

W07_Ethik_Transparenz

Transparenz

Erarbeitet von

Marie von Lobenstein, M.A.

Lernziele	1
Inhalt	2
Einstieg	2
Blackbox KI	2
Definition und Erläuterung KI-Transparenz	2
Dimensionen von KI-Transparenz	3
Warum ist KI-Transparenz wichtig?	4
Take-Home Message: Was solltest du heute mitnehmen?	4
Quellen	4
Disclaimer	5

Lernziele

- Du kannst erklären, was mit Transparenz gemeint ist
- Du kannst Dimensionen von KI-Transparenz benennen
- Du kannst erläutern, wieso es wichtig ist über KI-Transparenz zu sprechen

Inhalt

Einstieg

Habt ihr euch schonmal gefragt, wie die TikTok Algorithmen euch so präzise Videos empfehlen oder wie ChatGPT zu den Antworten kommt, die es auswirft? OpenAI erläutert die Methode hinter ChatGPT auf ihrer Website. (ChatGPT, 2022) Sie erklären zum Beispiel, dass sie ChatGPT mit Hilfe von Reinforcement Learning trainiert haben.

Einblendung Website ChatGPT

Reinforcement Learning ist eine Art des maschinellen Lernens, bei dem ein Agent durch Interaktion mit einer Umgebung lernt, Entscheidungen zu treffen. Ein erstes Modell von ChatGPT wurde durch überwachte Feinabstimmung trainiert. Das bedeutet, dass menschliche KI-Trainer Gespräche lieferten, in denen sie beide Seiten spielten – den Nutzer und einen KI-Assistenten. Sie gehen dazu noch mehr ins Detail und machen transparent, wie ihr KI-System funktioniert. Diese Transparenz ist das Thema, über das ich heute mit euch sprechen möchte.

Quelle [1]

Die Expert*innengruppe für KI (HLEG) der EU-Kommission hat Transparenz, in den im April 2019 veröffentlichten ethischen Leitlinien, als eine von sieben Schlüsselanforderungen für die Verwirklichung einer „vertrauenswürdigen KI“ genannt. „Transparenz“ ist einer der fünf wichtigsten Grundsätze, die in einer großen Zahl der ethischen Leitlinien, die für KI auf globaler Ebene entwickelt wurden, betont werden.

Quelle [2]

Der zunehmende Einsatz von KI für Entscheidungsfindungen, also zum Beispiel im Finanz- oder Rechtswesen, hat eine lebhafte Debatte über den Nutzen und den potenziellen Schaden „selbstlernender“ Technologien ausgelöst. Um negative Folgen zu vermeiden und verantwortliche KI-Systeme zu schaffen, denken viele, dass wir die „Blackbox“ der KI-Entscheidungsfindung öffnen und transparenter machen müssen.

Blackbox KI

Ein „Blackbox“-KI-System ist übrigens ein Algorithmus, dessen interne Funktionsweise und Entscheidungsprozesse für Benutzer*innen oder Entwickler*innen eben nicht transparent ist. Ein Beispiel für teilweise intransparente KI-Systeme sind neuronale Netzwerke. So eine Blackbox kann es schwierig oder unmöglich machen, zu verstehen, warum die KI bestimmte Entscheidungen fällt oder Vorhersagen trifft. Blackbox-KI-Modelle werden zwar nicht in Deutschland, aber in anderen Ländern, in Bereichen wie dem Finanzwesen, dem Gesundheitswesen und der Strafjustiz eingesetzt, wo viel auf dem Spiel steht und die Erklärbarkeit sehr wichtig ist.

Definition und Erläuterung KI-Transparenz

Der Begriff „KI-Transparenz“ bezieht sich auf den Aspekt, ob die innere Funktionsweise eines KI-Systems für den Menschen offen und erklärbar ist. Dies ist ein wichtiges Konzept

im Bereich der KI, da viele KI-Systeme komplex sind und auf eine Art und Weise funktionieren, die für Menschen schwer zu verstehen ist. Außerdem funktioniert die Sortierung, die KI-Algorithmen vornehmen, anders als die Sortierung von Menschen. Dies kann ein Problem darstellen, da KI-Systeme zunehmend dazu verwendet werden, Entscheidungen von großer Tragweite zu treffen, z. B. wenn sie in der medizinischen Diagnostik eingesetzt werden.

Quelle [3]

Zum Thema Transparenz gibt es verschiedene Meinungen. Die einen sagen, Transparenz ist ein wichtiges Kriterium, wenn es darum geht, wie die breite Öffentlichkeit KI-Entscheidungen als legitim und akzeptabel wahrnimmt. Transparenz ist demnach also wichtig, um Vertrauen zu haben in die genutzten KI-Systeme, genug Vertrauen, sie womöglich auch Entscheidungen treffen zu lassen, die einen weitreichenden Einfluss haben. Andere vertreten die Meinung, dass es eine offene Frage ist, ob Transparenz das entscheidende Kriterium ist oder ob nicht eher Zuverlässigkeit und die Möglichkeit des Einspruchs Vertrauen generieren.

Dimensionen von KI-Transparenz

1. Input-Transparenz: Die Input-Transparenz von KI bezieht sich darauf, dass die Daten und Inputs, die zum Training und Betrieb eines KI-Systems verwendet werden, sichtbar und verständlich sind. Dazu gehört die Möglichkeit, zu sehen, welche Daten zum Trainieren des Systems verwendet wurden, sowie die Möglichkeit zu verstehen, wie diese Daten gesammelt, beschriftet und vorverarbeitet wurden. Die Transparenz der Eingaben ist wichtig, denn sie ermöglicht die Bewertung der Qualität und Repräsentativität der zum Training eines KI-Systems verwendeten Daten sowie die Identifizierung und Korrektur von Fehlern oder Verzerrungen in den Daten.

2. Output-Transparenz: Die Output-Transparenz von KI bezieht sich darauf, dass die Denk- und Entscheidungsprozesse eines KI-Systems verständlich und erklärbar sind. Dazu gehört die Möglichkeit zu sehen, wie ein KI-System zu einem bestimmten Ergebnis oder einer bestimmten Entscheidung gekommen ist, sowie die Möglichkeit, die Faktoren zu bewerten, die dieses Ergebnis beeinflusst haben. Bei TikTok ist es z. B. so, dass der Algorithmus vor allem auf die verwendeten Sounds, die Hashtags und die Captions der Videos zurückgreift, mit denen sich ein Nutzer oder eine Nutzerin beschäftigt hat, und empfiehlt ähnliche Inhalte. Die Transparenz der Ergebnisse ist wichtig, weil sie die Bewertung der Leistung und Fairness eines KI-Systems ermöglicht und dazu beitragen kann, Fehler oder Verzerrungen bei der Entscheidungsfindung zu erkennen und zu korrigieren.

Quelle [4]

3. Eine weitere Transparenzanforderung, die viel diskutiert wird, ist, dass Menschen die Wahl haben sollten, ob sie mit einem Menschen oder eine KI interagieren. Dazu müssen Menschen allerdings auch merken, ob sie mit einem KI-System oder einer Person interagieren, oder darauf hingewiesen werden, beispielsweise bei Chatbots, dass sie mit einer KI kommunizieren. Es muss also transparent sein, mit wem ich interagiere, Mensch oder KI.

Warum ist KI-Transparenz wichtig?

Es gibt mehrere Gründe, warum KI-Transparenz wichtig ist. Ein Grund ist, dass sie dazu beitragen kann, das Vertrauen in KI-Systeme zu stärken und das Risiko von Verzerrungen oder Fehlern zu verringern. Ein weiterer Grund ist, dass sie zu einem besseren Verständnis der Funktionsweise von KI-Systemen beitragen, was für Forscher*innen, Entwickler*innen und Nutzer*innen von KI-Systemen nützlich sein kann. In der KI-Forschung ist Transparenz außerdem wichtig, da das Fehlen von Details zu den Methoden und dem Programmcode des Algorithmus den wissenschaftlichen Wert der Erkenntnisse untergraben kann. Der wissenschaftliche Fortschritt hängt von der Fähigkeit unabhängiger Forscher*innen ab, die Ergebnisse einer Forschungsstudie zu überprüfen, die wichtigsten Ergebnisse der Studie unter Verwendung des Materials zu reproduzieren und in künftigen Studien darauf aufzubauen. Das gilt auch für KI-Forschung, die ebenso transparent sein muss.

Quelle [5]

Take-Home Message: Was solltest du heute mitnehmen?

Wir müssen über KI-Transparenz nachdenken und darüber diskutieren, da sie elementar ist für die Fähigkeit von Einzelpersonen und Organisationen, die Entscheidungsprozesse und Ergebnisse von KI-Systemen zu verstehen und zu interpretieren. Wir müssen uns fragen, ob Transparenz das entscheidende Kriterium ist, wenn es darum geht, Vertrauen in KI-Systeme zu haben. Darüber hinaus kann mangelnde Transparenz zum Beispiel zu Misstrauen und Bedenken hinsichtlich der Auswirkungen von KI auf die Gesellschaft führen, einschließlich Fragen im Zusammenhang mit Voreingenommenheit, Datenschutz oder auch Kontrolle.

Quellen

Quelle [1] *ChatGPT: Optimizing Language Models for Dialogue*. (2022, November 30). OpenAI. <https://openai.com/blog/chatgpt/>

Quelle [2] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>

Quelle [2] AI, H. (2019). High-level expert group on artificial intelligence. Ethics guidelines for trustworthy AI, 6.

Quelle [3] Diakopoulos, N. (2020). Transparency. In *The Oxford handbook of ethics of AI* (pp. 197–213). Oxford University Press.

Quelle [3] Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.

Quelle [4] *How TikTok recommends videos #ForYou*. (2019, August 16). Newsroom | TikTok. <https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you>

Quelle [5] Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Shraddha, T., Kusko, R., Sansone, S.-A., Tong, W., Wolfinger, R. D., Mason, C. E., Jones, W., Dopazo, J., Furlanello, C., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., ... Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829), E14–E16.
<https://doi.org/10.1038/s41586-020-2766-y>

Disclaimer

Transkript zu dem Video „W07_Ethik_Transparenz“, Marie von Lobenstein.

Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz [CC-BY](#) 4.0 veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.