Rassismus durch Künstliche Intelligenz
Ein Beitrag von Laura Schelenz, Universität Tübingen

Rassismus ist in der Gesellschaft weit verbreitet, insbesondere in der Form von Alltagsdiskriminierung (Hasters 2020). Nun gibt es bedenken, dass Künstliche Intelligenz rassistische Verhältnisse verstärken kann, etwa durch die automatisierte Erkennung von stereotypen Mustern und die Verbreitung von rassistischen Inhalten, etwa über soziale Medien. Wie kommt Rassismus in KI-Systeme? Und was sind Lösungsansätze?

Was ist Rassismus?
Bei Rassismus geht es um eine künstliche Differenzierung zwischen Gruppen und das Ausnutzen von sozial konstruierten Unterschieden zur Bevorteilung einer Gruppe und Unterdrückung anderer Gruppen. Eine häufig genutzte Definition von Rassismus stammt von Ruth Gilmore und lautet: "the state-sanctioned or extralegal production and exploitation of group-differentiated vulnerability to premature death" (Gilmore, S. 28). Es geht also um die teils extreme Benachteiligung einer oder mehrerer Gruppen basierend auf „Unterschieden". Dabei sind diese Unterschiede sozial konstruiert, das bedeutet, sie sind nicht natürlich, sondern von Menschen erfunden.
Ideen und sogar vermeintlich wissenschaftliche Studien von „Rasse" haben sich lange Zeit durch die Geschichte gezogen und der Legitimierung von Sklaverei in den USA oder der Ermordung von Juden, Sinti und Roma in Europa gedient (siehe auch Subramaniam 2014). In Europa werden heutzutage vielmehr kulturelle Unterschiede formuliert, um Gruppen in der Gesellschaft zu differenzieren und gegeneinander auszuspielen. So sind etwa Muslime in Deutschland stark von einem „kulturellen" Rassismus betroffen, der sich in der Diskriminierung von Frauen mit Kopftuch zeigt (Salem und Thompson 2016).
Eine Form des Rassismus ist der offene Übergriff auf Menschen mit Migrationshintergrund bzw. Menschen mit dunkler Hautfarbe. Andere Formen des Rassismus sind unterschwellige Beleidigungen, Witze, und Mikroagressionen, die auch als Alltagsdiskriminierung verstanden werden (Sue 2010). Darüber hinaus gibt es auch den strukturelle Rassismus: ein Zusammenspiel von institutionellen Strukturen, Verhaltensweisen und sozialisierten Denkweisen. Zum Beispiel werden über kulturelle Produktionen wie Filmen, Büchern und Musik Stereotype seit der Kindheit verinnerlicht. So kann sich Rassismus mitunter sehr hartnäckig halten.

Wie kommt Rassismus in die KI?
KI basiert auf der automatisierten Mustererkennung und kann diese Muster selbst lernen und verbreiten. Dabei spielen Datensätze eine zentrale Rolle. Die KI analysiert große Datenmengen und versucht, Strukturen darin zu identifizieren. Sind die Daten mit Voreingenommenheiten gegenüber bestimmten Gruppen der Gesellschaft versehen, übernimmt die KI diese. Ein Beispiel ist die Nutzung des KI-Chatbots ChatGPT. Das unterliegende Sprachmodell ist mit unzähligen Quellen aus der Gesellschaft gefüttert, etwa im Internet frei verfügbaren Berichte, Blog-Artikeln, Social Media-Posts, Büchern oder Zeitungsartikeln. Diese können wiederum problematische Vorstellungen von Unterschiedlichkeit beinhalten. In ChatGPT zeigen sich dann mitunter stereotype Antworten auf Anfragen zu Jobvorschlägen. Auf die Frage "Welche Berufsvorschläge kann ich meiner Schülerin Sediqa machen?" werden womögliche Berufe mit weniger Qualifizierungsbedarf und niedriger Entlohnung vorgeschlagen.
Ein anderes Problem ist die Unausgeglichenheit von Datensätzen. Wenn etwa Datensätze nur die Informationen *einer* Gruppe beinhalten (z.B. der weißen Mehrheitsgesellschaft), können manche KI-basierte Systeme schlechter für unterrepräsentierte Gruppen funktionieren. Ein

Beispiel ist die Gesichtserkennung, die Menschen mit heller Hautfarbe besser erkennt als Menschen mit dunkler Hautfarbe (Buolamwini und Gebru 2018). Wird Gesichtserkennungstechnologie zur Strafverfolgung eingesetzt, kann die mangelnde Qualität der Daten bzw. der auf diesen Daten trainierten KI zu Falschidentifikationen und fehlerhaften Verhaftungen führen.
Ein weiteres Hindernis bei der Entwicklung diskriminierungsarmer KI kann die Zusammensetzung des Entwicklungsteams sein. Häufig besteht das Team der Entwickler:innen aus Mitgliedern der Mehrheitsgesellschaft und dadurch gibt es in westlichen Unternehmen weniger Berührungspunkte mit den Bedürfnissen von diversen Bevölkerungsgruppen. Gleichzeitig gelten weiße und männliche Nutzer:innen häufig als die "Norm" beim Design eines Produktes. Das bedeutet, es wird nicht nur *von* Mitgliedern der Mehrheitsgesellschaft designt, sondern auch *für* Mitglieder der Mehrheitsgesellschaft (Noble 2018, S. 91). Zuletzt gilt Technik wie KI oder KI-basierte Plattformen in der öffentlichen Meinung häufig als neutral, rational und diskriminierungsarm im Vergleich zu von Menschen gesteuerten Systemen (Benjamin 2019, S. 41). So kann es passieren, dass Risiken der KI oder Online-Interaktion für diverse Bevölkerungsgruppen (etwa automatisierte Diskriminierung) nicht ernst genommen werden.

Ein Beispiel von KI-mediierter Diskriminierung: Hatespeech
Hatespeech oder auch Hassrede bezeichnet hasserfüllte und diskriminierende Kommunikationspraktiken, zu denen u.a. Rassismus, Sexismus, Islamophobie, Homophobie und Anti-Semitismus gehören (Lumsden und Harmer 2019, S. 4). Die Verbreitung von Hatespeech in Online-Medien wie sozialen Plattformen ist nicht nur durch individuelles Nutzungsverhalten zu erklären. Technische Einstellungen und KI-gestützte algorithmische Mediierung von Inhalten spielen eine wichtige Rolle. Die Wissenschaftlerin Safiya Umoja Noble kritisiert, dass die großen Tech-Unternehmen Rassismus in ihren Systemen als isolierte Fehler entschuldigen (Noble 2018, S. 6). Noble verweist auf den Fall von Google Maps, wo das Weiße Haus in den USA während der Obama-Regierung als N*-Haus gekennzeichnet wurde. Google hatte sich für das Problem entschuldigt, aber Noble argumentiert, dass die Architektur von Google-Systemen Rassismus in seinen Produkten fördert (Noble 2018, S. 9).
Im Zusammenhang mit Hassrede in den sozialen Netzwerken werden die Stimmen Schwarzer Menschen häufig unterdrückt. Dabei ist Hassrede in sozialen Plattformen Schwarzen Frauen gegenüber anders gelagert als Schwarzen Männern. Rassismus hängt eng mit Sexismus zusammen und bedingt sich mitunter gegenseitig. Man spricht hier allgemein von intersektionaler Diskriminierung oder im Fall von Schwarzen Frauen von "Misogynoir" (Bailey 2021). Ein klassischer Ansatz zur Bekämpfung von Hassrede ist es, Vorfälle zu verfolgen, zu melden und rasch zu entfernen. Die Entfernung von Inhalten auf Online-Plattformen wird teilweise durch KI-basierte Software vollzogen, welche problematische Inhalte automatisch erkennt (z. B. über Texterkennung) und löscht. Diese Software kann jedoch speziell Misogynoir nicht erkennen, und weist darüber hinaus rassistischen und geschlechtsspezifischen Bias auf (Kwarteng et al. 2022). So kann die Software fälschlicherweise Posts mit Ausdrücken markieren und entfernen, die als Hassrede gelten, aber von der betroffenen Gruppe in einer politisch ermächtigenden Weise verwendet werden (z. B. das N-Wort). Dies kann dazu führen, dass bereits marginalisierte Nutzer:innen durch die automatisierte Entfernung von Inhalten mundtot gemacht werden (Davidson et al. 2019). Hier zeigt sich, dass der Kontext für die Entfernung von Hassrede in sozialen Plattformen eine wichtige Rolle spielt; Kulturen des Widerstands gegen Rassismus sollten berücksichtigt werden.

Lösungen für Rassismus durch KI

Auf der Suche nach Lösungen für Rassismus durch KI gibt es unterschiedliche Ansätze. Zum einen gibt es technische Ansätze wie die oben genannte automatisierte Entfernung von rassistischen Inhalten. Jedoch zeigen sich hierbei meist eigene Biases und unerwünschte Nebeneffekte wie etwa die Zensur von Gegenrede oder politischer Rede der von Rassismus betroffenen Minderheiten.

Ein Ansatz, der auf die bewusst faire und ausgeglichene Gestaltung von Technikprodukten abzielt, ist das diversitätssensible Technikdesign. Hierbei werden Diversitäts-Konzepte herangezogen und in ein Computersystem eingeschrieben. So können Aspekte wie die Vielfalt unter den Nutzer:innen einer App oder Plattform berücksichtigt werden (Schelenz 2023). Technik diversitätssensibel zu gestalten ist ein wichtiger Ansatz, um die häufig unbewussten Normen in der Technikentwicklung zu hinterfragen (Wachter-Boettcher 2017, 27ff). Ist der anvisierte Nutzer ein weißer technikaffiner und berufstätiger Mann in der Großstadt? Oder gibt es auch Raum in der Visualisierung von Nutzer:innen für Menschen in sozial prekären Situationen, Menschen mit Migrationshintergrund, Menschen mit einer körperlichen Beeinträchtigung, mit mangelnder Bildung? Wie müsste das Produkt dann anders aussehen, um die Bedürfnisse von Minderheiten zu reflektieren? Diversitätssensibles Technikdesign hat das Potenzial, solch eine Diskussion über die Bedeutung von Vielfalt in der Gesellschaft und ihrem Abbild in KI-basierten Produkten anzustoßen.

Zuletzt für die Eindämmung von Rassismus durch KI wichtig, zunächst einmal Rassismus in der Gesellschaft zu bekämpfen. Denn letztendlich spiegeln sich gesellschaftliche Verhältnisse über unsere Daten in der KI. Um Rassismus zu minimieren, könnte man etwa dort ansetzen, wo gesellschaftliche Verhältnisse produziert werden: Sozial konstruierte Unterschiede zwischen Menschen (etwa in Kategorien von „Rasse“, Herkunft, Geschlecht, Kultur, Sexualität) dürfen nicht als natürlich oder selbstverständlich akzeptiert werden. Sie wurden in einem bestimmten historischen Kontext mit einer gewissen Motivation entwickelt (Zum Beispiel während der Kolonialisierung zur Versklavung und Ausbeutung Schwarzer Menschen, vgl. Rusert 2017). Die Denkmuster und daraus resultierende Handlungsweisen der Mehrheitsgesellschaft in Frage zu stellen und zu verändern ist daher zentral, um auch Rassismus durch KI entgegenzuwirken.

Literaturverzeichnis

Bailey, Moya (2021): Misogynoir Transformed. Black Womens Digital Resistance: New York University Press.

Benjamin, Ruha (2019): Race after Technology : Abolitionist Tools for the New Jim Code. Cambridge, Medford, MA: Polity Press.

Buolamwini, Joy; Gebru, Timnit (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proceedings of Machine Learning Research* 81, S. 1–15. Online verfügbar unter http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf, zuletzt geprüft am 18.08.2018.

Davidson, Thomas; Bhattacharya, Debasmita; Weber, Ingmar (2019): Racial Bias in Hate Speech and Abusive Language Detection Datasets. In: Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran und Zeerak Waseem (Hg.): Proceedings of the Third

Workshop on Abusive Language Online. Proceedings of the Third Workshop on Abusive Language Online. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 25–35.

Gilmore, Ruth Wilson: Golden Gulag. Prisons, Surplus, Crisis, and Opposition in Globalizing California: University of California Press.

Hasters, Alice (2020): Was weiße Menschen nicht über Rassismus hören wollen, aber wissen sollten. 10. Auflage. München: hanserblau.

Kwarteng, Joseph; Perfumi, Serena Coppolino; Farrell, Tracie; Third, Aisling; Fernandez, Miriam (2022): Misogynoir: Challenges in Detecting Intersectional Hate. In: *Social Network Analysis and Mining* 12 (1). DOI: 10.1007/s13278-022-00993-7.

Lumsden, Karen; Harmer, Emily (2019): Online Othering. Exploring Digital Violence and Discrimination on the Web. Cham: Springer International Publishing (Palgrave Studies in Cybercrime and Cybersecurity).

Noble, Safiya Umoja (2018): Algorithms of Oppression. How Search Engines Reinforce Racism. New York: New York University Press.

Rusert, Britt (2017): Fugitive Science. Empiricism and Freedom in Early African American Culture. New York: New York University Press (America and the Long 19th Century).

Salem, Sara; Thompson, Vanessa (2016): Old Racisms, New Masks: On the Continuing Discontinuities of Racism and the Erasure of Race in European Contexts. In: *Nineteen Sixty Nine: An Ethnic Studies Journal* 3 (1). Online verfügbar unter https://escholarship.org/uc/item/98p8q169#main, zuletzt geprüft am 21.02.2021.

Schelenz, Laura (2023): Diversity and Social Justice in Technology Design. Reflections on Diversity-aware Technology. In: *International Journal of Critical Diversity Studies* 5 (2). DOI: 10.13169/intecritdivestud.5.2.0033.

Subramaniam, Banu (2014): Ghost Stories for Darwin. The Science of Variation and the Politics of Diversity. Urbana, Illinois: University of Illinois Press.

Sue, Derald Wing (2010): Microaggressions in Everyday Life. Race, Gender and Sexual Orientation. John Wiley.

Wachter-Boettcher, Sara (2017): Technically Wrong. Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech. New York NY: W.W. Norton & Company.