

Verengung öffentlicher Diskursräume auf Social Media durch Biases in KI, digitale Gewalt und Desinformation: Eine intersektional-feministische Analyse

Soziale Medien galten einst als Orte demokratischer Öffentlichkeit, als digitale Arenen, in denen Teilhabe, Partizipation und Meinungsfreiheit realisiert werden können. Doch seit langem zeigt sich: Diese Räume sind von massiven Machtasymmetrien, Ausschlussmechanismen und digitalen Gewaltformen geprägt. Hassrede, algorithmische Verzerrungen (Biases) und Desinformation verengen den Raum öffentlicher Debatten. Besonders betroffen sind marginalisierte Gruppen: Frauen, queere Personen, People of Colour (PoC), Menschen mit Behinderungen, religiöse Minderheiten oder Menschen aus finanziell schwächeren Schichtungen – Gruppen, die ohnehin strukturellen Diskriminierungen ausgesetzt sind.

Ein intersektional-feministischer Blick auf diese Entwicklungen verdeutlicht, dass es nicht nur um technische Herausforderungen geht, sondern um tief verwurzelte gesellschaftliche Ungleichheiten, die durch künstliche Intelligenz (KI) und Plattformlogiken nicht neutral reproduziert, sondern verstärkt werden. Oder wie es Safiya Umoja Noble (2018: 3) formuliert: „The internet was supposed to be a great equalizer – instead, it mirrors and magnifies our worst inequalities.“

Der vorliegende Beitrag untersucht, wie Biases in KI, digitale Gewalt und Desinformation zur Verengung öffentlicher digitaler Diskursräume beitragen, welche intersektionalen Dynamiken dabei wirksam werden und welche politischen sowie zivilgesellschaftlichen Handlungsperspektiven denkbar sind.

In der Praxis zeigt sich also, dass digitale Gewalt, algorithmische Verzerrungen und Desinformation in sozialen Medien zunehmend zur Verengung der Diskursräume beitragen. Statt mehr Partizipation und Gleichheit entstehen Ausschluss, Silencing und Radikalisierung. Um diese Mechanismen zu verstehen, ist der Begriff der *Intersektionalität*, geprägt von Kimberlé Crenshaw (1989), zentral. Er beschreibt, dass Diskriminierung und Unterdrückung nicht additiv entlang einzelner Kategorien (wie Geschlecht/gender, Hautfarbe/race oder Klasse/class) verlaufen, sondern sich überschneidende und verstärkende Effekte entstehen. Eine intersektional-feministische Analyse der oben beschriebenen gewaltförmigen Phänomene macht schnell deutlich, dass es sich dabei nicht um technische Nebeneffekte handelt, sondern um die digitale Fortsetzung tief verwurzelter gesellschaftlicher Herrschafts- und Ausschlussverhältnisse (Mosene 2024).

Social-Media-Plattformen basieren auf Aufmerksamkeitsökonomien: Sichtbar wird, was Klicks, Likes und Interaktionen verspricht. Algorithmen priorisieren Inhalte, die Empörung hervorrufen – oft Hassrede, Desinformation oder (sexualisierte) Gewalt (Noble 2018; Benkler/Faris/Roberts 2018). Diese Mechanismen führen dazu, dass diskriminierende Inhalte amplifiziert und marginalisierte Stimmen unsichtbar gemacht werden. Hier zeigt sich, was Safiya Noble als „algorithms of oppression“ beschreibt: KI-Systeme sind keine

neutralen Werkzeuge, sondern tief eingebettet in gesellschaftliche Herrschaftsstrukturen.

Digitale Gewalt äußert sich in vielfältigen Formen: Cybermobbing, Stalking, Doxxing, bildbasierte sexualisierte Gewalt oder Hasskampagnen gegen Aktivist*innen. Carolin Emcke (2016) beschrieb Hassrede als „Gatekeeper“: Wer sich äußert, muss mit Angriffen rechnen, und wer betroffen ist, bleibt oft stumm. Dadurch entstehen Strukturen, die ohnehin schon marginalisierte Stimmen verdrängen. Eine Vielzahl an (weitgehend nichtrepräsentativen) Studien bestätigen diese Dynamik: Amnesty International konnte für (zu der Zeit noch) Twitter schon im Jahr 2018 nachweisen, dass Women of Color (Schwarze, asiatische, lateinamerikanische oder Frauen mit gemischtem ethnischen Hintergrund) mit einer um 34 Prozent höheren Wahrscheinlichkeit in beleidigenden, verletzenden oder feindseligen Inhalten erwähnt wurden als *weiße* Frauen. Schwarze Frauen waren dabei überproportional betroffen: Sie wurden mit einer um 84 Prozent höheren Wahrscheinlichkeit als *weiße* Frauen in feindseligen Tweets erwähnt (Amnesty International 2018). Auch Plan International (2020) zeigte, dass junge Frauen weltweit im Netz wegen ihres Geschlechts zur Zielscheibe werden – insbesondere, wenn sie zusätzlich PoC, queer oder be_hindert sind. Eine kanadische Erhebung von 2025 dokumentierte, dass 61 Prozent der Frauen und gender-diversen Personen geschlechtsspezifisch motivierte Gewalt im Netz erlebt haben; über die Hälfte reduzierte daraufhin ihre Online-Präsenz oder zensierte sich selbst (Canadian Women’s Foundation 2025). Ähnliche Zahlen finden sich in der Studie *Lauter Hass, leiser Rückzug* des Kompetenznetzwerks gegen Hass im Netz (2024): So berichten 49 Prozent der Befragten, bereits online beleidigt worden zu sein, und fast jede*r Zweite zieht sich als Folge von Hasskommentaren aus digitalen Debatten zurück. Junge Frauen, Menschen mit Migrationshintergrund oder queere Personen sind besonders betroffen – wobei Frauen dabei überdurchschnittlich oft sexualisierte Gewalt im Internet erleben. Diese Studie verdeutlicht zudem, dass Hass im Netz nicht nur individuelles Leid verursacht, sondern auch die demokratische Öffentlichkeit bedroht: 57 Prozent äußern aus Angst vor Angriffen ihre politische Meinung online seltener, 82 Prozent sehen die Meinungsvielfalt im Internet gefährdet. Digitale Gewalt, das wird deutlich, ist nicht nur individuell schädigend, sondern wirkt strukturell: Sie entzieht bestimmten Gruppen die Möglichkeit, gleichberechtigt an digitaler Öffentlichkeit und demokratischen Diskursen teilzunehmen.

Die technischen Grundlagen sozialer Medien verstärken diese Dynamiken. Inhalte, die Empörung und Hass erzeugen, werden priorisiert, während differenzierte oder marginalisierte Perspektiven eher untergehen (Milli et al. 2025). Diese Algorithmen sind auf Engagement-Maximierung optimiert, nicht auf das Wohlergehen von Nutzer*innen. Inhalte, die Empörung oder Polarisierung bedienen, werden bevorzugt ausgespielt, und das bedeutet konkret: antifeministische, rassistische, queerfeindliche Narrative erhalten mehr Sichtbarkeit als sachliche, differenzierte Gegenreden. Diese Logik verstärkt das Risiko für Betroffene digitaler Gewalt und wirkt wie ein Verstärker für strukturell ohnehin marginalisierte Stimmen, die aus dem digitalen Diskurs gedrängt werden (Weale 2024). Feministische oder antirassistische Beiträge werden dagegen häufiger fälschlich entfernt oder herabgestuft; das zentrale Problem liegt hier in den (teil-)automatisierten Moderationssystemen, die Plattformen zur Erkennung rechtswidriger oder „problematischer“ Inhalte einsetzen. Diese Systeme arbeiten oft auf der Basis sprachlicher und kontextueller Muster – Muster, die wiederum auf historischen Datensätzen beruhen, in denen marginalisierte Sprachen und Perspektiven unterrepräsentiert sind. Insbesondere Sprachen des sogenannten Globalen Südens oder queere Codesprache werden schlecht erfasst, was zu ungleichem Schutz führt (Shahid/Elsawah/Vashistha 2025). Inhalte aus feministischer, queerer oder antirassistischer Perspektive werden deshalb überproportional häufig als „unangemessen“, „hasserfüllt“ oder „extremistisch“ eingestuft, während sexistische, transfeindliche oder rassistische Aussagen algorithmisch durchrutschen (Golunova 2025).

So geraten Betroffene von digitaler Gewalt in einen absurden Zustand: Ihre Strategien der Gegenwehr und Selbstermächtigung werden zensiert, während die Gewalt, gegen die sie sich richten, online verbleibt, oft

unter dem Deckmantel von „Meinungsfreiheit“. Auch diese algorithmische Moderation ist mithin keineswegs neutral, sondern reproduziert Herrschaftsverhältnisse und übt epistemische Gewalt aus, indem sie marginalisiertes Wissen nahezu unsichtbar macht. Die so propagierte Meinungsfreiheit führt so zu faktischer Zensur.

Parallel dazu wirkt Desinformation als systematischer Angriff auf Öffentlichkeiten. Gender-spezifische Desinformation nutzt Geschlechterstereotype aus und zielt oft darauf ab, Einzelpersonen, insbesondere Frauen und queere Personen, aus öffentlichen Räumen zu verdrängen, die öffentliche Meinung zu beeinflussen, die Legitimität von Personen oder Institutionen zu untergraben und damit auch demokratische Teilhabe zu behindern. Diese Desinformation kann das Vertrauen schwächen, zur Selbstzensur zwingen und die Sicherheit von Frauen online und offline gefährden (Bradshaw 2024). Misogynie dient oftmals als „Einstiegspforte“ für Radikalisierung: Online-Communities wie Incels (Involuntary Celibate) oder die „Manosphere“ⁱ verbreiten frauenfeindliche Narrative, die in Kombination mit Verschwörungsmythen und Rassismus in Gewalt umschlagen können (Agena/Rahner 2021). Antifeminismus und Queerfeindlichkeit bilden weitere Ankerpunkte, an denen öffentliche Teilhabe eingeschränkt wird; schon seit einigen Jahren beobachten wir eine zunehmende Radikalisierung in der traditionellen Verschränkung von Misogynie, Antifeminismus, Antisemitismus und Rassismus, basierend auf reaktionären, konservativen Wertekontexten und geschlossenen, nicht selten verschwörungsmythologisch aufgeladenen Weltbildern. Alltagssexismus in Form von Sprache, Memesⁱⁱ und Herabwürdigungen, verstärkt in und durch verschwörungsmythologische(n) Gruppen, ist nicht selten das erste Eingangstor zu einer massiven Radikalisierung breiter Gruppen von Gesellschaft und kann bis zu menschenfeindlichen Attentaten und Femiziden reichen (Mosene 2021). Das bestätigt auch der Policy Brief der OSZE aus 2022, der aufzeigt, dass ein wachsender Bestand an Forschungsergebnissen darauf hinweisen, dass sich gewalttätig-misogyne Einstellungen und Haltungen, die gewalttätigen Extremismus und Radikalisierung unterstützen, stark überschneiden (OSCE 2022).

Die materielle Architektur der Voreingenommenheit

Im Kontext der Inhaltenmoderation wird die Verbindung zwischen KI und Materialität zunehmend sichtbar: Ihre Leistungsfähigkeit hängt nicht nur von Rechenleistung und Datenspeicherung ab, sondern auch von unsichtbarer menschlicher Arbeit, die meist von bereits marginalisierten Gruppen geleistet wird. Die sozialen und ökonomischen Bedingungen dieser Arbeit spiegeln die globalen Machtasymmetrien wider, die bereits in den Trainingsdatensätzen angelegt sind. Gemeint ist damit, was Kate Crawford in *Atlas of AI* (2021: 8) prägnant zusammenfasst: „AI is neither artificial nor intelligent. It is made from natural resources, fuel, human labor, data, and classified by humans. It is a registry of power.“ KI ist demzufolge nicht nur Code, sondern ebenso Bergbau für seltene Erden im Globalen Süden, wasser- und energieintensive Rechenzentren, und die oft unsichtbare Arbeit von Millionen Menschen, die Daten sammeln, labeln oder Inhalte moderieren.

Besonders in der Datenannotation und Content Moderation, den Basisarbeiten für maschinelles Lernen, sind es prekär beschäftigte Mitarbeiter*innen im Globalen Süden, die einen unverzichtbaren, aber unsichtbaren Teil der KI-Wertschöpfungskette ausmachen. Die globale KI-Industrie folgt so denselben extraktiven Logiken wie andere kapitalistische und koloniale Strukturen: Wertschöpfung findet vor allem im Globalen Norden statt, während Risiken, Kosten und Belastungen insbesondere im Globalen Süden anfallen. Damit sind die Macht- und Abhängigkeitsverhältnisse, die KI in ihren Ergebnissen reproduziert, bereits tief in ihrer

Produktionsweise eingeschrieben (Çetin 2021).

Diese Entwicklungen verschärfen sich durch globale Ungleichheiten. Plattformen konzentrieren ihre Schutzmaßnahmen auf den Globalen Norden, während Nutzende im Globalen Süden mit wesentlich weniger Regulierung und schlechterer Moderation konfrontiert sind. Gerade dort, wo politische Repression und digitale Gewalt ineinandergreifen, bleiben Betroffene schutzlos.

Gleichzeitig entstehen Gegenbewegungen. Netzfeministische Hashtag-Kampagnen wie #MeToo, #NiUnaMenos oder #SayHerName zeigen, dass digitale Räume auch für Widerstand genutzt werden können. Sie erzeugen Sichtbarkeit, Solidarität und politische Wirkung – oft gegen massiven Widerstand, algorithmische Unsichtbarmachung und digitale Angriffe. In Argentinien führte #NiUnaMenosⁱⁱⁱ seit 2015 zu spürbarem gesellschaftlichem und politischem Wandel, in den USA forderte #SayHerName Sichtbarkeit für Schwarze Frauen im Kontext von Polizeigewalt. Diese Bewegungen zeigen: Digitale Räume sind umkämpft, sie können sowohl Ausschlussräume sein als auch Orte feministischer Gegenöffentlichkeit.

Politisch haben die EU mit dem Digital Services Act (DSA) und der KI-Verordnung (AI Act) Fortschritte erzielt, etwa durch Transparenzpflichten oder Risikoprüfungen (Kleemann/Hirsbrunner/Aden 2023). Doch aus feministischer Sicht bleiben Lücken: Der DSA adressiert strukturelle Macht- und Diskriminierungsverhältnisse nur am Rande. Es fehlt eine ernsthafte Auseinandersetzung mit epistemischer Gewalt – etwa, wenn feministische Inhalte fälschlich entfernt werden, während sexistische Posts durchrutschen.

Ein intersektionaler feministischer Blickwinkel zeigt auf, dass es technische, politische, zivilgesellschaftliche und individuelle Handlungsebenen braucht. Technisch erfordert es Transparenz, Diversität und Bias-Diagnostik in der KI-Entwicklung. Politisch sind inklusive und globale Regulierungen notwendig, die Macht- und Diskriminierungsverhältnisse ernst nehmen. Zivilgesellschaftlich gilt es, feministische Gegenöffentlichkeiten zu stärken und solidarische Netzwerke zu fördern. Individuell können Selbstschutzstrategien wichtig sein, doch ohne kollektive Solidarität bleibt der Schutz lückenhaft.

Die Verengung öffentlicher Diskursräume durch digitale Gewalt, algorithmische Verzerrungen und Desinformation bedroht die demokratische Öffentlichkeit fundamental. Sie ist Ausdruck bestehender Ungleichheiten, die sich im digitalen Raum technologisch verfestigen. Nur wenn feministische und intersektionale Perspektiven in Technologiegestaltung, Regulierung und Öffentlichkeit systematisch berücksichtigt werden, können digitale Räume plural, sicher und demokratisch werden.

Katharina Mosene ist Politikwissenschaftlerin. Sie setzt sich dafür ein, intersektionale feministische Ansätze im Bereich der Internet Governance zu integrieren. Ihr Fokus liegt auf der Identifizierung von tradierten Vorurteilen (Biases) im Bereich der Künstlichen Intelligenz sowie auf ethischen Fragen beim Einsatz von Algorithmen in Wirtschaft und Gesellschaft. Zudem erforscht sie Lösungsansätze zur Bekämpfung von digitaler Gewalt, Hassrede und Antifeminismus, mit dem Ziel, Chancengleichheit im digitalen Raum zu ermöglichen. Am Alexander von Humboldt Institut für Internet und Gesellschaft ist sie Co-Projektkoordinatorin und Wissenschaftliche Mitarbeiterin des Forschungsprojekts Human in the Loop. Am Leibniz-Institut für Medienforschung | Hans-Bredow-Institut ist sie zudem für strategische Forschungs- und Veranstaltungs-Kooperationen zuständig, außerdem ist sie Gleichstellungsbeauftragte des Instituts. Neben

ihrem freiberuflichen Engagement im Bereich Cyber Security, Digitale Bildung und kritische Medienbildung ist sie Gründungsmitglied von netzforma* e.V. – Verein für feministische Netzpolitik.

Literatur

Agena, Gesinde/Rahner, Judith 2021: Antifeminismus, gewaltbereiter Rechtsextremismus und Geschlecht, in: Heinrich-Böll-Stiftung vom 14.04.2021, <https://heimatkunde.boell.de/de/2021/04/14/antifeminismus-gewaltbereiter-rechtsextremismus-und-geschlecht>.

Amnesty International 2028: Toxic Twitter – A Toxic Place for Women, in: Amnesty International vom 21.03.2018, <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1>.

Benkler, Yochai/Faris, Robert/Roberts, Hal 2018: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics, Oxford.

Bradshaw, Samantha 2024: Disinformation and Identity-Based Violence Muscatine, <https://stanleycenter.org/wp-content/uploads/2024/10/Disinformation-and-Identity-Based-Violence-Bradshaw.pdf>.

Canadian Women's Foundation 2025: Challenging Gendered Digital Harm: Research Report on Impacts and Solutions to Digital Harm Facing Women, Gender-Diverse People, and Gender Equality Organizations, Toronto, https://canadianwomen.org/wp-content/uploads/2025/05/25-18_CWF_GenderedDigitalHarm_ENG_Mainreport_v14.pdf.

Çetin, Raziye Buse 2021: Weisheit des Nichtwissens und dekoloniale KI, in Gunda-Werner-Institut (Hrsg.): Dekolonisierung des Digitalen: Feminismus und intersektionale Technologie, <https://www.gwi-boell.de/de/2021/02/11/weisheit-des-nichtwissens-und-dekoloniale-ki>.

Crawford, Kate 2021: Atlas of AI, New Haven.

Crenshaw, Kimberlé 1989: Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, in: *The University of Chicago Legal Forum*, H. 1, S. 139-167.

Emcke, Carolin 2016: Gegen den Hass, Frankfurt am Main.

Golunova, Valentina 2025: A Misogynistic Glitch? A Feminist Critique of Algorithmic Content Moderation, in: *Law, Technology and Humans*, Jg. 7, H. 2, S. 40-52.

Kleemann, Steven/Hirsbrunner, Simon/Aden, Hartmut 2023: Fairness, Erklärbarkeit und Transparenz bei KI-Anwendungen im Sicherheitsbereich – ein unmögliches Unterfangen? In: *vorgänge. Zeitschrift für Bürgerrechte und Gesellschaftspolitik* Nr. 242 = Jg. 62, H. 2, S. 29-47.

Kompetenznetzwerk gegen Hass im Netz (Hrsg.) 2024: Lauter Hass – leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung, Berlin, https://kompetenznetzwerk-hass-im-netz.de/download_lauterhass.php.

Milli, Smitha et al. 2025: Engagement, user satisfaction, and the amplification of emotionally charged, out-group hostile content in social media ranking algorithms, in: *PNAS Nexus*, Jg. 4, H. 3, S. 1-10, <https://doi.org/10.1093/pnasnexus/pgaf062>.

Mosene, Katharina 2021: Antifeminismus und die Fortschreibung von Marginalisierungen in digitalen Räumen, in: *blog interdisziplinäre geschlechterforschung* vom 07.10.2021, <https://www.gender-blog.de/beitrag/antifeminismus-digitale-raeume/>.

Mosene, Katharina 2024: Ein Schritt vor, zwei zurück: Warum Künstliche Intelligenz derzeit vor allem die Vergangenheit vorhersagt, in: *HIIG Digital Society Blog* vom 15.10.2024, <https://www.hiig.de/warum-ki-derzeit-vor-allem-vergangenheit-vorhersagt/>.

Noble, Safiya Umoja 2018: Algorithms of Oppression. New York.

Organisation for Security and Co-operation in Europe (OSCE) 2022: The Linkages between Violent Misogyny and Violent Extremism and Radicalization that Lead to Terrorism: Policy Brief, Wien, <https://www.osce.org/files/f/documents/d/c/525297.pdf>.

Plan International 2020: Free to be Online. Girls' and young women's experiences of online harassment, Surrey, Woking, https://www.plan.de/fileadmin/website/05._Ueber_uns/Maedchenberichte/Maedchenbericht_2020/Free_to_be_online.

Shahid, Farhana/Elsawah, Mona/Vashistha, Aditya 2025: Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages, [Preprint] DOI: 10.48550/arXiv.2501.13836.

Weale, Sally 2024: Social media algorithms ‘amplifying misogynistic content’, in: *The Guardian* vom 06.02.2024, <https://www.theguardian.com/media/2024/feb/06/social-media-algorithms-amplifying-misogynistic-content>.

Anmerkungen

i Die Manosphere (deutsch: Mannosphäre) ist ein vorwiegend antifeministisches Netzwerk. Es umfasst verschiedene Internetforen und Blogs, in denen hegemoniale Männlichkeit im Fokus stehen.

ii Ein Meme ist ein digitaler Inhalt, der sich viral verbreitet und oft humorvoll oder satirisch ist. Es kann ein Bild, Video oder Text sein, der leicht kopiert und weiterverbreitet werden kann.

iii *Ni una menos* ist eine lateinamerikanische Frauenbewegung, die sich gegen Gewalt an Frauen und Femizide einsetzt.

<https://www.humanistische-union.de/publikationen/vorgaenge/vorg-253/publikation/verengung-oeffentlicher-diskursraeume-auf-social-media-durch-biases-in-ki-digitale-gewalt-und-desinformation-eine-intersektional-feministische-analyse/>

Abgerufen am: 16.08.2026