KI und Gewalt im digitalen Raum

14.11.2024 Anja Roß

Künstliche Intelligenz wird als Schlüsseltechnologie der Gegenwart bezeichnet: Sie kann aus Daten lernen und dadurch Muster erstellen, die Probleme lösen, Vorhersagen oder Entscheidungen treffen und Aktionen schneller ausführen. Es gibt kaum noch Lebensbereiche, die nicht durch KI bzw. algorithmische Entscheidungssysteme (AES) unterstützt werden. Aber wie sieht der Einsatz von KI aus, wenn es um Gewalt im digitalen Raum geht? Hilft sie uns, Gewaltverbrechen aufzuklären und Betroffene besser zu schützen, oder wird sie als weiteres Werkzeug eingesetzt, um Hass und Hetze im Netz weiter voranzutreiben?

Digitale versus analoge Gewalt

Durch die Digitalisierung und die scheinbar unbegrenzten Möglichkeiten, die damit verbunden sind, nimmt auch Gewalt neue Formen an und verbreitet sich zunehmend im Internet. Digitale Gewalt ist dabei als eigenständiges Phänomen zu betrachten und nicht einfach als verlängerter Arm analoger Gewalt.

Denn digitale Gewalt unterscheidet sich grundlegend von analoger Gewalt: Digitale Gewalt kann sich sekundenschnell über geografische Grenzen hinweg verbreiten und ist nicht auf einen bestimmten physischen Ort oder ein direktes Umfeld festgelegt. Inhalte, die einmal im Internet verbreitet wurden, können dauerhaft bestehen bleiben. Auch wenn das ursprüngliche Material, z. B. aufgrund einer Strafanzeige, gelöscht worden ist, kann es durch Screenshots oder Weiterverbreitung weiterhin zugänglich sein. Während bei Gewalt im analogen Raum Täter*innen oftmals bekannt sind und aus dem näheren Umfeld des Opfers kommen, können Menschen im Internet anonym handeln, was deren Identifikation erschwert und das Gefühl von Straflosigkeit verstärken kann. Digitale Gewalt kann durch das Internet eine Vielzahl von Täter*innen involvieren, die sich an Shitstorms oder Mobbing beteiligen.

Deepfakes als neue Form KI-generierter Gewalt

Technologien wie künstliche Intelligenz ermöglichen zudem neue Formen der Gewalt, die über herkömmliche Erscheinungsformen hinausgehen und neue Dimensionen annehmen. Eine besonders populäre und besorgniserregende Form sind sogenannte Deepfakes (Englisches Wort aus den Begriffen „deep learning“ und „fake“). Deepfakes sind digital manipulierte Medieninhalte, meist Videos oder Bilder, bei denen Künstliche Intelligenz verwendet wird, um das Gesicht oder die Stimme einer Person täuschend echt durch eine andere zu ersetzen. Diese Technologie nutzt maschinelles Lernen, um realistische Fälschungen zu erstellen, die oft schwer von echten Inhalten zu unterscheiden sind. Deepfakes können für harmlose Zwecke wie Unterhaltung eingesetzt werden, werden aber zunehmend für politische Desinformation, Betrug oder sexualisierte Gewalt genutzt (Böttcher et al. 2024).

Zum Missbrauchspotenzial von Deepnudes

Eine besondere Form von Deepfakes sind sogenannte Deepnudes (zusammengesetzt aus dem Wort "deepfake" und dem englischen Wort "nude" für nackt). Deepnudes sind KI-generierte Nacktbilder, die sich einfach mit kostenlosen Apps herstellen lassen. Um ein manipuliertes Nacktbild zu erstellen, muss lediglich ein Foto einer bekleideten Person hochgeladen werden. Wenige Sekunden später produziert die Seite oder App ein manipuliertes Bild, auf dem es wirkt, als sei die Person in Unterwäsche, Bikini oder nackt. Um die manipulierten Bilder zu optimierten, gibt es eigens dafür eingerichtete Foren, in welchen Nutzer*innen darüber diskutieren, wie die besten Ergebnisse erzielt werden können. Von solchen Deepfakes betroffen sind in erster Linie Mädchen, Frauen oder queere Personen. Während die Betroffenen mit der erlebten Scham und Demütigung oftmals alleingelassen werden, bergen Deepnudes zusätzlich ein großes Missbrauchspotenzial und können für Mobbing oder Erpressung genutzt werden (vgl. Köver 2023).

Social Bots als Mittel antifeministischer Gewalt

Ein weiteres Beispiel für KI-generierten Hass bzw. die Unterstützung digitaler Gewalt durch künstliche Intelligenz zeigt sich am Beispiel feministischer Online-Aktivist*innen, die vehement von antifeministischer Gewalt betroffen sind (vgl. Roß 2024). In meiner Studie wurde deutlich, dass antifeministische Akteur*innen gezielt KI-generierte Tools einsetzen, um ihren Hass weiter zu verbreiten und den Eindruck zu vermitteln, dass sie viele sind (vgl. Roß 2024). Die Rede ist hier von Social Bots, von Programmen, die sich mithilfe von Algorithmen in sozialen Netzwerken bewegen und dort Beiträge posten, in diesem Fall Hasspostings. Durch ihre Aktivität wird Nutzer*innen suggeriert, dass hinter den Beiträgen echte Menschen stehen. In Wirklichkeit folgen die Bots jedoch einer im Programm festgelegten Leitlinie und versuchen unter Vernetzung mit anderen Bots und realen Nutzer*innen durch Kommentare, Likes und das Teilen von Postings die Reichweite bestimmter Beiträge zu erhöhen (vgl. Graber/Lindemann 2018).

Die Macht der Bots

Stephanie, ein*e feministisch*e Aktivist*in, erzählt mir im Interview, wie sich der Hass durch Social Bots verändert hat:

> Und dann war da einfach der - also der - der Hass so schnell und so organisiert und krass auch mit Bots, und so (Stephanie, Pos. 373). […] Früher war das so, dann hatte man zwei, drei, vier blöde Kommentare aufm Blog, oder Antworten auf Twitter - oder zwei Mails, oder so. Und heute -- hat man dann direkt, irgendwie -- weiß ich nicht, hundert Bots, die einem ein und dieselbe Nachricht schicken, dadurch aber das komplette System irgendwie lahmlegen (Stephanie, Pos. 467).

Auch wenn die Aktivist*innen KI-generierten Hass identifizieren können und sich dadurch, wie in der Studie gezeigt wird, nicht einschüchtern lassen, so stellt sich die Frage, welche Macht Social Bots haben und inwiefern sie Diskurse manipulieren und Menschen einschüchtern können (vgl. Roß 2024).

KI im Kampf gegen (geschlechtsspezifische) Gewalt

In vielen Institutionen wird bereits mit KI gearbeitet, um digitale Gewalt und Hasskriminalität zu bekämpfen. So setzen deutsche Medienanstalten, deren Aufgabe es

ist, Vielfalt und Meinungsfreiheit in den privaten Rundfunk- und Onlinemedien zu gewährleisten, auf den Einsatz Künstlicher Intelligenz. Eine eigens für die Medienaufsicht entwickelte KI soll die Arbeit der Medienaufsicht beschleunigen und verbessern, indem etwa das Internet nach Verstößen gegen die Menschenwürde, volksverhetzende Äußerungen oder die Verwendung von verfassungsfeindlichen Kennzeichen durchsucht wird. In NRW werden z. B. rechtsradikale und jugendgefährdende Inhalte automatisch aus 10.000 Seiten pro Tag herausgefiltert (vgl. Klüppel 2023).
Auch im präventiven Bereich findet KI seine Anwendung. Mithilfe von Algorithmen des maschinellen Lernens können KI-Systeme menschliches Verhalten, Social-Media-Hilfe-Nachrichten oder Notfalltelefonanrufe analysieren und mögliche Fälle (geschlechtsspezifischer) Gewalt erkennen. Künstlicher Intelligenz wird dabei das Potenzial zugesprochen, Lösungen in den Bereichen Prävention, Erkennung und Opferhilfe zu bieten.

Daten, Algorithmen und (digitale) Gewalt

Wenn über KI und Gewalt im digitalen Raum gesprochen wird, ist es unumgänglich auch über Biases in Daten und diskriminierende Algorithmen zu sprechen. Besonders deutlich wird dies aktuell im Kontext europäischer Migrationspolitik: An europäischen Außengrenzen werden KI-Systeme, wie z. B. „iBorder Ctrl“, eingesetzt, um herauszufinden, ob Geflüchtete die Wahrheit sagen. Diese Emotionserkennungssysteme, deren Einsatz in anderen Bereichen verboten ist, sind jedoch stark fehlerhaft und diskriminierend. So schätzen sie einer Studie zufolge Schwarze Basketballspieler auf Fotos als aggressiver ein als *weiße* Sportler (vgl. Rohrbach 2024).

Ähnliche Beispiele lassen sich im Kontext KI-generierter Gesichtserkennung wiederfinden. Da diese vorwiegend an *weißen* männlichen Probanden trainiert wird, weist sie insbesondere bei Frauen und People of Colour höhere Fehlerdaten auf. Dies kann zur Folge haben, dass z. B. eine Schwarze Frau zu Unrecht eines Verbrechens verdächtigt wird, weil sie der gesuchten Person ähnlich sieht. Und auch KI-Systeme, die Vorhersagen zur Rückfälligkeit von Straftäter*innen treffen, werden häufig mit historischen Daten trainiert, die bestehende Vorurteile und diskriminierende Praktiken widerspiegeln. Solche fehlerhaften Einschätzungen können härtere Strafen oder vermehrt Polizeikontrollen zur Folge haben.

Diese Beispiele machen deutlich, dass Auswahl der Daten, deren Qualität und Zusammensetzung großen Einfluss auf die Entscheidungen haben, die durch Algorithmen getroffen werden. Denn „Algorithmen, die auf Künstlicher Intelligenz basieren, […] finden Muster innerhalb von Datensätzen, die unsere eigenen Vorurteile widerspiegeln, und stellen dadurch unsere Vorurteile als allgemeingültige Wahrheit dar bzw. verstärken diese“ (Howard/Borenstein 2018, 1524). Es ist also wichtig, diese Perspektive aufzugreifen und zu verdeutlichen, an welcher Stelle der Entwicklung algorithmischer Entscheidungssysteme bereits Ungleichheiten und Diskriminierungen eingeschrieben sind, die auch zu Gewalt führen können.

Literatur

Behme, Pia; Böttcher, Martin; Kogel, Dennis & Terschüren, Hagen (2024): Digitale Gewalt. KI-generierte Nacktbilder und langsame Gesetze. https://www.deutschlandfunkkultur.de/digitale-gewalt-nacktbilder-durch-ki-und-spaete-gesetze-dlf-kultur-ef457a59-100.html, zuletzt geprüft am 14.11.24.

Graber, Robin & Lindemann, Thomas (2018): Neue Propaganda im Internet. Social Bots und das Prinzip sozialer Bewährtheit als Instrumente der Propaganda. In: Sachs-Hombach, Klaus & Zywietz, Bernd (Hrsg.): Fake News, Hashtags & Social Bots. Neue Methoden populistischer Propaganda. Wiesbaden: Springer VS, S. 51–68. https://doi.org/10.1007/978-3-658-22118-8_3

Howard, Ayanna & Borenstein, Jason (2018): The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social

Inequity. In: Science and Engineering Ethics 24 (5), S. 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Köver, Chris (2023): Deepfakes in Spanien: Gefälschte Nacktbilder von Mädchen sorgen für Aufschrei, https://netzpolitik.org/2023/deepfakes-in-spanien-gefaelschte-nacktbilder-von-maedchen-sorgen-fuer-aufschrei/, zuletzt geprüft am 14.11.24

Klüppel, Manuela (2023): KI gegen Hass und Gewalt im Netz, https://www1.wdr.de/nachrichten/rheinland/duesseldorf-ki-gegen-hass-im-internet-100.html, zuletzt geprüft am 14.11.24.

Rohrbach, Lena (2024): Künstliche Intelligenz und Menschenrechte – aber nicht für alle, https://www.gwi-boell.de/de/2024/09/11/kuenstliche-intelligenz-und-menschenrechte-aber-nicht-fuer-alle-die-ki-verordnung-der-eu, zuletzt geprüft am 14.11.24.

Roß, Anja (2024): Feminismus im Netz. Eine qualitative Studie zum Umgang mit digitaler Gewalt. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-44614-7

Weitere Literatur zum Thema:

Guijarro-Santos, Victoria (2020): Effiziente Ungleichheit. In: netzforma* e.V. (Hg.): Wenn KI dann feministisch. Impulse aus Wissenschaft und Aktivismus, S. 47–64, https://netzforma.org/wp-content/uploads/2021/01/2020_wenn-ki-dann-feministisch_netzforma.pdf, zuletzt geprüft am 14.11.24.

Neu, Matthias; Müller, Melanie; Pothen, Biju & Zingel, Moritz (2022): Anwendungsfelder und Herausforderungen der Künstlichen Intelligenz. Wiesbaden: Springer Gabler. https://doi.org/10.1007/978-3-658-38891-1

Zhang, Shunyuan; Mehta, Nitin; Singh, Param Vir & Srinivasan, Kannan (2021): Can an AI algorithm mitigate racial economic inequality? An analysis in the context of Airbnb. Rotman School of Management Working Paper No. 3770371. http://dx.doi.org/10.2139/ssrn.3770371

Zitation

Anja Roß: KI und Gewalt im digitalen Raum, in: blog interdisziplinäre geschlechterforschung, 14.11.2024, www.gender-blog.de/beitrag/ki-und-digitale-gewalt/, DOI: https://doi.org/10.17185/gender/20241114

DOI: 10.17185/gender/20241114
URN: urn:nbn:de:hbz:465-20241114-103445-4