Wie voreingenommen ist künstliche Intelligenz?

In unserer zunehmend digitaler werdenden Welt vertrauen wir immer mehr auf die Entscheidungen von KI-Systemen. Die von künstlicher Intelligenz (KI) getroffenen Entscheidungen werden dabei meist als neutral oder objektiv eingestuft, da diese auf Daten und Fakten basieren. Dies ist zunächst auch richtig, jedoch werden KI-Systemen die zugrunde liegenden Daten von Menschen zur Verfügung gestellt. Und das ist ein Problem.

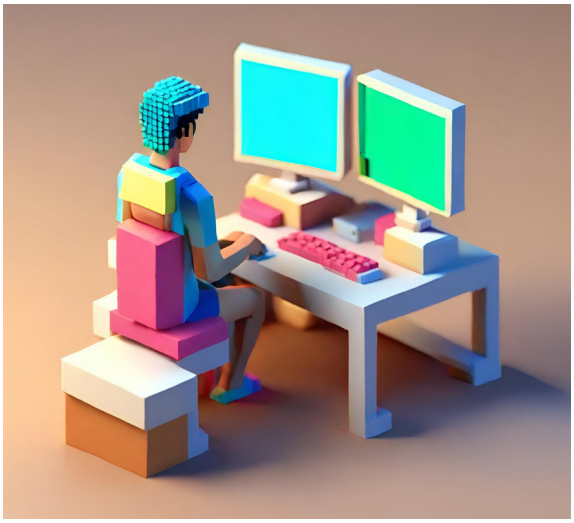

Bild generiert mit MagicMedia /Canva

Hierin liegt der kritische Punkt: Da Menschen eben nicht neutral und objektiv sind, können sich diese Verzerrungen auch unbeabsichtigt im Datensatz der KI wiederfinden. Dies kann zur Folge haben, dass KI-Systeme gesellschaftliche Ungleichheiten, Vorurteile und Stereotype nicht nur widerspiegeln, sondern auch verstärken und so zu Diskriminierungen führen, die rassistischer, sexistischer oder anderweitig gesellschaftlich schädlicher Natur sein können.

Die Relevanz der Auseinandersetzung mit KI wird noch deutlicher, wenn man bedenkt, dass Prognosen zufolge bis zum Jahr 2026 rund 90 Prozent der im Internet verfügbaren Inhalte KI-generiert sein könnten.[1] Diese Entwicklung zeigt nicht nur das immense Wachstum und die Durchdringung von KI-Technologien in unserem Alltag, sondern auch die Notwendigkeit, ethische und gesellschaftliche Implikationen dieser Technologien ernst zu nehmen und zu adressieren.

Die Bewältigung von Bias in KI-Systemen erfordert eine kritische Auseinandersetzung mit den inhärenten Vorurteilen und eine reflektierte Anwendung der Technologie, um Stereotypen zu erkennen und zu mindern. Lösungen umfassen nicht nur technische Anpassungen, sondern auch philosophische Überlegungen zu Fairness sowie rechtliche Rahmenbedingungen, die Unternehmen zur Verantwortung ziehen und Diversität sowie Transparenz fördern sollen.

Gründe für Bias

Einer der Hauptgründe für Bias in KI-Anwendungen ist die Zusammensetzung der Trainingsdaten. Wenn die Daten, die zum Trainieren einer KI verwendet werden, nicht ausreichend divers oder bestimmte Gruppen von Menschen über- oder unterrepräsentiert sind,

[1] *Europol (2022): Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg. Abgerufen am 7. Februar 2024, https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes*

kann dies dazu führen, dass die KI verzerrte Muster lernt. Dies kann beispielsweise zur Folge haben, dass eine KI zur Bilderkennung, die vornehmlich mit Bildern von hellhäutigen Personen trainiert wurde, dunkelhäutige Personen nicht korrekt erkennt.

Ein weiterer Grund für Bias liegt in den Algorithmen selbst. Diese können bestimmte Merkmale überbewerten, was dazu führt, dass sie diesen Merkmalen ein unverhältnismäßig großes Gewicht bei der Entscheidungsfindung beimessen. Dies kann zur Vernachlässigung anderer wichtiger Faktoren führen.

Darüber hinaus können die Entwicklerinnen und Entwickler, die die KI-Systeme programmieren, unbeabsichtigt ihre eigenen Vorurteile in die Systeme einfließen lassen. Diese menschlichen Bias können sich manifestieren, wenn Entscheidungen über die verwendeten Daten und die Gestaltung der Modelle getroffen werden.

Schließlich trägt die Natur vieler KI-Systeme als „Black Boxes“ zum Problem bei. Da die Entscheidungsfindungsprozesse dieser Systeme oft nicht transparent und nachvollziehbar sind, ist es schwierig, Bias zu erkennen und zu adressieren. Ohne klare Einblicke, wie eine KI zu ihren Schlussfolgerungen kommt, bleibt das Problem des Bias oft verborgen und unkorrigiert (siehe auch „Paint the Black Box White“: *https://www.bitkom.org/sites/default/files/file/import/171012-KI-Gipfelpapier-online.pdf,* ab Seite 160).

Paint the Black Box White (PDF ab Seite 160)

Probleme durch KI-Bias

Die Probleme und Auswirkungen von Bias in KI-Systemen sind weitreichend und können bestehende gesellschaftliche Ungleichheiten, Vorurteile und Stereotype nicht nur abbilden, sondern auch verstärken. Dies führt potenziell zu Diskriminierungen, die sich in rassistischen, sexistischen oder anderweitig schädlichen Formen manifestieren können.

Die Unterrichtsreihe macht dies durch eine Visualisierung deutlich, aber selbstverständlich verfügen auch KI-Anwendungen mit Textausgabe über diesen Bias.

Humans are biased. Generativ AI ist even worse (engl.)

Geschlechterstereotypen prägen die Wahrnehmung von Fähigkeiten und Rollen in unserer Gesellschaft, was in KI-Systemen oft durch die stereotype Darstellung von Berufsrollen zum Ausdruck kommt. Höher bezahlte und einflussreiche Positionen, wie die von CEOs, werden meist mit Männern assoziiert, während niedriger bezahlte Berufe oft Frauen zugeschrieben werden. Diese Verzerrungen fördern nicht nur eine ungleiche Wahrnehmung der Geschlechter, sondern zementieren auch Ungleichheiten in der realen Arbeitswelt.[2]

Das Familienstereotyp, das die Kernfamilie als Norm darstellt, ist ein weiteres Beispiel für einen Bias in KI-Systemen. Diese Sichtweise kann zur Marginalisierung alternativer Familienstrukturen führen und Heteronormativität fördern, was wiederum andere Beziehungsformen und sexuelle Orientierungen diskriminieren kann.

People of Color (PoC) können von rassistischen Stereotypen und Kriminalitätszuschreibungen betroffen sein. KI-Systeme, die solche Stereotype replizieren, tragen zur Verfestigung von Rassismus und zur Diskriminierung ethnischer sowie religiöser Gruppen bei. Dies schließt auch die Verbreitung von National- und Herkunftsstereotypen ein, die kulturelle Klischees fördern sowie zu einer verzerrten und vereinfachten Darstellung von Personen unterschiedlicher Herkunft führen.

[2] *Nicoletti, L. & Bass, D. (2023): Humans are biased. Generative AI is even worse. Bloomberg Technology. Abgerufen am 20. März 2024, https://www.bloomberg.com/graphics/2023-generative-ai-bias/?embedded-checkout=true*

Die Verstärkung von unrealistischen Schönheitsidealen und Sexismus durch die Darstellung von Frauen mit bestimmten äußeren Merkmalen ist ein weiteres Problem. Solche Darstellungen können schädliche Auswirkungen auf das Selbstbild und die Wahrnehmung der Geschlechterrollen haben.

Ungenügend geschulte KI-Modelle und schlechte Datengrundlagen haben in der Vergangenheit bereits zu unfairen Beurteilungen von Personen in der realen Welt geführt. So mussten etwa vor einigen Jahren in den Niederlanden Tausende Menschen aufgrund einer automatisierten Auswertung von diskriminierenden Daten, zu Unrecht staatliche Sozialleistungen zurückzahlen. Die Betroffenen kämpften jahrelang um Aufklärung.[3] In diesem Fall konnten die Datengrundlage und die Wichtung angeblicher Risikofaktoren für Sozialbetrug gut nachvollzogen werden – bei den heutigen KI-Systemen ist dies kaum noch möglich.

Notwendigkeit und Handlungsempfehlungen

Daraus folgt die Notwendigkeit, sich mit dem Thema Bias in KI-Systemen auseinanderzusetzen und eine kritische, reflektierte Nutzung dieser Technologien zu fördern. Es gilt zu untersuchen, wie häufig der Bias in Erscheinung tritt, welche Vorurteile und Stereotype besonders häufig auftreten, welche Prompt-Strukturen zu Verzerrungen führen und wie Prompts gestaltet werden können, um Verzerrungen zu reduzieren.

Damit Schülerinnen und Schüler dies im Rahmen dieser Unterrichtsreihe erarbeiten können, sind einige Voraussetzungen zu erfüllen. Es ist beispielsweise wichtig, dass bei der Verwendung von KI-Anwendungen im schulischen Kontext datenschutzrechtliche Bestimmungen eingehalten werden. Zu diesem Zweck sollten Lehrkräfte sich vor der Verwendung einer KI-Anwendung über die Datenschutzbestimmungen informieren und sicherstellen, dass die Anwendung den jeweiligen Anforderungen entspricht. Insbesondere sollten sensible Daten wie Name, Adresse oder Kontaktdaten nicht in die Anwendung eingegeben werden. Die Nutzung von KI-Tools ohne Anmeldung ist zu bevorzugen. Mögliche bildgenerierende KI-Anwendungen sind beispielsweise Dall-E 3, Stable Diffusion XL, Leonardo AI und Ideogram. Grundsätzlich kommen jedoch alle KI-Tools infrage, sofern sie den Datenschutzbestimmungen entsprechen.

Die Schülerinnen und Schüler sollten zudem bereits grundlegende Kenntnisse in der Bedienung und Nutzung von KI-Tools besitzen, auch die Lehrkraft benötigt in diesem Kontext zumindest erste Kenntnisse im Umgang mit generativer KI. Sofern diese Kenntnisse nicht vorhanden sind, ist es sinnvoll, zuvor einige KI-Systeme auszuprobieren und herauszufinden, wie sich unterschiedliche Prompts auf die Ergebnisse auswirken. Wenn in diesem Punkt Schülerinnen und Schüler möglicherweise sogar mehr Erfahrungen haben als die Lehrkraft, ist dies für das das Unterrichtsziel, das darin besteht, die Voreingenommenheit der KI-Anwendungen zu identifizieren, unerheblich.

[3] *Dachwitz, I. (2021): Niederlande zahlen Millionenstrafe wegen Datendiskriminierung; Netzpolitik.org. Abgerufen am 20. März 2024, https://netzpolitik.org/2021/kindergeldaffaere-niederlande-zahlen-millionenstrafe-wegen-datendiskriminierung/*

Lösungsansätze

Die Lösung dieses Problems „KI-Bias“ erfordert mehr als nur technische Anpassungen; sie erfordert eine Auseinandersetzung mit tiefgreifenden philosophischen Fragen nach der Definition von Fairness und Bias. Die Frage, wie beispielsweise das Geschlechterverhältnis bei der Darstellung von CEOs aussehen sollte, verdeutlicht die Komplexität des Problems. Sollte die Darstellung die reale Verteilung (neun männliche CEOs zu einer weiblichen CEO bei den Fortune-500-Unternehmen)[4] widerspiegeln oder ein idealisiertes Verhältnis anstreben, um Ungleichheiten entgegenzuwirken? Die Antwort auf diese Frage ist nicht einfach und wirft weitere Fragen nach der fairen Darstellung anderer sozialer Gruppen und Identitäten auf.

Europäischer Ansatz für künstliche Intelligenz

Regierungen und Aufsichtsgremien haben die Möglichkeit, Gesetze und Vorschriften zu erlassen, um der Verbreitung von Vorurteilen durch KI entgegenzuwirken. So könnten Unternehmen mehr in die Verantwortung und im Zweifel zur Rechenschaft gezogen werden. Dies könnte auch die Festlegung von Standards für Trainingsdatensätze, die Förderung von Transparenz und Diversität sowie die Einrichtung von Beschwerdestellen umfassen.[5] Die EU hat dazu erste Bestrebungen angestellt (*https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence*).

KI-Regulierung: Was sollen künstliche Intelligenzen dürfen?

Interessant könnte eine Diskussion darüber sein, welche Gremien demokratisch legitimierte Kriterien oder Regelungen treffen dürften um dem KI-Bias entgegenzuwirken. Dazu ist als Hintergrundinformation der Beitrag „KI-Regulierung: Was sollen künstliche Intelligenzen dürfen?“ im Spektrum der Wissenschaften erhellend: *https://www.spektrum.de/news/ki-regulierung-was-soll-kuenstliche-intelligenz-duerfen/2165157*

Darin heißt es: „Mangel an gesichertem Wissen erschwert die Regulierung deutlich: Die neue Technologie ist noch zu wenig erforscht, um langfristige Folgen zu überblicken und realistisch einschätzen zu können.“[6] Somit sind nicht allein die KI-Bias, sondern die Anwendungen an sich so komplex, dass sich neben der Frage „Wer darf regulieren?“ die spannende Frage nach dem „Wie“ stellt.

---

[4] *LIS - The London Interdisciplinary School (11. August 2023): How AI image generators make bias Worse [Video]. YouTube. Abgerufen am 20. März 2024, https://www.youtube.com/watch?v=L2sQRrf1Cd8*

[5] *Ebenda*

[6] *Wolfangel, E. (2021): Was sollen künstliche Intelligenzen dürfen? In: Spektrum der Wissenschaften. Abgerufen am 20. März 2024, https://www.spektrum.de/news/ki-regulierung-was-soll-kuenstliche-intelligenz-duerfen/2165157*

Impressum

DGUV Lernen und Gesundheit, Bias: Wie objektiv ist KI?, März 2024
Herausgegeben von: Deutsche Gesetzliche Unfallversicherung e.V. (DGUV), Glinkastraße 40, 10117 Berlin, **Chefredaktion:** Kathrin Baltscheit (V.i.S.d.P.), DGUV, Berlin
Redaktion: Stefanie Richter, Universum Verlag GmbH, Wiesbaden, *www.universum.de*
E-Mail Redaktion: *info@dguv-lug.de*
Text: Manuel Flick, Berlin

Internet-hinweis · Arbeits-blätter · Arbeits-auftrag · Präsentation · Video · Didaktisch-methodischer Hinweis · Lehr-materialien · Distanz-unterricht