

Zwischen Brandherd und Feuerlöscher: Künstliche Intelligenz in digitalen Debatten und demokratischen Diskursen

Diana Rieger & Mario Haim

Das Internet hat sich von einem zunächst als egalisierend verstandenen Medium zu einem fragmentierten Raum mit sinkender Diskursqualität entwickelt. Wurde anfangs vor allem der Zugang zu Informationen und Partizipation betont, treten heute Hassrede, Polarisierung und ungleiche Teilhabe stärker in den Vordergrund. KI-Anwendungen verstärken diese Entwicklungen, indem sie nicht nur zur Kuratierung von Inhalten dienen, sondern zunehmend auch für die Generierung und Verbreitung mit herangezogen werden. Generative KI ermöglicht dabei Manipulationen, Deepfakes oder eine einfache und adaptive Verschleierung von beispielsweise extremistischen Inhalten. Dadurch erhöht sich der Anteil von sogenanntem Borderline-Content, der rechtlich schwer zu fassen, aber gesellschaftlich verhältnismäßig wirksam ist. Neben direkter Meinungsbeeinflussung birgt auch die Einbindung verzerrter Daten in das Training von Sprachmodellen langfristige Gefahren. Als Handlungsoptionen diskutieren wir sowohl direkte Eingriffe wie Inhaltsmoderation, Prebunking oder Downranking als auch indirekte Maßnahmen wie rechtliche Regulierung und die Stärkung demokratischer Aushandlungsprozesse. Der Beitrag plädiert für interdisziplinären Austausch und demokratisch legitimierte Rahmenbedingungen, um KI-gestützte digitale Öffentlichkeiten aktiv mitzugestalten.

Empfohlene Zitierung:

Rieger, Diana/Haim, Mario (2025). Zwischen Brandherd und Feuerlöscher: Künstliche Intelligenz in digitalen Debatten und demokratischen Diskursen. In: Institut für Demokratie und Zivilgesellschaft (Hg.). Wissen schafft Demokratie. Schwerpunkt Demokratiegefährdung online, Band 18. Jena, 36–51.

Schlagwörter:

Hass im Netz, Fake News, Radikalisierung, Online-Kommunikation, Internet



**„DIE KI-
GEFÄHRDUNGSPOTENZIALE
MACHEN ES NACH
UNSEREM DAFÜRHALTEN
UNBEDINGT NOTWENDIG,
DASS SICH DEMOKRATISCHE
GESELLSCHAFTEN AUF
LEGITIMATIONSPROZESSE
FÜR DAS INTERNET
VERSTÄNDIGEN, UM NICHT
NUR REAKTIV DURCH GESETZE
AUF UMWÄLZUNGEN ZU
REAGIEREN, SONDERN AUCH
PROAKTIV GESTALTEN ZU
KÖNNEN.“**

Diana Rieger & Mario Haim

Einleitung

Eine wissenschaftliche Beschäftigung mit dem Internet führte bisher meist zu einem Zwiespalt. Auf der einen Seite standen Befunde, die einen breiteren Zugang zu Informationen, eine multimodale und damit inklusivere Aufbereitung sowie die einfache Möglichkeit der Partizipation beschrieben (z. B. Papacharissi 2004). Das Internet könne also Katalysator eines konstruktiven und breiten Austauschs werden. Dem gegenüber standen Befunde, dass vor allem ressourcenstarke Akteur*innen online aktiv partizipieren und einen Großteil der Aufmerksamkeit auf sich ziehen (z. B. Hindman 2008). Insbesondere der Anonymität und der häufig fehlenden Strafverfolgung werden zudem Entwicklungen hin zu Hassrede, Polarisierung oder Radikalisierung zugeschrieben (Rieger et al. 2024). Dem Internet wurde dabei eine stagnierende oder gar destruktive Rolle zuteil. Deutschland und die Europäische Union (EU) nahmen alsdann international Vorreiterrollen für die Regulierung des Internets ein (z. B. Klaus 2022). Bis vor wenigen Jahren schien sich die Wissenschaft deshalb unter dem Strich auf eine eher egalisierende Wirkung des Internets – im Sinne eines Beitrages zu einer breiteren Demokratisierung – zu einigen.

Seit einiger Zeit lässt sich eine gegenläufige Entwicklung beobachten. Intermediäre weichen zunehmend davon ab, Strafverfolgung im Internet aktiv zu unterstützen. Stattdessen verkünden sie libertäre Ideale und stellen sich wieder einmal lautstark auf den Standpunkt, in der Regel nicht Urheber und demnach nicht verantwortlich für die Inhalte auf ihren Plattformen zu sein. Dieser Trend geht einher mit Änderungen großer Intermediäre, Faktenchecks wieder zugunsten sogenannter Community Notes zu streichen. Obgleich ein solches Vorgehen den Regularien der EU widersprechen dürfte (Farrand 2025), ist zu erwarten, dass sich derartige Entscheidungen maßgeblich auf das Klima in Online-Debatten auswirken (Haim und Neuberger 2022; Lynch 2016). Das Internet, so die aktuelle kommunikationswissenschaftliche Beobachtung, ist zu einem lauten und fragmentierten Umschlagplatz von Meinungen ohne allzu hohe Diskursqualität geworden. Diese Beobachtungen spiegeln sich beispielsweise in Forschung, die zeigt, dass durch fehlende Regulation oder Inhaltsmoderation mehr Hassrede auf Plattformen zu finden ist (Hickey et al. 2025; Rieger et al. 2021).

Das Internet, so die aktuelle kommunikationswissenschaftliche Beobachtung, ist zu einem lauten und fragmentierten Umschlagplatz von Meinungen ohne allzu hohe Diskursqualität geworden.

Diese Entwicklungen flankieren eine Umwälzung kommunikativer Grundannahmen, insofern Anwendungen Künstlicher Intelligenz (KI) den öffentlichen Online-Raum mitgestalten. Unter KI

verstehen wir ein Konstrukt, das technologische wie soziale Komponenten umfasst, um maschinell Aufgaben zu erledigen, die üblicherweise menschliche Intelligenz erfordern (vgl. Russell und Norwig 2020, 19). Für diesen Beitrag gehen wir von KI in manifestierter Form unterschiedlicher Anwendungen im Rahmen digitaler Debatten und demokratischer Diskurse aus, also den Austausch von politischen Informationen, der zur individuellen Meinungsbildung beiträgt. Einerseits geschieht das unmittelbar bei Diskussionen in sozialen Netzwerken, in den Kommentarbereichen traditioneller Massenmedien, mit modernen KI-Assistenzsystemen wie ChatGPT sowie in Messenger-Apps wie WhatsApp oder Telegram. Andererseits zählen wir die unterschiedlichen Möglichkeiten der KI-gestützten Auffindbarkeit politischer Informationen dazu, also etwa Suchmaschinen, Aggregatoren oder Filter- und Empfehlungsalgorithmen.

Kommunikative Umwälzungen bestehen nun darin, dass solche KI-unterstützten Anwendungen zuletzt Informationen im Internet primär kuratiert haben, nun jedoch verstärkt selbst in die Rollen von Rezipient*innen und Produzent*innen geschlüpft sind. KI wird gleichermaßen als Urheber, mögliche Lösung sowie als vermeintlich neutraler Intermediär dieser Entwicklungen gehandelt. Umso schwieriger gestalten sich gesellschaftliche Aushandlungsprozesse darüber, welche Rolle KI einnehmen kann, soll oder muss (Friemel und Neuberger 2023; Jungherr und Schroeder 2023). Der vorliegende Beitrag verfolgt daher zwei Ziele: Erstens sollen durch KI entstehende Gefährdungspotenziale für einen demokratischen Online-Diskurs diskutiert und eingeordnet werden. Zweitens sollen Handlungsoptionen zwischen passiver Reaktion und proaktivem Einsatz auf verschiedenen Einflussebenen für demokratische Prozesse skizziert werden.

KI wird gleichermaßen als Urheber, mögliche Lösung sowie als vermeintlich neutraler Intermediär dieser Entwicklungen gehandelt.

Gefährdungspotenziale

Durch die Masse an KI-unterstützten Anwendungen ist die Rolle von KI in Diskursen entsprechend vielfältig. Lag der Fokus bisher noch auf der Kuratierung von Informationen, etwa bei Suchmaschinen, so dient insbesondere generative KI heute der Produktion von Diskussionsbeiträgen. Das zeigt sich beispielsweise in der Manipulation von Wahlen durch KI-Unterstützung und der Generierung und Weiterverbreitung falscher oder intoleranter Inhalte. KI kann als Urheber (etwa von Deepfakes) oder Verstärker (etwa durch Social Bots oder Astroturfing) entsprechender Inhalte auftreten (García-Orosa 2022; Hajli et al. 2021). Auch für die Betonung visueller Charakteristika (etwa arischer Phänotypen) wird generative KI eingesetzt (Hiller und de las Casas 2025). Derart generierte und/oder verstärkte Inhalte passen durch die systematische Nutzung von Emotion, Überraschung oder Personalisierung üblicherweise hervorragend zu den Filter- und Empfehlungsmechanismen, nach

denen soziale Netzwerke oder Suchmaschinen Inhalte kuratieren, und verbreiten sich also umso besser – nicht nur durch KI-Akteure, sondern in weiterer Folge auch über menschliche und institutionelle Akteure (z. B. Nachrichtenmedien). So verhilft der Empfehlungsalgorithmus bei YouTube Verschwörungsinhalten und extremistischen Inhalten zu Aufmerksamkeit, selbst wenn die Ausgangsvideos zivilgesellschaftliche Aufklärungskampagnen sind (Schmitt et al. 2018; Zieringer und Rieger 2023; für TikTok, siehe Matlach et al. 2025).

Unter dem Begriff Mainstreaming wird diskutiert, wie radikale oder extremistische Akteure ihre Inhalte möglichst unauffällig online einem breiten Publikum zugänglich machen (ohne dabei gelöscht zu werden) (Rothut et al., 2024). KI kann hier eingesetzt werden, um problematische Inhalte zu generieren, sie zu verschleiern oder als Dog Whistles (=Anspielungen/Codes, die nur von der Ingroup erkannt und verstanden werden) zu nutzen. Die Einsatzmöglichkeiten reichen von der sprachlichen Tarnung extremer Inhalte (z. B. durch Paraphrasierung oder Codewörter) über multimodale Verschleierung (z. B. durch Einbettung relevanten Texts in Grafiken, Insider-Emojis, Botschaften auf der Tonspur bei unauffälligem visuellen Material) und Moderations-Evasionen (z. B. kleine maschinelle Änderungen, die Suchfilter in die Irre führen; LLM-Prompts, die scheinbar neutralen Kontext erstellen) bis hin zu Automatisierung und Reichweitenaufbau (wie z. B. die bereits erwähnten Social Bots und Deep-fakes, aber auch die Verteilung von ‚sauberem‘ Teaser-Inhalt mit Einladungen oder Weiterleitungen in schwerer moderierbare Räume). Erwartbar ist also mehr sogenannter Borderline-Content (Macdonald und Vaughan 2024) – Inhalt, der nicht ganz offen zu Hass aufruft, Desinformationen nur andeutet oder haarscharf an der Grenze zur Gewaltverherrlichung auftritt und somit in einer rechtlichen Grauzone verbleibt.

Unter dem Begriff Mainstreaming wird diskutiert, wie radikale oder extremistische Akteure ihre Inhalte möglichst unauffällig online einem breiten Publikum zugänglich machen (ohne dabei gelöscht zu werden).

Darüber hinaus ist KI in den vergangenen Jahren stärker in die Rolle Rezipierender gerückt, insofern sie Debatten aufnehmen und interpretieren soll. Hassrede und Desinformation kann so noch stärker und noch individualisierter auf Diskursverläufe zugeschnitten werden. Jüngere Studien zeigen Effekte auf ausgewählte Weltansichten, wenn in persuasiven Gesprächen KI individuell auf Argumente eingeht (z. B. Costello et al. 2024; Salvi et al. 2025).

Weitere Gefährdungspotenziale durch KI konnten wir bereits in der Vergangenheit beobachten, etwa beim Microtargeting im Rahmen der US-Wahl 2016 oder bei der algorithmisch kuratierten Isolation von Diskursen (Echokammern), die als ein Treiber gesellschaftlicher Polarisierung,

insbesondere bei bipolaren Entscheidungen, gelten (z. B. bei den Brexit-Kampagnen 2016). Auch die Debatte um als Fake News bezeichnete Imitate vertrauenswürdiger Nachrichtenmedien wurde bereits geführt, erhielt jedoch neuen Schub durch die sogenannte Doppelgänger-Kampagne. Ab 2022 tauchten dabei in westlichen Ländern rund um nationale Wahlen Dutzende Webseiten auf, die glaubwürdige Medien optisch nachahmten, zwischen gestohlenen Meldungen Desinformation platzierten und offensichtlich weitgehend per KI erstellt und betrieben wurden. Später wurde die Kampagne der russischen Social Design Agency zugeschrieben.

Relevant ist neben der unmittelbaren Meinungsbeeinflussung auch das langfristige Gefährdungspotenzial systematischer KI-Trainingsbeeinflussung. Moderne KI, insbesondere Large Language Models (LLMs), werden mit riesigen Mengen internetbasierter und möglichst glaubwürdiger Quellen trainiert, die jedoch Verzerrungen wie Stereotype, einseitige Normen oder ungleiche Repräsentationen enthalten (vgl. Gallegos et al. 2024). Da viele Modelle auf ähnlichen

Relevant ist neben der unmittelbaren Meinungsbeeinflussung auch das langfristige Gefährdungspotenzial systematischer KI-Trainingsbeeinflussung. Moderne KI werden mit riesigen Mengen internetbasierter und möglichst glaubwürdiger Quellen trainiert, die jedoch Verzerrungen enthalten.

Datensätzen beruhen, erfordert es unternehmerischen Mut, neue Quellen zu erschließen. Umso schwerer ist es, massenhaft glaubwürdig wirkende, aber desinformationsgefüllte Lookalikes auszusortieren. Gelangen sie ins Training, können Modelle deren Narrative übernehmen, verbreiten und etwa zur Unterdrückung von Minderheiten beitragen.

Diese Vielschichtigkeit von KI, positive wie negative Wirkpotenziale entfalten zu können, und die zahlreichen Ebenen, auf denen KI Einfluss auf Diskurse nehmen kann, machen den Austausch über einen angemessenen Umgang sehr schwierig. Nicht zuletzt deshalb offenbaren sich Unzulänglichkeiten gesellschaftlicher Aushandlungsprozesse, die eher ein Re-Agieren – anstelle eines proaktiven Ansatzes – vorgeben (vgl. Frischlich et al. 2017).

Handlungsoptionen

Wir sehen zunächst die Notwendigkeit, den Handlungsspielraum im Umgang mit KI einzuteilen. Für die Frage, welchen Gegenmaßnahmen KI dienen *kann*, verweisen wir auf aktuelle(re) Fachliteratur, um nicht in einen Duktus technologischen Determinismus abzuschweifen (siehe u. a. Battista und Mangone 2025; Cupać et al. 2024; Jungherr 2023; Katzenbach 2021; Paltieli 2023). Stattdessen legen wir den Fokus auf mögliche Handlungsoptionen, unabhängig von ihrer Implementierung. Auf der

einen Seite stehen konkrete Handlungsoptionen, die direkten Einfluss auf Gefährdungspotenziale von KI in Debatten und Diskursen nehmen können und sollen. Auf der anderen Seite sehen wir nachhaltigere Handlungsoptionen, die einen entsprechenden Einfluss nur vorzeichnen und indirekt negative Konsequenzen von KI adressieren. Diese Optionen verstehen wir perspektivisch gar als Muss im Umgang mit den unterschiedlichen Gefährdungspotenzialen von KI.

Betrachten wir die Handlungsoptionen mit direktem Einfluss lässt sich feststellen: Die Forschung hat sich viel mit der automatisierten Erkennung von Beiträgen beschäftigt, die spezifizierte Charakteristika erfüllen, beispielsweise KI-generierte Einzelbeiträge, die mutmaßlich bestimmte Straftatbestände (z. B. Aufruf zu Gewalt) erfüllen. Ebenfalls hierunter fallen KI-orchestrierte Beitragswellen, für die zahlreiche Social-Media-Posts mit bestimmtem Falschinformationen verbreitet werden. Diese Handlungsoptionen gleichen indes einem Katz-und-Maus-Spiel, bei dem ständig Möglichkeiten ausgelotet werden müssen, um neuen Gefährdungspotenzialen entgegenwirken zu können und nicht zuletzt Fragen der dafür notwendigen Kompetenz an entsprechenden Stellen zu bedenken sind.

Aus unserer Sicht mindestens genauso wichtig ist die normative Auslegung derartig direkter Handlungsoptionen: Was soll unterbunden werden? Die Aufrechterhaltung von Vielfalt wird bereits im Kontext der Plattformregulierung als Leitidee gesehen (Schneiders et al. 2024). Ginge es schlicht um die Einhaltung bereits geltender Gesetze, so würde das wohl die automatisierte Identifikation von Gewaltaufrufen

Aus unserer Sicht mindestens genauso wichtig ist die normative Auslegung derartig direkter Handlungsoptionen: Was soll unterbunden werden?

legitimieren – wie sonst sollen Intermediäre derartigen Ansprüchen bei den immensen Datenaufkommen nachkommen. Doch die möglichen Nebenwirkungen, die sich aus so einer Auslagerung der Tatbestandserkennung an Algorithmen ergeben, sind kaum zu unterschätzen. Verzerrungen in den Trainingsdaten würden für uneinheitliche Detektion sorgen. Mithilfe von KI wäre es ein Leichtes, derartige Systeme durch die vorher bereits genannten Strategien zu umgehen. Auch deshalb hat die EU bereits Intermediäre als Betreiber relevanter Plattformen in die Pflicht genommen und nicht nur die Erkennung, sondern auch die Berichterstattung darüber verbindlich implementiert (Digitale-Dienste-Gesetz, ehemals Netzwerkdurchsetzungsgesetz). Man muss die Wirksamkeit der aktuellen Umsetzung jedoch infrage stellen, verpflichtet sie letztlich ja nur zu einem Reporting über die Anzahl und Geschwindigkeit detektierter Fälle. Was aber ist, beispielsweise, mit KI-orchestrierter Meinungsmache? Kein Gesetz sieht vor, dass bestimmte Meinungen nicht zum Ausdruck gebracht werden dürfen. Ein solches Gesetz, das in Teilen wohl einer Zensur gleichkäme, wäre in falschen Händen mehr als nur besorgniserregend. Ein juristischer Mittelweg ist hierbei schmal (z.B. Wachter et al.

2021). Zahlreiche Studien aber zeigen, dass (auch nicht strafbare) Meinungsäußerungen in Form von Hassrede die Offenheit und Diskursteilhabe verringern und so Pluralität in öffentlichen Debatten reduzieren (Rieger et al. 2024; Rothut et al. 2023; Schulze et al. 2024). Während die Befundlage also deutliche Hinweise für oder gegen bestimmte Maßnahmen erlauben würde, herrscht vor allen Dingen Uneinigkeit über den politischen Willen. Um zielgerichteter Handlungsoptionen ausloten zu können, bedarf es mehr normativen Konsenses – und dafür wiederum mehr normativen Diskurses.

Während die Befundlage deutliche Hinweise für oder gegen bestimmte Maßnahmen erlauben würde, herrscht vor allen Dingen Uneinigkeit über den politischen Willen. Um zielgerichteter Handlungsoptionen ausloten zu können, bedarf es mehr normativen Konsenses.

Hierzu wird häufig diskutiert, inwieweit implizite Hassrede (Rieger et al. 2021), Furchtrede (Greipl et al. 2024) oder Borderline-Content sich verbreiten kann, wenn gängige KI-Systeme vor allem auf die (einfachere) Detektion von ‚klassischer‘ expliziter Hassrede ausgelegt sind (Macdonald und Vaughan 2023). Solche Systeme sind weniger in der Lage, die Multimodalität von Hass wie auch verdeckte Formen (z. B. negative Stereotypisierungen) zu erkennen. Gleichzeitig geben Befunde Anlass zur Vermutung, dass vor allem Borderline-Content langfristig zu einer Normalisierung beitragen kann (Rothut et al. 2024; Saha et al. 2023). Entsprechend wird der Einsatz von KI vor allem zur Inhaltsmoderation diskutiert. Dies kann durch Chat-Assistenten umgesetzt werden, die politische Vielfalt im Diskurs aufrechterhalten sollen (Argyle et al. 2023), persuasiv gegen bestimmte Ansichten (hier: Verschwörungsglauben) argumentieren (Costello et al. 2024) oder als Mediator in Gruppen mit divergierenden Ansichten fungieren (Tessler et al. 2024).

Eine häufiger diskutierte Handlungsoption ist das sogenannte Prebunking, bei dem Resilienz durch a priori verbreitete Hinweise auf kursierende Falschinformationen erzeugt wird (Lewandowsky und Van der Linden 2021). Durch die Exposition mit abgeschwächten Formen von beispielsweise Verschwörungsinhalten oder Desinformation (oder ihrer Strukturen) soll diese Präventivmaßnahme die Widerstandsfähigkeit von Nutzenden stärken, die noch nicht mit Falschinformationen konfrontiert wurden (Roozenbeek et al. 2021). Die Funktion einer ‚Impfung‘ (Inokulation; siehe Compton 2025; Compton und Braddock 2025) erfüllt dabei eine KI. Ein bekanntes Pionierprojekt ist Google’s Jigsaw Redirect Kampagne, die beim Aufruf von problematischen YouTube-Videos zuvor mögliche Gegenbotschaften setzt. Ähnlich können auch weitere Prebunking-Werkzeuge zum Einsatz kommen, die gegen Falschinformationen in alternativen Nachrichten, Verschwörungsinhalten oder extremistischen Inhalten eingesetzt werden (Biddlestone et al. 2025).

Im Gegensatz dazu legen indirekte Handlungsoptionen einen Umgang nahe, der perspektivisch ausgewählten Gefährdungspotenzialen entgegenwirken soll. Gesetze sind sicherlich ein probates Mittel, das insbesondere in Deutschland und der EU zunehmend spezialisierte Form annimmt. Zu nennen sind die Datenschutzgrundverordnung, das Netzwerkdurchsetzungsgesetz, der Medienstaatsvertrag, der Data Act, der AI Act oder der sich abzeichnende Media Act. Zentral ist, dass nicht Gesetze entwickelt, sondern auch deren Durchsetzung entsprechend mitgeplant wird, was nach unserem Dafürhalten in den genannten Beispielen durchaus der Fall und somit begrüßenswert ist.

Darüber hinaus fehlt es dem modernen Internet vor allem an legitimierten Aushandlungsprozessen (Haim und Neuberger 2022). Öffentliche Infrastruktur bedarf demokratisch legitimierter Teilhabe, um Akzeptanz und letztlich auch Vertrauen aufbauen zu können. Derartige Legitimation entsteht üblicherweise entweder aus Ernennung und Zustimmung (Input-Legitimation, Beispiel: gewähltes Europäisches Parlament), aus der Nützlichkeit generierter Leistungen der Ernannten (Output-Legitimation, Beispiel: Europäische

Öffentliche Infrastruktur bedarf demokratisch legitimierter Teilhabe, um Akzeptanz und letztlich auch Vertrauen aufbauen zu können.

Zentralbank) oder durch den Rahmen eines bereits legitimierten Prozesses (Throughput-Legitimation, Beispiel: Volksabstimmungen). Während das Internet als Ganzes keine dieser Definitionen erfüllt (Deuze und McQuail 2020), sind dennoch positive Beispiele ersichtlich. Wikipedia etwa hat einen vertrauensvollen epistemischen Prozess geschaffen, bei dem offene Teilhabe und Transparenz als Grundlage seiner Legitimation dienen. Kommt es zu Unstimmigkeiten (vgl. Frost-Arnold 2018), werden institutionalisierte Diskursräume auf Community-Seiten und Konferenzen für einen Austausch darüber, nicht aber über die Inhalte genutzt. So hat sich die Legitimation des Prozesses stets verstärkt, wovon Wikipedia als Produkt profitiert. Andere Online-Beispiele, denen ein legitimierter Prozess zugeschrieben werden kann, sind dezentrale Plattform-Infrastrukturen wie das Fediverse (z. B. Mastodon), Open-Source-Communitys (z. B. Mozilla) oder beteiligungsorientierte Demokratieplattformen (z. B. Petitionssammlungen).

Wenngleich das moderne Internet als Ganzes zum überwiegenden Teil aus kommerziellen Angeboten besteht und viel Macht bei wenigen sehr großen Akteuren liegt, lassen sich aus unserer Sicht einige Aspekte der genannten Positiv-Beispiele für demokratisch legitimierte Aushandlungsprozesse aufgreifen. Mehr noch: Die KI-Gefährdungspotenziale machen es nach unserem Dafürhalten unbedingt notwendig, dass sich demokratische Gesellschaften auf Legitimationsprozesse für das Internet verständigen, um nicht nur reaktiv durch Gesetze auf Umwälzungen zu reagieren, sondern auch proaktiv gestalten zu können.

Derartige Aushandlungsprozesse könnten an gesellschaftliche Ideale einer normativ wünschenswerten Öffentlichkeit anknüpfen. Deliberative Ideale etwa legen einen konsensorientierten und rationalen Diskurs nahe, in dem Journalismus als vermittelnde Kraft auftritt, wohingegen liberale Ideale Öffentlichkeit verstärkt als Resonanzboden unterschiedlicher Perspektiven verstehen, auf dem der Journalismus Orientierung bieten soll (vgl. Donges und Jarren 2022; Ferree et al. 2002; Martinsen 2009). Auf solchen normativen Idealen aufbauend ließen sich sodann Präventionsstrategien entwickeln, um einen inklusiven, egalitären oder vielfältigen Diskurs zu forcieren.

Eine Präventionsstrategie besteht darin, das Design der Plattformen zu verändern, etwa durch Friction (Naderer et al. in Druck). Friction-Elemente sind eingesetzte Gestaltungsmerkmale, die das Online-Verhalten der Nutzenden verlangsamen und dadurch die kognitive Reflexion fördern. Beispiele sind das aufwändige Kündigung von Zeitschriften-Abos (bei welchem mehrere Aktionen nötig sind), das wiederholte Nachfragen, ob eine Datei wirklich gelöscht werden soll oder die Limitierung von bestimmten Aktionen an einem Tag (z. B. Swipes bei Tinder). Sie erhöhen die Kontrolle der Nutzenden über Inhalte, denen sie begegnen wollen, reduzieren so mögliches impulsives Verhalten, bieten Lernmöglichkeiten und können für schädliche Inhaltsformen sensibilisieren (Ingegno 2023).

Eine Präventionsstrategie besteht darin, das Design der Plattformen zu verändern, etwa durch Friction. Friction-Elemente sind eingesetzte Gestaltungsmerkmale, die das Online-Verhalten der Nutzenden verlangsamen und dadurch die kognitive Reflexion fördern.

Eine andere Strategie sind algorithmische Eingriffe durch Technologieunternehmen, um Sichtbarkeit und Reichweite problematischer Inhalte zu verringern (sogenannte Downranking/Algorithmic Demotion) (Macdonald und Vaughan 2024; Urman et al. 2024). Anstelle einer vollständigen Entfernung, die Fragen im Zusammenhang mit der Meinungsfreiheit aufwerfen kann (Kozyreva et al. 2023), reduzieren Downranking-Strategien die algorithmische Verstärkung bestimmter Beiträge (z. B. mit Kombinationen von Schlüsselwörtern, die polarisierte Reaktionen zeigen) oder Konten (z. B. die algorithmische Funktion des Friendings) und schränken so deren Verbreitung ein, ohne sie direkt zu löschen. Hier könnte auch Shadowbanning ein KI-gesteuertes Werkzeug sein, welches Inhalte sanktioniert, die konträr zu deliberativen Normen stehen (Gillespie 2022). Während Shadowbanning die verdeckte Einschränkung von geposteten Inhalten eines Nutzenden beschreibt (mit der Folge eingeschränkter Sichtbarkeit), meint Downranking die Abwertung von Beiträgen durch den Feed-Algorithmus (mit der Folge sinkender Reichweite).

Beide Strategien unterstreichen erneut die notwendige demokratische Legitimation von Prozessen, in denen konzeptuelle Klarheit hergestellt werden muss, was eigentlich zu problematischem oder Borderline-Inhalt zählt und wer unter welchen Prämissen dazu legitimiert ist, die Sichtbarkeit oder Verbreitung von Inhalten einzuschränken. MacDonald und Vaughan (2023) plädieren daher für Transparenz-Praktiken in der Inhaltsmoderation. Hierbei finden sich zudem Anknüpfungspunkte an den Forschungsstrang der sogenannten explainable AI, der sich mit den Möglichkeiten beschäftigt, KI in die Lage zu versetzen, getroffene Entscheidungen nachvollziehbar zu begründen (z. B. Hoffman et al. 2023).

Ausblick

Wir beobachten das Internet derzeit als lauten und fragmentierten Umschlagplatz von Meinungen ohne allzu hohe Diskursqualität, der durch KI auf unterschiedlichen Ebenen grundlegenden Umwälzungsprozessen unterliegt. Orchestriert von Akteur*innen mit unterschiedlichsten Intentionen geriert sich KI als Urheber und Verstärker von Hass, Desinformation, Verschwörungserzählungen und, zunehmend, vielen Inhalten, die sich nicht eindeutig derartigen Kategorien zuordnen lassen. Gegenmaßnahmen werden vielfach beforscht, zuletzt fällt deren Einsatz aber schwerer. Das liegt einerseits an den mächtigen Intermediären, die ihre aktive Unterstützung der Strafverfolgung im Lichte politischer Verschiebungen auf ein Minimum reduzieren. Andererseits hadern Gesetzgeber mit den genannten inhaltlichen Veränderungen sowie einem fehlenden normativen Konsens darüber, wie eigentlich ein wünschenswerter Online-Diskursraum aussehen könnte.

Die Formierung eines öffentlichen Raumes stellt eine ständige Gratwanderung dar. Was Öffentlichkeit leisten soll, darf keine rein juristische Fragestellung sein. Sie darf auch nicht rein wissenschaftlich beantwortet werden und schon gar nicht darf sie dem Gusto der Regierenden obliegen. Derzeit geben insbesondere kommerzielle Interessen die Rahmung dieser Öffentlichkeit vor, was durchaus als im Einklang mit der Logik mancher Mediensysteme verstanden werden kann.

Die Formierung eines öffentlichen Raumes stellt eine ständige Gratwanderung dar. Was Öffentlichkeit leisten soll, darf keine rein juristische Fragestellung sein. Sie darf auch nicht rein wissenschaftlich beantwortet werden und schon gar nicht darf sie dem Gusto der Regierenden obliegen.

Für Deutschland und die EU gilt das indes nur begrenzt. Hier ist der Bedarf nach (1) demokratisch legitimierten Arenen zur Aushandlung angemessener Rahmenbedingungen groß. Um diese zu schaffen, bieten sich (2) moderne demokratische Beteiligungsverfahren, etwa digitale Bürgerräte

(z. B. Dienel et al. 2024) an. Es erscheint außerdem naheliegend, (3) europäische Intermediäre von entsprechender Relevanz zu fördern, wobei (4) ein steter Austausch unterschiedlicher Stakeholder (u. a. Recht, Wissenschaft, Intermediäre, Politik, Zivilgesellschaft) von Bedeutung ist. Auch (5) die Idee, ein öffentlich-rechtliches soziales Netzwerk zu schaffen, wurde bereits geäußert (z. B. Zuckerman 2020). Schließlich plädieren wir aus wissenschaftlicher Perspektive für (6) mehr interdisziplinären Austausch, um beispielsweise informatische, psychologische, juristische und sozialwissenschaftliche Perspektiven regelmäßiger unter einen Hut zu bekommen. Dafür bieten sich passende Förderformate an, wie sie unlängst im Rahmen der deutschen Internetinstitute (bidt, CAIS, Weizenbaum) aufkamen. Europa kann hier Maßstäbe setzen – mit klaren Werten, mutigen Strukturen und dem Willen zum demokratischen Aushandlungsprozess. Die digitale Öffentlichkeit der Zukunft ist vor allen Dingen eine gesellschaftliche Frage.

Diana Rieger ist Professorin für Kommunikationswissenschaft am Institut für Kommunikationswissenschaft und Medienforschung an der Ludwig-Maximilians-Universität München. In mehreren Drittmittelprojekten (z. B. MOTRA, RadiGame, gefördert von BMFTR, BMI und BMBFSFJ) beschäftigt sie sich mit der Erforschung von Indikatoren einer diskursiven Online-Radikalisierung, der Ausgestaltung und Wirkung von toxischen Online-Inhalten (ToxicAIment, gefördert vom bidt und Dis-Ident, gefördert von BMFTR) sowie mit möglichen (digitalen) Gegenstrategien, bzw. digitalen Informations- und Hilfsangeboten (TATE, gefördert von der EU).

Mario Haim ist Professor für Kommunikationswissenschaft am Institut für Kommunikationswissenschaft und Medienforschung an der Ludwig-Maximilians-Universität München. Seine Forschungsinteressen konzentrieren sich auf algorithmische Einflüsse, beispielsweise in der politischen Kommunikation, im Journalismus, in der Gesundheitskommunikation oder in der Mediennutzung, sowie auf (computergestützte) Methoden und Metawissenschaft. Weitere Informationen: <https://haim.it/>.

Literaturverzeichnis

- Argyle, Lisa P./Bail, Christopher A./Busby, Ethan C./Gubler, Joshua R./Howe, Thomas/Rytting, Christopher/Sorensen, Taylor/Wingate, David (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences of the United States of America* 120 (41), 1–8. <https://doi.org/10.1073/pnas.2311627120>.
- Battista, Daniele/Mangone, Emiliana (2025). Technological Culture and Politics: Artificial Intelligence as the New Frontier of Political Communication. *Societies* 15 (4), 1–15. <https://doi.org/10.3390/soc15040075>.

- Biddlestone, Mikey/Roozenbeek, Jon/Suiter, Jane/Culloty, Eileen/van Der Linden, Sander (2025). Tune in to the prebunking network! Development and validation of six inoculation videos that prebunk manipulation tactics and logical fallacies in misinformation. *Political Psychology*, 1–29. <https://doi.org/10.1111/pops.70015>.
- Compton, Josh (2025). Inoculation theory. *Review of Communication* 25 (1), 1–13. <https://doi.org/10.1080/15358593.2024.2370373>.
- Compton, Josh/Braddock, Kurt (2025). Inoculation theory and conspiracy, radicalization, and violent extremism. In: Sergei A. Samoilenko/Solon Simmons (Hg.). *The Handbook of Social and Political Conflict*. 1. Aufl. Wiley, 405–413. <https://doi.org/10.1002/9781119895534.ch36>.
- Costello, Thomas H./Pennycook, Gordon/Rand, David G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385 (6714). <https://doi.org/10.1126/science.adq1814>.
- Cupać, Jelena/Schopmans, Hendrik/Tuncer-Ebetürk, İrem (2024). Democratization in the age of artificial intelligence: introduction to the special issue. *Democratization* 31 (5), 899–921. <https://doi.org/10.1080/13510347.2024.2338852>.
- Deuze, Mark/McQuail, Denis (2020). *Media & Mass Communication Theory*. 7th Edition. Sage.
- Dienel, Hans-Liudger/von Blanckenburg, Christine/Bach, Nicolas (2024). Mini Publics Online: Erfahrungen mit Online-Bürgerräten und Online-Planungszellen. In: Norbert Kersting/Jörg Radtke/Sigrid Baringhorst (Hg.). *Handbuch Digitalisierung und politische Beteiligung*. Wiesbaden, Springer VS. https://doi.org/10.1007/978-3-658-31480-4_37-1.
- Donges, Patrick/Jarren, Otfried (2022). *Politische Kommunikation in der Mediengesellschaft. Eine Einführung. Studienbücher zur Kommunikations- und Medienwissenschaft*. 5. Aufl. Wiesbaden, Springer VS.
- Farrand, Benjamin (2025). Online platforms, intermediary responsibility, and human rights: Digital copyright as a site of multiple contestations in the EU. In: Ben Wagner/Matthias C. Kettemann/Kilian Vieth-Ditlmann/Susanah Montgomery (Hg.). *Research Handbook on Human Rights and Digital Technology*. Elgar, 54–68. <https://doi.org/10.4337/9781035308514.00010>.
- Ferree, Myra Marx/Gamson, William A./Gerhards, Jürgen/Rucht, Dieter (2002). Four models of the public sphere in modern democracies. *Theory and Society* 31 (3), 289–324. <https://doi.org/10.1023/A:1016284431021>.
- Friemel, Thomas N./Neuberger, Christoph (2023). The public sphere as a dynamic network. *Communication Theory* 33 (2–3), 92–101. <https://doi.org/10.1093/ct/qtad003>.
- Frischlich, Lena/Rieger, Diana/Morten, Anna/Bente, Gary in Kooperation mit der Forschungsstelle Terrorismus / Extremismus des Bundeskriminalamtes (2017). *Videos gegen Extremismus? „Counter-Narrative“ auf dem Prüfstand*. Wiesbaden, Bundeskriminalamt.
- Frissen, Thomas/d’Haenens, Leen/Opgenhaffen, Michael (2021). Extreme right and mis/disinformation. In: Howard Tumber/Silvio Waisbord (Hg.). *The Routledge Companion to Media Disinformation and Populism*. 1st ed. London, Routledge, 268–278. <https://doi.org/10.4324/9781003004431-29>.
- Frost-Arnold, Karen (2018). Wikipedia. In: David Coady/James Chase (Hg.). *The Routledge Handbook of Applied Epistemology*. 1st ed. London, Routledge, 28–40. <https://doi.org/10.4324/9781315679099-7>.
- Gallegos, Isabel O./Rossi, Ryan A./Barrow, Joe/Tanjim, Mehrab M./Kim, Sungchul/Dernoncourt, Franck/Yu, Tu/Zhang, Ruiyi/Ahmed, Nasreen K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50 (3), 1097–1179. https://doi.org/10.1162/coli_a_00524.
- García-Orosa, Berta (2022). Digital Political Communication: Hybrid Intelligence, Algorithms, Automation and Disinformation in the Fourth Wave. In: Berta García-Orosa (Hg.). *Digital Political Communication Strategies. Multidisciplinary Reflections*, Palgrave Macmillan Cham, 3–23. https://doi.org/10.1007/978-3-030-81568-4_1.
- Gillespie, Tarleton (2022). Do not recommend? Reduction as a form of content moderation. *Social Media+ Society* 8 (3). <https://doi.org/10.1177/20563051221117552>.
- Greipl, Simon/Hohner, Julian/Schulze, Heidi/Schwabl, Patrick/Rieger, Diana (2024). “You are doomed!” Crisis-specific and dynamic use of fear speech in protest and extremist radical social movements. *Journal of Quantitative*

- Description: Digital Media 4, 1-46. <https://doi.org/10.51685/jqd.2024.icwsm.8>.
- Haim, Mario/Neuberger, Christoph (2022). The paradox of knowing more and less: Audience metrics and the erosion of epistemic standards on the internet. *Studies in Communication and Media* 11 (4), 566–589. <https://doi.org/10.5771/2192-4007-2022-4-566>.
- Hajli, Nick/Saeed, Usman/Tajvidi, Mina/Shirazi, Farid (2021). Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence. *British Journal of Management* 33 (3), 1238–1253. <https://doi.org/10.1111/1467-8551.12554>.
- Hickey, Daniel/Fessler, Daniel M./Lerman, Kristina/Burghardt, Keith (2025). X under Musk's leadership: Substantial hate and no reduction in inauthentic activity. *PLoS one* 20 (2). <https://doi.org/10.1371/journal.pone.0313293>.
- Hindman, Matthew (2008). *The myth of digital democracy*. Princeton, University Press.
- Hoffman, Robert R./Mueller, Shane T./Klein, Gary/Litman, Jordan (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5. <https://doi.org/10.3389/fcomp.2023.1096257>.
- Ingegno, Massimo (2023). Friction in design: The good, the bad, and the.. Dark. LinkedIn – Make It Toolkit. Online verfügbar unter <https://www.linkedin.com/pulse/friction-design-good-bad-dark-make-it-toolkit/> (abgerufen am 18.09.2025).
- Jungherr, Andreas (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society* 9 (3). <https://doi.org/10.1177/20563051231186353>.
- Jungherr, Andreas/Schroeder, Ralph (2023). Artificial intelligence and the public arena. *Communication Theory* 33 (2–3), 164–173. <https://doi.org/10.1093/ct/qtad006>.
- Katzenbach, Christian (2021). „AI will fix this“ – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society* 8 (2). <https://doi.org/10.1177/20539517211046182>.
- Klausa, Torben (2022). Graduating from 'new-school' – Germany's procedural approach to regulating online discourse. *Information, Communication & Society* 26 (1), 54–69. <https://doi.org/10.1080/1369118X.2021.2020321>.
- Lewandowsky, Stephan/Van Der Linden, Sander (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* 32 (2), 348–384. <https://doi.org/10.1080/10463283.2021.1876983>.
- Macdonald, Stuart/Vaughan, Katy (2024). Moderating borderline content while respecting fundamental values. *Policy & Internet* 16 (2), 347–361. <https://doi.org/10.1002/poi3.376>.
- Martinsen, Renate (2009). Öffentlichkeit als „Mediendemokratie“ aus der Perspektive konkurrierender Demokratietheorien. In: Frank Marcinkowski/Barbara Pfetsch (Hg.). *Politik in der Mediendemokratie*. Wiesbaden, Springer VS, 37–69.
- Matlach, Paula-Charlotte/Castillo, Allison/Drath, Charlotte/Hevesi, Eva F. (2025). Recommending hate: How TikTok's search engine algorithms reproduce societal bias. Institute for Strategic Dialogue. Online verfügbar unter [How-TikToks-Search-Engine-Algorithms-Reproduce-Societal-Bias.pdf](https://www.instituteforstrategicdialogue.com/wp-content/uploads/2025/09/How-TikTok-s-Search-Engine-Algorithms-Reproduce-Societal-Bias.pdf) (abgerufen am 18.09.2025).
- Naderer, Brigitte/Rothut, Sophia/Rieger, Diana (in Druck). Preventing and countering online radicalization in the digital age: Actors, strategies and challenges. In: Iwan Awan/Pelham Carter (Hg.). *International handbook on counter-radicalization*. De Gruyter.
- Lynch, Michael Patrick (2016). *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York, Liveright Publishing.
- Paltieli, Guy (2023). Re-imagining democracy: AI's challenge to political theory. In: Simon Lindgren (Hg.). *Handbook of Critical Studies of Artificial Intelligence*. Cheltenham, Edward Elgar Publishing, 333–342. <https://doi.org/10.4337/9781803928562.00036>.
- Papacharissi, Zizi (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6 (2), 259–283. <https://doi.org/10.1177/1461444804041444>.

- Rieger, Diana/Greipl, Simon/Schmid, Ursula K./Hohner, Julian/Schulze Heidi (2024). Hassrede als Merkmal von (Online-) Radikalisierung. In: Tobias Rothmund & Eva Walther (Hg.). *Psychologie der Rechtsradikalisierung. Theorien, Perspektiven, Prävention*. Stuttgart, Kohlhammer, 125–134.
- Rieger, Diana/Kümpel, Anna Sophie/Wich, Maximilian/Kiening, Toni/Groh, Georg (2021). Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media + Society* 1–14. <https://doi.org/10.1177/20563051211052906>.
- Rothut, Sophia/Sacher, Anna-Luisa/Strohmeier, Rebecca/Reinemann, Carsten (2023). Meinungsfreiheit in Gefahr? Wie politische Einstellungen und individuelle Erfahrungen die Wahrnehmung der Meinungsfreiheit in Deutschland prägen. *Studies in Communication and Media* 12 (1), 48–91. <https://doi.org/10.5771/2192-4007-2023-1-48>.
- Rothut, Sophia/Schulze, Heidi/Rieger, Diana/Naderer, Brigitte (2024). Mainstreaming as a meta-process: A systematic review and conceptual model of factors contributing to the mainstreaming of radical and extremist positions. *Communication Theory* 34 (2), 49–59. <https://doi.org/10.1093/ct/qtae001>.
- Rozenbeek, Jon/Maertens, Rakoel/McClanahan, William/Van Der Linden, Sander (2021). Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited. *Educational and Psychological Measurement* 81 (2), 340–362. <https://doi.org/10.1177/0013164420940378>.
- Russell, Stuart/Norvig, Peter (2020). *Artificial Intelligence: A Modern Approach*. 4. Aufl. Harlow, Pearson.
- Saha, Punyajoy/Garimella, Kiran/Kalyan, Narla Komal/Pandey, Saurabh Kumar/Meher, Pauras Mangesh/Mathew, Binny/Mukherjee, Animesh (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences* 120 (11). <https://doi.org/10.1073/pnas.2212270120>.
- Salvi, Francesco/Horta Ribeiro, Manoel/Gallotti, Riccardo/West, Robert (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behavior* 9, 1645–1653. <https://doi.org/10.1038/s41562-025-02194-6>.
- Schmitt, Josephine B./Rieger, Diana/Rutkowski, Olivia/Ernst, Julian (2018). Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube Recommendation Algorithms. *Journal of Communication* 68 (4), 780–808. <https://doi.org/10.1093/joc/jqy029>.
- Schneiders, Pascal/Stegmann, Daniel/Stark, Birgit/Zieringer, Lisa/Reinemann, Carsten (2024). Meinungsmacht unter der Lupe: Ein Ansatz für eine vielfaltssichernde, holistische Plattformregulierung. In: Marlis Prinzing/Josef Seethaler/Mark Eisenegger/Patrik Ettinger (Hg.). *Regulierung, Governance und Medienethik in der digitalen Gesellschaft*. Wiesbaden, Springer Fachmedien, 97–120.
- Schulze, Heidi/Greipl, Simon/Hohner, Julian/Rieger, Diana (2024). Social media and Radicalization: An Affordance Approach for Cross-Platform Comparison. *M&K Medien & Kommunikationswissenschaft* 72 (2), 187–212. <https://doi.org/10.5771/1615-634X-2024-2-187>.
- Tessler, Michael Henry/Bakker, Michiel A./Jarrett, Daniel/Sheahan, Hannah/Chadwick, Martin J./Koster, Raphael/Evans, Georgina/Campbell-Gillingham, Lucy/Collins, Tantum/Parkes, David C./Botvinick, Matthew/Summerfield, Christopher (2024). AI can help humans find common ground in democratic deliberation. *Science* 386 (6719). <https://doi.org/10.1126/science.adq2852>.
- Wachter, Sandra/Mittelstadt, Brent/Russell, Chris (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41. <https://doi.org/10.1016/j.clsr.2021.105567>.
- Zieringer, Lisa/Rieger, Diana (2023). Algorithmic Recommendations' Role for the Interrelatedness of Counter-Messages and Polluted Content on YouTube – A Network Analysis. *Computational Communication Research* 5 (1), 109–140. <https://doi.org/10.5117/CCR2023.1.005.ZIER>.
- Zuckerman, Ethan (2020). *The Case for Digital Public Infrastructure*. Knight First Amendment Institute at Columbia University. Online verfügbar unter <https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure> (abgerufen am 18.09.2025).



**„WIR BEOBACHTEN
DAS INTERNET DERZEIT
ALS LAUTEN UND
FRAGMENTIERTEN
UMSCHLAGPLATZ VON
MEINUNGEN OHNE ALLZU
HOHE DISKURSQUALITÄT,
DER DURCH KI AUF
UNTERSCHIEDLICHEN
EBENEN GRUNDLEGENDEN
UMWÄLZUNGSPROZESSEN
UNTERLIEGT.“**

Diana Rieger & Mario Haim