Kartenlesegerät, Fragezeichen, nachdenkende Figur); Schlagwort („vertrauenswürdig”)

Was versteht man unter „vertrauenswürdiger KI”?

Unter dem Begriff **vertrauenswürdige KI** oder auch „trustworthy AI” findet man eine ganze Menge Treffer. Falls du mal danach suchst, wirst du feststellen, dass es viele Richtlinien und Leitfäden gibt, die sich mit diesem Thema beschäftigen [1-2].

Quelle [1-2]

Einblendung: Icons (Mappe mit Papieren, Papierstapel); Schlagwort („vertrauenswürdige KI“)

Exemplarisch stelle ich daher vor, was die Hochrangige Expertengruppe für Künstliche Intelligenz der Europäischen Union unter vertrauenswürdiger KI versteht [3]. Du kannst ja im Rückgriff auf das Eingangsbeispiel mal überlegen, ob dir diese Kriterien sinnvoll erscheinen. Und ob du diese bei deiner Einschätzung, ob du der Entscheidung vertraust, die dir mitgeteilt wurde, einbeziehen würdest.

Quelle [3]

Einblendung: Icon (denkende Figur mit Hand am Kinn); Cover der Leitlinien vor weißem Hintergrund

Laut der Expertengruppe ist Vertrauenswürdigkeit eine Art Grundvoraussetzung dafür, dass wir KI-Systeme überhaupt entwickeln und nutzen. Zumindest für die Nutzung sollte dir das plausibel erscheinen, denn unser Vertrauen in ein System beeinflusst, ob und wie wir ein System nutzen [4]. Der Expertengruppe geht es aber nicht nur um die Bestandteile des KI-Systems allein, sondern auch um das System im Kontext – also quasi das sozio-technische Gesamtkonzept. Und das Vertrauen entweder schaffen oder auch zerstören kann.

Quelle [4]

Einblendung: Icons (Daumen hoch, Figur mit Idee, Welt mit Figuren drum herum), Schlagworte („sozio-technisches System“)

Vertrauenswürdige KI definiert die Expertengruppe dabei als **rechtmäßig, ethisch und robust**. Sie soll alle anwendbaren Gesetze und Bestimmungen sowie ethische Grundsätze und Werte einhalten und technisch und sozial robust sein. Diese Robustheit bezieht sich beispielsweise darauf, dass die Systeme zuverlässig arbeiten, vor Angriffen geschützt sind und im Vorfeld Sicherheitsmaßnahmen getroffen wurden, um unerwünschte negative Auswirkungen des KI-Einsatzes zu verhindern. Gerade der letzte Punkt ist naheliegend, wenn man sich die Skalierbarkeit von KI-Entscheidungen bzw. KI-Unterstützung bei Entscheidungen vor Augen führt. In den USA gab es beispielsweise in Michigan den Fall, dass mehreren tausend Menschen fälschlicherweise vorgeworfen wurde, sie hätten bei der Arbeitslosenunterstützung betrogen [5-6], anderen wurde die Leistung „nur” fälschlicherweise vorenthalten. Das macht deutlich, wie weitreichend die Konsequenzen eines KI-Einsatzes sein können – positiv wie negativ.

Quelle [5-6]

Einblendung: Icons (Nummern von 1 bis 3, Waagschale, Computercode mit Lupe, Gruppe Menschen, Checkliste, Ritter, Vorhängeschloss, Glühbirne, entsetztes Gesicht, durchgestrichenes Geld); Schlagworte („rechtmäßig”, „ethisch”, „robust”)

Aber zurück zu der Expertengruppe …

Ethische Grundsätze für vertrauenswürdige KI

Zusätzlich zu ihrer Definition von vertrauenswürdiger KI hat die Expertengruppe auch noch vier ethische Grundsätze für vertrauenswürdige KI-Systeme erarbeitet. Und zwar

1. **Achtung der menschlichen Autonomie:** Bei diesem Punkt geht es unter anderem darum, dass Menschen in der Lage sein müssen, ihre Selbstbestimmung auszuüben. KI-Systeme sollten Menschen beispielsweise nicht täuschen, manipulieren oder nötigen. In diesem Kontext taucht aber auch die Idee auf, dass die Zuweisung von Funktionen zwischen Mensch und KI-Systemen nach menschenzentrierten Grundsätzen erfolgen sollte. Erhalte ich eine KI-gestützte Entscheidung, soll in diesem Entscheidungsprozess eine menschliche Aufsicht und Kontrolle der KI-Arbeitsprozesse erfolgt sein. Die Menschen, die an der Entscheidung beteiligt sind, sollen zudem sinnvolle Entscheidungsspielräume haben. Es geht also nicht darum, dass die Entscheidung, die ich bekomme, einfach blind abgenickt wird oder werden musste.
2. **Schadensverhütung:** KI-Systeme sollten bestehende Situationen nicht verschlimmern. Sie sollten also eigentlich keine Schäden verursachen oder diese noch verschärfen. Idealerweise sollten sie sich sogar gar nicht negativ auf den Menschen auswirken.
3. **Erklärbarkeit:** Hierunter fällt beispielsweise, dass man Entscheidungen den direkt oder indirekt betroffenen Personen in größtmöglichem Umfang erklären kann. Die

Prozesse müssen transparent gestaltet werden und auch die Fähigkeiten und der Zweck des KI-Systems soll offen kommuniziert werden. Das ähnelt auch den Aspekten der Transparenz im Kontext von Vertrauen in KI-Systeme und unserer Interaktion mit ihnen.

4. **Fairness:** Dabei handelt es sich um einen recht vielfältigen Begriff, der unterschiedlich ausgelegt werden kann. Die Expertengruppe verweist hier insbesondere auf die Verpflichtung, eine gleiche und gerechte Verteilung von Vorteilen und Kosten zu gewährleisten, sowie Personen und Gruppen beispielsweise vor Diskriminierung und Stigmatisierung zu schützen. In diesen Bereich fällt aber auch, dass sich Betroffene gegen KI-gestützte Entscheidungen wehren können.

Einblendung: Icons (1. Durchgestrichener Teufel, Figur mit Smartphone in der Hand, nachdenkende Figur; 2. durchgestrichener Daumen runter, durchgestrichenes ängstliches Gesicht; 3. Figur mit Idee, Code mit Lupe; 4. Waage, Gruppe); Schlagworte („Achtung der menschlichen Autonomie", „Schadensverhütung", „Erklärbarkeit", „Fairness")

Da Fairness in Bezug auf KI ein interessantes, und wie gesagt, recht vielseitiges Konzept ist, schauen wir uns noch eine andere Dimension von Fairness an. Fairness kann beispielsweise auch bezeichnen, wie eine einzelne Person die Behandlung durch andere Individuen oder Institutionen wahrnimmt und einschätzt [7]. Diese Wahrnehmung kann als ein wichtiger Aspekt betrachtet werden, wenn KI beispielsweise in der Politik oder der Verwaltung eingesetzt werden soll. Vor allem, wenn es um die Legitimität des Entscheidungsprozesses und seiner Ergebnisse geht [7]. Entscheidungen können beispielsweise als unfair wahrgenommen werden, wenn das Ergebnis nicht den eigenen Vorstellungen von Fairness oder einer fairen Verteilung entspricht. Die Black-Box-Problematik verschärft das Problem möglicherweise, weil Betroffene nicht nachvollziehen können, wie eine sie betreffende Entscheidung zu Stande gekommen ist.

Quelle [7]

Einblendung: Icons (Glühbirne, Brandenburger Tor, Papierstapel, wütendes Gesicht, Figur mit Fragezeichen), Schlagworte („Fairness", „=", „Wahrnehmung der Behandlung durch andere")

In diesem Zusammenhang wird dann auch wieder – wie von der Expertengruppe ausgewiesen – wichtig, Entscheidungen adäquat erklären zu können. Erinnere dich doch nochmal an das Beispiel aus den USA zurück: Die Personen, die fälschlicherweise des Betrugs bezichtigt wurden, hätten sicherlich gerne eine Erklärung für diesen Vorwurf erhalten, um auch nachvollziehen zu können, wo diese Einschätzung hergekommen ist. In diesem Fall ging es um Arbeitslosenunterstützung, also quasi die Verteilung von öffentlichen Gütern bzw. Leistungen, aber auch bei anderen gesellschaftlich folgenreichen Entscheidungen wie zum Beispiel der Genehmigung von Asylanträgen wäre eine Unterstützung durch Entscheidungssysteme denkbar [7]. Hält man sich dies vor Augen, wird deutlich, dass es sinnvoll ist, über eine angemessene Gestaltung von KI-Systemen und ihre Einbindung in sozio-technische Kontexte nachzudenken – und das nicht nur aus der

Perspektive derjenigen, die mit den Systemen arbeiten, sondern auch aus der Sicht von denjenigen, die mit KI-unterstützen Entscheidungen konfrontiert werden.

Quelle [7]

Einblendung: Icon (Figur mit Idee, erschrecktes Gesicht, denkende Figur, Ausrufezeichen); Schlagworte („gesellschaftlich folgenreiche Entscheidungen")

Anforderungen an vertrauenswürdige KI

Die Expertengruppe hat auch Anforderungen formuliert, die bei der Entwicklung, Einführung und Nutzung von vertrauenswürdigen KI-Systemen erfüllt werden sollen. Da einige der Anforderungen auf den eben besprochenen ethischen Grundsätzen aufbauen, werde ich sie nur kurz benennen.

Einblendung: Icon (Glühbirne, Sanduhr)

Bei vertrauenswürdiger KI soll auf den **Vorrang menschlichen Handelns und auf menschliche Aufsicht** geachtet werden. Die Entscheidung darf nicht rein durch eine Maschine getroffen werden und unsere Grundrechte müssen dabei gewahrt werden. Hinzu kommt die bereits erwähnte **technische Robustheit und Sicherheit der Systeme** sowie der **Schutz der Privatsphäre und des Datenqualitätsmanagements.** In den letztgenannten Bereich fällt unter anderem auch, dass die Qualität der Datensätze sicherzustellen ist, auf denen das KI-System basiert, um beispielsweise Verzerrungen oder Benachteiligung zu vermeiden. Auch **Transparenz** ist eine wichtige Anforderung für vertrauenswürdige KI. Die Expertengruppe fasst hierunter auch den Aspekt der Kommunikation: Nutzende dürfen bei einem KI-System nicht das Gefühl bekommen, sie hätten es hier mit einem Menschen zu tun. Sie haben das Recht zu wissen, dass sie mit einem System interagieren. Als weitere Anforderungen führt die Expertengruppe dann noch **Vielfalt, Nichtdiskriminierung und Fairness**, **gesellschaftliches und ökologisches Wohlergehen** und **Rechenschaftspflicht** an.

Einblendung: Icons (durchgestrichener Laptop, Lupe, Gruppe, Dosentelefon); Schlagworte („Vorrang menschlichen Handelns & menschliche Aufsicht", „technische Robustheit & Sicherheit", „Schutz der Privatsphäre & des Datenqualitätsmanagements", „Transparenz", „Vielfalt, Nichtdiskriminierung & Fairness", „gesellschaftliches & ökologische Wohlergehen", „Rechenschaftspflicht")

Diskussion & Fazit

Gehen wir nun doch nochmal zurück zu unserem Eingangsbeispiel, bei dem du statt eine Rückzahlung zu erhalten sogar noch nachzahlen musstest. Keine so angenehme Entscheidung, mit der du da konfrontiert warst. Würde ein Befolgen der eben ausgeführten Richtlinien dazu führen, dass du den KI-Einsatz im Entscheidungsprozess akzeptierst und darauf vertraust, dass alles richtig gelaufen ist?

Einblendung: Icon (Münzen, erschrecktes Gesicht, Fragezeichen)

Ich weiß nicht, wie es dir geht, aber wenn ich mir vorstelle, dass alle Richtlinien eingehalten wurden – also beispielsweise, dass die KI fair ist, dass man mir die Entscheidung erklären kann, dass ich weiß, dass ein Mensch nochmal draufgeguckt hat und dass ich mich auch dagegen wehren kann – dann fühlt sich das, für mich zumindest, recht gut an. Man muss aber auch dazu sagen, dass die Richtlinien der Expertengruppe (wie viele andere Richtlinien und Leitfäden auch) erstmal nur Vorschläge sind und nicht rechtlich bindend. Außerdem sind es nicht die einzigen Richtlinien, die sich mit ethischen Aspekten des KI-Einsatzes beschäftigen. Es gibt viele andere, die den Fokus zum Teil auch etwas anders legen, und Reviews, die versuchen, einen gemeinsamen Nenner zu finden. Eine Übersicht von 2019 listet beispielsweise insgesamt 11 verschiedene Cluster ethischer Prinzipien auf, von denen fünf Prinzipien in mehr als der Hälfte der betrachteten Richtlinien genannt werden: Transparenz, Gerechtigkeit und Fairness, Nichtschädigung, Verantwortung und Privatsphäre [8].

Quelle [8]

Einblendung: Icons (denkende Figur, Checkliste, Daumen hoch, Ausrufezeichen, druchgestrichene Waage, Papiermappe, 11); Schlagworte („Transparenz", „Gerechtigkeit & Fairness", „Nichtschädigung", „Verantwortung", „Privatsphäre")

Die Vielzahl der existierenden Richtlinien macht deutlich, dass keine breite Einigkeit herrscht, wie der Einsatz von KI-Systemen ethisch gut vertretbar gestaltet werden soll und viele der angeführten ethischen Aspekte sind umstritten [9]. Außerdem sind sie vielschichtig und können zum Teil unterschiedlich interpretiert werden [9]. Teilweise sind sie auch recht subjektiv: Ob ich eine Entscheidung beispielsweise als fair empfinde, hängt auch ein Stück weit von mir persönlich ab.

Quelle [9]

Einblendung: Icon (streitende Figuren, nachdenkende Figur)

Auch wenn es keinen allgemein gültigen Konsens zu geben scheint, wie beispielsweise vertrauenswürdige KI gestaltet sein sollte, ist es doch wichtig, darüber zu sprechen und zu diskutieren. Nur so können wir entscheiden, was für uns akzeptabel ist und eingesetzt werden kann. Und nur so haben wir die Möglichkeit, die potentiellen Vorteile von KI auch wirklich nutzen zu können – und zwar in einem ethisch vertretbaren Rahmen.

Einblendung: Icon (Glühbirne, Sprechblasen, Daumen hoch)

Quellen

Quelle [1] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys, 55*(2), Artikel Nummer 39. https://doi.org/10.1145/3491209

Quelle [2] Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*, 447-464. https://doi.org/10.1007/s12525-020-00441-4

Quelle [3] Hochrangige Expertengruppe für Künstliche Intelligenz (2019). *Ethik-Leitlinien für eine vertrauenswürdige KI.* Europäische Kommission. Verfügbar unter: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (zuletzt abgerufen am 18.1.2024)

Quelle [4] 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 [5] Egan, P. (2017, July 30). Data glitch was apparent factor in false fraud charges against jobless claimants. *Detroit Free Press*. Verfügbar unter: https://eu.freep.com/story/news/local/michigan/2017/07/30/fraud-charges-unemployment-jobless-claimants/516332001/ (zuletzt abgerufen am 18.1.2024)

Quelle [6] Zweig, K. (2023). *Die KI war's! Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz.* Heyne.

Quelle [7] Marcinkowski, F., & Starke, C. (2019). Wann ist Künstliche Intelligenz (un-)fair? In J. Hofmann, N. Kersting, C. Ritzi, & W. J. Schünemann (Hrsg.), Politik in der digitalen Gesellschaft (S. 269-288). De Gruyter. https://doi.org/10.1515/9783839448649-014

Quelle [8] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Quelle [9] Stahl, B. C. (2023). Grauzonen zwischen Null und Eins. KI und Ethik. *Aus Politik und Zeitgeschichte, 73*(42), 17-22. Verfügbar unter https://www.bpb.de/shop/zeitschriften/apuz/kuenstliche-intelligenz-2023/ (zuletzt abgerufen am 9.01.2024)

Disclaimer