Kommunikations- und Medienethik

Stapf | Heesen | Martena [Hrsg.]

Zwischen Hassrede, Framing und generativer Künstlicher Intelligenz

Medien und Sprache aus ethischer Perspektive

Nomos

Kommunikations- und Medienethik

Herausgegeben von
Alexander Filipović
Christian Schicha
Ingrid Stapf

Band 24

Ingrid Stapf | Jessica Heesen | Laura Martena [Hrsg.]

Zwischen Hassrede, Framing und generativer Künstlicher Intelligenz

Medien und Sprache aus ethischer Perspektive

Nomos

Die in diesem Sammelband präsentierten Beiträge durchliefen ein Peer-Review-Verfahren zur Qualitätssicherung und wurden von unabhängigen Fachvertreter:innen begutachtet.

Für die finanzielle Unterstützung, die die Bereitstellung dieses Bandes open access ermöglicht hat, bedanken wir uns bei der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft (DGPuK), der Universität Greifswald, der Hochschule Mittweida und der Hochschule Bonn-Rhein-Sieg.

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2026

Publiziert von
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden
www.nomos.de

Gesamtherstellung:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden

ISBN (Print): 978-3-7560-4020-9
ISBN (ePDF): 978-3-7489-7032-3

DOI: https://doi.org/10.5571/9783748970323

Onlineversion
Inlibra

Inhaltsverzeichnis

II. Automatisierte Sprache: Potenziale und Risiken Künstlicher Intelligenz

III. Öffentliche Sprache: Ethische Perspektiven auf demokratische Debattenkultur

Einleitung

Laura Martena / Ingrid Stapf / Jessica Heesen

Der Ausdruck „Kommunikation" stammt von „communio" ab, was „mitteilen", aber auch „teilen", „gemeinsam machen" oder gar „vereinigen" heißen kann. Kommunikation, im Wesentlichen sprachlich verfasste, ermöglicht einerseits den Ausdruck des Eigenen, andererseits Differenz und Pluralität, und vermag idealiter gerade darin Gemeinsamkeit zu schaffen. Damit wird sie zum wesentlichen Medium moralischer Aushandlungsprozesse und demokratischer Verständigung.

Im demokratischen Diskurs selbst wurde Sprache lange als vermeintlich neutrales Werkzeug behandelt, das Inhalte transportiert, ohne selbst Thema zu werden. In den vergangenen Jahren hat sich dies jedoch gewandelt. Der Sprachgebrauch ist heute selbst zu einem zentralen Gegenstand moralischer und politischer Konflikte geworden. Immer neue Debatten um Formulierungen, die früher als unproblematisch galten, jetzt aber als diskriminierend wahrgenommen werden und deshalb aus dem Wortschatz gestrichen werden sollen, oder auch die Forderungen nach geschlechtergerechter Sprache verweisen darauf. Gleichzeitig versuchen insbesondere rechtsextreme Akteure, oft unter Berufung auf die Redefreiheit, gezielt in den privaten und öffentlichen Sprachgebrauch zu intervenieren. Im Rahmen ihrer „metapolitischen" Strategien, die sie ursprünglich linken Kontexten entnommen haben, arbeiten sie darauf hin, diskreditierte Begriffe zu rehabilitieren oder neue zu etablieren, um Wahrnehmungsbereitschaften und Bewertungsmaßstäbe im Sinne ihrer eigenen politischen Agenda zu verschieben (vgl. Ruppert-Karakas 2024). Diese gegenläufigen Dynamiken – die Sensibilisierung für diskriminierende und verletzende Sprache einerseits, ihre gezielte ideologische Instrumentalisierung andererseits – lenken den Blick auf neue Weise auf die wirklichkeitskonstituierende Macht der Sprache. Sprache wird in dieser Sicht zum politischen Kampfplatz.

Diese Entwicklungen vollziehen sich in einem medialen Umfeld, das sich im Zuge der Digitalisierung tiefgreifend verändert hat. Soziale Medien, die anfänglich mit hohen Erwartungen an eine Demokratisierung des Diskurses verbunden waren, weil jeder nun zugleich Sender:in und Empfänger:in sein können sollte, sind Arenen der Polarisierung geworden.

Hassrede, Beleidigungen und Diffamierungen prägen die kommunikative Praxis vieler Plattformen. Obwohl Dienste wie Instagram, TikTok oder X durch ihr Content-Management maßgeblich mitbestimmen, welche Inhalte sichtbar werden, verstehen sie sich selbst meist als neutrale Vermittler und weisen Verantwortung von sich. Die bislang ohnehin nicht sehr effektiven Mechanismen der Moderation und Regulierung sind angesichts der schieren Menge an Inhalten prekär geblieben und politisch umstritten: Was die einen als Voraussetzung für faire Diskurse begreifen, sehen andere gerade als Eingriff in die Meinungsfreiheit, der den demokratischen Diskurs untergrabe. In den USA werden Moderationen öffentlichen Sprachgebrauchs unter Donald Trump derzeit zunehmend wieder abgeschafft (vgl. Debatin in diesem Band), während die Europäische Union mit dem *Digital Services Act* die Funktionen einer demokratischen Öffentlichkeit stützen will.

In dieses Spannungsfeld ist mit Generativer Künstlicher Intelligenz eine neue, potenziell disruptive Kraft getreten. Einerseits versprechen Große Sprachmodelle enorme Potenziale für journalistische Arbeit und politische Kommunikation. Im professionellen Medienbereich werden sie bereits jetzt verwendet, um z. B. Wetter-, Börsen- oder Sportmeldungen zu verfassen. Auch in Film, Kunst und Musik ist Generative Künstliche Intelligenz in Produktionen integriert und Teil öffentlich zugänglicher Inhalte, ebenso in Form virtueller Influencer in den Sozialen Medien. Neben den Potenzialen werden seit einigen Jahren auch die Gefahren und Möglichkeiten des Missbrauchs von Künstlicher Intelligenz diskutiert. So neigen Modelle, die synthetische Inhalte generieren, zu sogenannten Halluzinationen, das heißt dazu, überzeugend formulierte Ergebnisse auszugeben, die objektiv gesehen aber falsch sind. Auch wenn dies ohne Absicht geschieht, können diese Misinformationen (vgl. Aïmeur/Amri/Brassard 2023) die Vertrauenswürdigkeit öffentlicher Kommunikation gefährden. Generative Künstliche Intelligenz wird zudem gezielt für Desinformation missbraucht. Diese lässt sich unter Verwendung entsprechender Tools schneller in großer Zahl erzeugen und leichter personalisiert verbreiten. Vor allem im Umfeld von Wahlen und insgesamt im Zusammenhang der öffentlichen und individuellen Meinungsbildung entfalten Deepfakes als Desinformation von besonders hoher Suggestions- und Täuschungskraft zunehmend an Wirkung (vgl. Muñoz 2024). Insbesondere dann, wenn künstlich generierte Mis- oder Desinformationen wiederum durch KI-Anwendungen verwendet werden, kann ein gefährlicher Kreislauf entstehen, der nur noch schwer durch menschliche Aufsicht kontrolliert und korrigiert werden kann.

Angesichts dieser Phänomene lässt sich von einer multiplen Krise demokratischer Debattenkultur sprechen. Sie betrifft die symbolischen Grundlagen politischer Verständigung, die institutionellen Bedingungen öffentlicher Kommunikation und die technischen Infrastrukturen gleichermaßen. Die beschriebene Krise ist zugleich die Stunde der Kommunikations- und Medienethik. Ethik ist immer auch „Krisenreflexion" (Riedel 1979). In diesem Sinn übt die Kommunikations- und Medienethik Kritik an bestehenden Kommunikationsordnungen und ihren Infrastrukturen, indem sie deren Legitimität, Angemessenheit und demokratische Tragfähigkeit befragt. Zugleich gibt Ethik Orientierung für die Praxis (Debatin 1997). Die Kommunikations- und Medienethik macht konkret normative Orientierungsangebote für diejenigen, die in Medien und Öffentlichkeit handeln – etwa in Gestalt von Kodizes, Qualitätsstandards oder berufsethischen Leitlinien. So kann Sie Maßstäbe geben, an denen sich die Gesellschaft und gesellschaftliche Verständigung in ihrer Diversität ausrichten kann. In der gegenwärtigen Konstellation kann die Kommunikations- und Medienethik einen Beitrag zur Sicherung und Erneuerung demokratischer Öffentlichkeit(en) leisten – und wird damit selbst zu einer wichtigen Stimme in der Auseinandersetzung um die Bedingungen verantwortlicher Kommunikation.

Der vorliegende Band versteht sich als ein solcher Beitrag. Er widmet sich dem Verhältnis von Medien und Sprache aus ethischer Perspektive und versammelt theoretische, normative und empirische Studien. Gemeinsam verfolgen sie das Ziel, sprachbezogene Konflikte, digitale Kommunikationspraktiken und KI-gestützte Informationsprozesse in ihren ethischen Implikationen sichtbar zu machen und Orientierungen für eine demokratische Kommunikationskultur zu entwickeln. Zusammen geben die Beiträge eine medienethische Bestandsaufnahme einer Gegenwart, in der Sprache nicht nur Mittel, sondern Gegenstand öffentlicher Konflikte ist, digitale Infrastrukturen die Kommunikation prägen und neue Technologien die Schaffung geteilter Welten herausfordern.

Teil 1: Die Beiträge sind drei Themenschwerpunkten zugeordnet. Die Autor:innen des ersten Teiles über *umstrittene Sprache* widmen sich Sprache als Medium und Gegenstand ethischer Auseinandersetzungen anhand konkreter Konfliktfelder. Ein erster Schwerpunkt liegt auf Hassrede. Während dieser Begriff im öffentlichen Diskurs in den vergangenen Jahren allgegenwärtig geworden ist, empirische Studien ihre Verbreitung aus Sicht Betroffener nahelegen und Gegenmaßnahmen gefordert werden, wird seltener gefragt, was „Hassrede" genauer ist. Das führt zu einer uneinheitli-

chen Verwendung, die auch die Einordnung entsprechender Studienergebnisse und die Evaluation von Maßnahmen erschwert. Dieser Frage widmet sich Inga Bones anhand einer kritischen Diskussion bestehender Begriffsbestimmungen. Sie kommt zu dem Schluss, dass Hassrede am besten als Form der Kommunikation zu verstehen ist, die eine Person oder Gruppe auf Basis stabiler und identitätsrelevanter Eigenschaften wie Religion, Ethnie, Geschlecht oder Herkunft auf tiefgreifende Weise angreift, abwertet oder diskriminiert.

Bernhard Debatin zeigt auf, welcher Strategien sich rechtsextreme Akteure in den USA bedienen, um Hassrede zu verbreiten. Besonders beliebt sei das sogenannte *Dog Whistling*. Durch kodierte und suggestive Aussagen werde dabei Hass auf Angehörige bestimmter Gruppen bei der eigenen Klientel weiter geschürt. Außenstehenden erschließe sich deren rassistischer und gewaltverherrlichender Gehalt dagegen nicht unmittelbar, sodass dieser sich im Zweifel leicht verleugnen und der Rassismusvorwurf gegen diejenigen kehren lasse, die ihn erheben. Dabei bedienten sich die entsprechenden Akteur:innen auch Generativer KI, um ihre Hassbotschaften weiter zu verbreiten.

Jörg-Uwe Nieland untersucht das Verhältnis von Sprache, Emotion und Konflikt in der organisierten Sportfankultur. Durch die Digitalisierung verlagern sich auch hier Auseinandersetzungen rivalisierender Fangruppen in digitale Arenen. Anhand einer Pilotstudie zum österreichischen Eishockey zeichnet der Autor nach, wie Fankommunikation hier stattfindet. Dabei lasse sich unter den Fans eine grundlegend agonistische Online-Kommunikation beobachten, die von Rivalität, Abwertung des Gegners und Aufwertung des eigenen Vereins und seiner Anhänger:innen gekennzeichnet sei. Bei alledem bleibe die Streitkultur aber zivilisiert. Dehumanisierende Feindbilder und Hassrede träten auf den Instagram-Kanälen der Vereine nicht offen zutage. Ob sich die entsprechende Kommunikation in andere Kanäle verlagert hat, bleibt offen.

Miriam Goetz und Vanessa Wimmers beleuchten spezifisch sexistische Hate Speech im Sinne geschlechtsspezifischer Diskriminierung, die vor allem Frauen – hier exemplarisch bei Instagram – betrifft. Userinnen berichteten in hohem Maße von Beleidigungen, sexualisierter Belästigung sowie Vergewaltigungs- und Morddrohungen in Kommentaren und Direktnachrichten. Instagram-spezifische Funktionen wie Direktnachrichten und die niedrige Zugangsschwelle zur Kontaktaufnahme begünstigten gezielte Angriffe. Dabei wirkten algorithmische Verstärkungsmechanismen und gruppendynamische Effekte normalisierend auf sexistische Inhalte. Mel-

de- und Moderationssysteme würden von den Nutzerinnen als ineffektiv wahrgenommen. Entsprechend würden härtere Sanktionen, bessere Filter, menschliche Moderation, geringere Anonymität und rechtliche Durchsetzung gefordert.

Wie gehen Medien mit der Politisierung der Sprache um? Dieser Frage widmet sich Beatrice Dernbach am Beispiel des Genderns. Sie rekonstruiert, wie unterschiedliche deutschsprachige Redaktionen sich im Streit um gendersensible Sprache positionieren – von konsequenter Nutzung des Genderns mit Sonderzeichen über pragmatische, eher zurückhaltende Lösungen bis zu klarer Ablehnung von Sonderzeichen und Betonung des generischen Maskulinums. Die Verfasserin zeigt, welche normativen Vorstellungen der Sprache und der eigenen gesellschaftlichen Rolle dahinterstehen, um schließlich für eine Versachlichung der Debatte zu plädieren. Dabei diskutiert sie auch, inwiefern der wachsende Einsatz Generativer Künstlicher Intelligenz in den Redaktionen zu einer solchen Versachlichung beitragen oder aber die Polarisierung weiter verstärken könne.

Teil 2: Auf die Analyse aktueller Sprachkonflikte folgt ein Blick auf *automatisierte Sprache,* und damit die technologischen Veränderungen, die die Produktion von Sprache selbst transformieren. Der zweite Teil des Bandes befasst sich mit ethischen Fragen, die speziell mit dem Aufstieg Generativer Künstlicher Intelligenz, insbesondere Großen Sprachmodellen verbunden sind. Er beleuchtet diese Problemstellungen auf der Ebene des öffentlichen Diskurses über KI, auf der gesamtgesellschaftlichen Ebene, hinsichtlich ihres Einsatzes in Redaktionen und auf der Mikroebene der Trainingsdaten. Und er fragt, wie KI-Anwendungen ihrerseits helfen können, einige dieser Probleme zu lösen.

Zunächst werfen Brigitte Huber und Julia Levasier Schlaglichter auf aktuelle Online-Diskussionen über KI im deutschsprachigen Raum seit der Veröffentlichung von ChatGPT. Anhand einer Untersuchung der Diskurse auf der Nachrichtenwebsite und dem Instagram-Account von *Die Zeit* zeigen sie, dass diese von ethischen Bedenken und Szenarien des Kontrollverlusts geprägt sind. Dazu zählt etwa die der Science Fiction entstammende Vorstellung, KI könne ein Bewusstsein entwickeln, aber auch die handfeste Sorge vor Diskriminierung und mangelndem Datenschutz. Aus wissenssoziologischer Sicht zeichnen sie nach, dass die Nutzer:innen dabei vor allem auf Erfahrungswissen und populärkulturelle Referenzen zurückgriffen, während wissenschaftliches Wissen selten rezipiert werde. Zugleich neigten die Nutzer:innen dazu, die Wissensansprüche anderer in Zweifel zu ziehen.

Michael Litschka stellt die Ergebnisse einer Studie zum aktuellen Einsatz von KI-Anwendungen, insbesondere Sprachmodellen, in österreichischen und internationalen Medienunternehmen aus Sicht der Akteur:innen dar. Da diese Erhebung neben ökonomischen Gesichtspunkten auch ethische Bedenken und ein Bedürfnis nach Regulierung offenbart, geht er im zweiten Schritt der Frage nach Möglichkeiten und Grenzen solcher Regulierungsversuche und ihrer ethischen Fundierung nach. Anhand der Gerechtigkeitstheorien von John Rawls und Amartya Sen liefert er den Befürworter:innen von Regulierung eine medienethische Begründung und erweitert die bestehenden KI-ethischen Diskurse um eine seltener eingenommene Perspektive.

Matthias Rath beleuchtet eine systematische Gefahr Generativer KI, die im öffentlichen Diskurs weniger präsent ist: eine „artifizielle Sepsis". Mit dieser Metapher beschreibt er, wie der unregulierte Einsatz solcher Tools zu einer epistemischen „Selbstvergiftung" führen könne. Diese drohe dann, wenn Generative KI-Modelle beginnen, wiederum maschinell erzeugte Inhalte, die schon heute einen Großteil der online verfügbaren Inhalte ausmachen, in ihre Such- und Trainingsprozesse einzuspeisen. Dadurch würde sich die Informationsökologie schrittweise von ihrer Begründung in realweltlicher Erfahrung entkoppeln. Als Gegenmodell schlägt der Autor eine „salutogenetische Bot-Nutzung" vor, die epistemische Resilienz stärken soll.

Die Gefahr „artifizieller Sepsis" verweist auf die zentrale Bedeutung der Qualität von Trainingsdaten, der sich die nächsten beiden Beiträge widmen. Wenn Jana Hecktor Qualitätskriterien von Trainingsdaten für KI-Anwendungen im Allgemeinen vorstellt und diese speziell auf LLMs anwendet, bezieht sie nicht nur die Auswahl bestehender Datensätze ein, sondern auch den Prozess der Datengewinnung – einschließlich der Rechte derjenigen Menschen, die bewusst oder unbewusst daran mitarbeiten. Konkret rückt sie Ausgewogenheit, Datenschutz, Diversität, Fairness, Korrektheit, Privatsphäre und Repräsentativität in den Fokus und buchstabiert aus, was damit im Kontext von LLMs verbunden wäre. Dabei könnten die Kriterien im Einzelfall durchaus in Widerspruch geraten und sollten fallbezogen gegeneinander abgewogen werden. Die Anwendung solcher Kriterien könnte helfen, offensichtlicher Diskriminierung durch Generative KI-Tools vorzubeugen.

Gibt es aber nicht auch subtilere, verdecktere und daher schwerer detektierbare Formen geschlechtsbezogener, rassistischer oder ableistischer Diskriminierung und Stereotypisierung? Wenn ja, wie ließen diese sich im

Maschinellen Lernen aufdecken und vermeiden? Anregungen findet Theresa Krampe in zeitgenössischen Kunstprojekten, die sich Generativer KI auf kreative und subversive Weise bedienen. Solche Projekte, die mit diversifizierten und sorgfältig kuratierten Datensätzen arbeiten, könnten nicht nur hegemoniale Vorstellungen irritieren, die unserem aktuellen Umgang mit KI oft zugrunde liegen, und *biases* zu minimieren helfen. Sie könnten auch dazu inspirieren, alternative und fairere KI-Zukünfte zu imaginieren.

Können automatisierte Systeme auch helfen, die Probleme zu lösen, die sie selbst mit erzeugen oder verstärken? Dieser Frage widmen sich Mario Anastasiadis und Hektor Haarkötter am Beispiel von Desinformation. Deren enorme Verbreitung im Netz droht die öffentliche Meinungsbildung zu verzerren, Diskursräume zu destabilisieren und kann das Vertrauen in demokratische Institutionen beschädigen. Um dagegen anzugehen, wurde die App NEBULA entwickelt, die Desinformation automatisch detektieren soll. Die Autoren stellen Ergebnisse der medienethischen Begleitforschung zur Entwicklung dieser App vor. NEBULA soll nicht nur Hinweise auf Desinformation ausgeben, sondern die Fähigkeit stärken, diese eigenständig zu erkennen.

Teil 3: Die Beiträge des dritten Teils über *öffentliche Sprache* weiten den Blick, indem sie die Frage nach den normativen Grundlagen demokratischer Streitkultur in Zeiten der Digitalisierung und des Aufstiegs Generativer Künstlicher Intelligenz neu stellen.

Petra Grimm, Susanne Kuhnert und Marcel Schlegel berichten von zwei Forschungs- und Lehrprojekten, die sich um die Auswirkungen Generativer KI auf das Mediensystem drehen. Im Fokus stehen Anwendungen zur artifiziellen Nachbildung menschlicher Stimmen. Ein erstes Projekt richtete sich an Medienstudierende, die anhand neuer didaktischer Szenarien, insbesondere realitätsnaher Fallstudien, ethische Aspekte des Umgangs damit abzuwägen lernen sollten. Ein zweites Projekt nahm erfahrene Medienpraktiker:innen in den Blick. Anhand qualitativer Interviews wurden zunächst Leitlinien zur KI-Nutzung erarbeitet, die Medienschaffenden helfen sollten, die ethische Legitimität eines spezifischen KI-Einsatzes zu überprüfen.

Christian Schicha stellt die Bedeutung der verfassungsrechtlich garantierten Meinungs- und Kommunikationsfreiheit in den Mittelpunkt. Ausgehend von Jürgen Habermas skizziert er ein Konzept deliberativer Öffentlichkeit, zu der alle Bürger:innen gleichberechtigt Zugang haben und in der Argumente statt Machtmittel zählen. Der digitale Strukturwandel drohe dieses Ideal zu untergraben, indem er die Entstehung fragmentierter, vermachteter und algorithmisch gesteuerter Teil- und Gegenöffentlichkeiten fördere. Dagegen

setzt der Autor das Ideal einer an Solidarität, Empathie, Respekt und wechselseitiger Anerkennung orientierten Streitkultur, die allein demokratische Selbstbestimmung und Gemeinwohlorientierung sichere.

Stefanie Averbeck-Lietz begründet den Pluralismus als Wert aus kommunikationsethischer Perspektive und verankert ihn in einem Universalismus der Würde. Mit Seyla Benhabib skizziert sie einen „interaktiven Universalismus", in dem moralischer Universalismus und die Anerkennung der Andersheit der Anderen sich nicht länger ausschließen, sondern wechselseitig bedingen und hervorbringen. Auch sie knüpft an Jürgen Habermas an, wenn sie diesem Gedanken eine kommunikationstheoretische Wendung gibt. Schließlich unterscheidet sie mit Ernst Manheim verschiedene Konzeptionen der Öffentlichkeit, die dieses Ideal ermöglichen oder untergraben. Der Beitrag skizziert die Aufgabe einer kommunikationsethischen Bildung, die Anerkennung von Verschiedenheit, Sensibilität für verletzte Geltungsansprüche und Widerstandskraft stärkt – auch mit Blick auf KI-gestützte Kommunikationsumgebungen.

Wie können konstruktive Debattenräume im Digitalen geschaffen werden, wenn kommerzielle Plattformen einzig an Gewinnmaximierung durch höhere Verweildauer der Nutzer:innen interessiert sind und dazu polarisierende Inhalte bevorzugen? Was haben die öffentlich-rechtlichen Medien dem entgegenzusetzen? Henning Eichler stellt mit dem *Public Spaces Incubator* (PSI) ein internationales Forschungs- und Entwicklungsprojekt vor, das die Sender wenn nicht ganz aus der Abhängigkeit von kommerziellen Plattformen lösen, so doch neue Räume für werteorientierte, konstruktive und respektvolle Online-Debatten eröffnen könne. Damit wollen sie ihrem Funktionsauftrag auch angesichts veränderter Erwartungen des Publikums an den Dialog mit Journalist:innen wieder gerecht werden. Eichler stellt die Ergebnisse einer qualitativen Studie zu ethischen Herausforderungen dar, die sich aus Sicht von Journalist:innen und anderen Beteiligten aus der bisherigen Arbeit in PSI-Debattenräumen ergeben, und ordnet die Befunde medienethisch ein.

Janis Brinkmann geht allgemeiner der Frage nach, wie sich journalistische Qualität unter Bedingungen eines „entgrenzten" und ausdifferenzierten Journalismus neu bestimmen lässt. Ausgangspunkt ist die Diagnose, dass klassische, am Informations- und Nachrichtenjournalismus orientierte Qualitätsbegriffe gegenüber alternativen Berichterstattungsmustern wie etwa dem investigativen, narrativen oder konstruktiven Journalismus an Grenzen stoßen. Auf öffentlichkeitstheoretischer Grundlage plädiert er für ein flexibles Set von Qualitätskriterien mit einem festen Kern (Aktualität, Relevanz,

Faktizität, Unabhängigkeit), das je nach Journalismus-Typ um spezifische Kriterien ergänzt wird. So soll Qualitätsbewertung gegenstandsadäquater, transparenter und inklusiver werden, ohne in Beliebigkeit abzugleiten.

Danksagung

Hervorgegangen ist dieser Band aus der 10. Jahrestagung der Fachgruppe Kommunikations- und Medienethik der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft (DGPuK) in langjähriger Kooperation mit dem Netzwerk Medienethik, dem Zentrum für Medien und die digitale Gesellschaft und der Akademie für Politische Bildung, die im Februar 2025 erstmals an der Akademie für Politische Bildung in Tutzing stattfand. Wir danken den Sprecher:innen der Fachgruppe Kommunikations- und Medienethik, Lars Rademacher und Claudia Paganini, dem Vorbereitungsteam der Tagung sowie allen Beteiligten an der Tagung. Insbesondere danken wir den Autor:innen des Bandes für ihre Beiträge. Besonderer Dank gebührt auch Kristína Janačková, Patrick Lessmeister, Khurshida Tarik und Philipp Staber bei der Unterstützung der redaktionellen Publikationsarbeit. Dem Nomos-Verlag danken wir für die Aufnahme des Bandes in die Reihe.

Literatur

Aïmeur, Esma / Amri, Sabrine / Brassard, Gilles (2023): Fake news, disinformation and misinformation in social media: a review, in: Social Network Analysis and Mining, 1–36, https://doi.org/10.1007/s13278-023-01028-5

Debatin, Bernhard (1997): Ethische Grenze oder Grenze der Ethik? Überlegungen zur Steuerungs- und Reflexionsfunktion der Medienethik, in: Günter Bentele / Michael Haller (Hg.): Aktuelle Entstehung von Öffentlichkeit. Akteure – Strukturen – Veränderungen. Konstanz, S. 281–290.

Muñoz, Katja (2024): Gegen den Strich: Künstliche Intelligenz und Wahlen, in: Internationale Politik, Ausg. 4, S. 36–41.

Riedel, Manfred (1979): Norm und Werturteil, Stuttgart.

Ruppert-Karakas, Sascha (2024): Die Politik des Zorns. Wie die Vordenker der Neuen Rechten den Umsturz vorbereiten, in: Blätter 5/2024, S. 89–98, (online unter: https://www.blaetter.de/ausgabe/2024/mai/die-politik-des-zorns – letzter Zugriff 28.01.2026).

I.
Umstrittene Sprache: Hate Speech und Gender-Diskurse

Was ist Hassrede? Versuch einer Schärfung eines umstrittenen Begriffs

Inga Bones

Zusammenfassung

Der Begriff Hassrede ist im öffentlichen Diskurs in den vergangenen Jahren allgegenwärtig geworden. Empirische Studien zeigen deutlich die Dimension ihrer Verbreitung und im politischen und regulatorischen Bereich werden Gegenmaßnahmen gefordert. Es wird jedoch seltener gefragt, was „Hassrede" genau ist. Das führt zu einer uneinheitlichen Verwendung, die auch die Einordnung entsprechender Studienergebnisse und die Evaluation von Maßnahmen erschwert. Der Beitrag diskutiert kritisch einschlägige Begriffsbestimmungen und zeigt, dass Hassrede als Form der Kommunikation zu verstehen ist, die eine Person oder Gruppe auf Basis stabiler und identitätsrelevanter Eigenschaften wie Religion, Ethnie, Geschlecht oder Herkunft auf tiefgreifende Weise angreift, abwertet oder diskriminiert.

1. Einleitung

Hassrede (englisch *hate speech*) und ihre Bekämpfung waren in den vergangenen Jahren immer wieder Thema in den deutschsprachigen Medien. Anlässe, Hassrede in den Fokus zu nehmen, gab es bedauerlicherweise viele: Der Kasseler Regierungspräsident Walter Lübcke etwa wurde vor seiner Ermordung durch einen Rechtsextremisten im Jahr 2019 zur Zielscheibe von hasserfüllten Kommentaren, Gewaltaufrufen und Morddrohungen. In den vergangenen Bundestagswahlkämpfen mussten sich zahlreiche Politiker:innen mit Hasskommentaren auseinandersetzen (vgl. Hoppenstedt 2021), während der Corona-Pandemie wurde gegen Wissenschaftler:innen wie Christian Drosten oder den Gesundheitsminister Karl Lauterbach gehetzt (vgl. Charisius 2021) und seit dem Anschlag der Hamas auf Israel am 7. Oktober 2023 ist ein Anstieg antisemitischer Hassrede zu verzeichnen (vgl. Hoppenstedt 2023, Rech 2024). Und immer wieder ist von Umfrageergebnissen zu lesen, denen zufolge ein großer Teil aller Internetnutzer:in-

nen, insbesondere jüngerer, schon einmal mit Hassrede konfrontiert oder sogar selbst von ihr betroffen war (vgl. Daniel 2023).

Vor dem Hintergrund solcher Meldungen drängt sich die Frage auf, was wir als Gesellschaft gegen Hassrede unternehmen können: Welche Maßnahmen zur Eindämmung von Hassrede sollten staatliche Institutionen, die Betreiber sozialer Netzwerke, die Anbieter von Messengerdiensten (und so weiter) ergreifen? Angesichts der Vielzahl der in den Medien verwendeten Ausdrücke – *Hasskommentar*, *Onlinehetze*, *Beschimpfung*, *Bedrohung*, *Beleidigung*, *fake news* – stellt sich aber auch eine andere, und grundlegendere, Frage: Was genau ist Hassrede eigentlich? Dieser Frage möchte ich in diesem Text nachgehen.

2. Woher stammt und wer gebraucht den Ausdruck „Hassrede"?

2.1 Woher stammt der Ausdruck „Hassrede"?

In seiner Verwendung als Kollektivum ist der Ausdruck „Hassrede" eine Lehnübersetzung aus dem Englischen („hate speech") und ein relativer Neuzugang des deutschen Alltagswortschatzes. In der deutschsprachigen (Buch-) Literatur tauchte er vor der Jahrtausendwende kaum auf. Zwischen 2000 und 2010 nahm die Häufigkeit seiner Verwendung zu, um schließlich in den 2010er Jahren sprunghaft anzusteigen. Dieser sprunghafte Anstieg ist für denselben Zeitraum auch für den Ausdruck *hate speech* in der englischsprachigen (Buch-) Literatur zu verzeichnen.[1] Eine Recherche mit dem *Google Books Ngram Viewer* zeigt aber auch, dass der Terminus *hate speech* sich im englischen Sprachraum früher als im deutschen zu verbreiten begonnen hat, nämlich bereits in den späten 1980er und frühen 1990er Jahren. Das deckt sich mit einer Untersuchung von Alexander Brown (2017: 424), demzufolge der Ausdruck *hate speech* von einer Gruppe US-amerikanischer Jurist:innen und Aktivist:innen der Post-Bürgerrechtsära nach 1965 geprägt wurde – und zwar in den achtziger und frühen neunziger Jahren.

Wegweisend war hier ein 1989 publizierter Aufsatz der Rechtswissenschaftlerin Mari Matsuda über rassistisch motivierte Beleidigungen und Drohungen. Matsudas Text basiert auf einem Vortrag, den sie ein Jahr zuvor auf einer Konferenz zum Spannungsverhältnis von Redefreiheit und

1 Die hier besprochenen Daten zur Verwendungshäufigkeit der Termini „Hassrede" und „hate speech" basieren auf einer Suche mit dem Google Books Ngram Viewer.

sprachlicher Gewalt an der *Hofstra University* gehalten hatte, und enthält ganze vierzig Vorkommnisse des Ausdrucks *hate speech*. 1993 schließlich erschien mit dem Sammelband *Words That Wound* ein weiterer für die gesellschaftliche Debatte um Hassrede zentraler Text, der unter anderem Beiträge von Matsuda, Richard Delgado und Kimberlé Crenshaw versammelt. Alle drei gehören auch zu den Begründer:innen der *critical race theory*, die sich zum Ziel gesetzt hatte, für strukturellen Rassismus und die Intersektionalität von Diskriminierung zu sensibilisieren. Die Verbreitung des Ausdrucks *hate speech* signalisiert also auch ein gesteigertes öffentliches Bewusstsein für gesellschaftliche Machtverhältnisse, strukturelle Diskriminierung und für solche Phänomene wie Alltagsrassismus.

2.2 Ist „Hassrede" ein umkämpfter Begriff?

Der Ausdruck „Hassrede" ist das, was man in der Sprachphilosophie einen Hybridausdruck nennt: ein Ausdruck, der nicht nur beschreibt, sondern auch eine Wertung enthält. Wer Hassrede äußert, tut per se etwas Schlechtes – anders als jemand, der eine Feststellung macht oder eine Befürchtung äußert. Wenn wir einen Sprechakt als Hassrede klassifizieren, dann machen wir deutlich, dass wir ihn für verwerflich halten, für etwas, dessen moralische oder rechtliche Sanktionierung im öffentlichen Interesse liegt. Letzteres erklärt auch, warum der Begriff der Hassrede umkämpft ist: Die Freiheit der Meinungsäußerung ist ein schützenswertes Gut und ein Grundrecht, das nicht nur in unserer deutschen Verfassung, sondern auch in der europäischen Charta der Grundrechte verankert ist. Welche Sprechakte eine Gesellschaft sanktionieren sollte, ist deshalb Gegenstand kontroverser Debatten. Eine von Verfechtern einer möglichst weitgehenden Redefreiheit manchmal geäußerte Sorge ist, dass Sanktionen von (vermeintlich oder tatsächlich) problematischen Sprechakten zu einem sogenannten *chilling effect* führen könnten, einer Form der Selbstzensur durch Personen, die zwar eine streitbare, aber durchaus keine extremistische politische Meinung vertreten (vgl. Simpson 2024). Andererseits, so eine mögliche Replik, droht eine Zunahme von Hassrede gerade diejenigen Menschen „mundtot" zu machen (englisch *to silence*), die sich aufgrund ihrer Zugehörigkeit zu Minderheitengruppen im gesellschaftlichen Diskurs ohnehin nur schwer Gehör verschaffen können (vgl. West 2012).

2.3 Wer gebraucht den Ausdruck „Hassrede“?

„Hassrede“ ist kein juristischer Begriff. Doch obwohl er im Strafgesetzbuch nicht vorkommt, weist er einige Bezüge zu juristischen Tatbeständen auf. Ein zentraler Bezugspunkt ist die Volksverhetzung, andere sind die sogenannten Ehrdelikte der Beleidigung, Verleumdung und üblen Nachrede. Trotz dieser Bezugspunkte ist es nicht sinnvoll, „Hassrede“ als simplen Sammelbegriff für die genannten Straftatbestände zu verstehen – mehr dazu später.

In Ermittlungsbehörden wie den Landespolizeien wurde im Jahr 2001 durch Beschluss der Innenministerkonferenz ein bundesweites Klassifikationssystem zur Erfassung politisch motivierter Kriminalität (PMK) eingeführt. Gleichzeitig wurde mit dem Oberthemenfeld „Hasskriminalität“ eine Möglichkeit geschaffen, solche Straftaten gesondert zu erfassen, die „durch gruppenbezogene Vorurteile motiviert begangen“ werden. Als gruppenbezogene Vorurteile gelten dabei „Vorurteile des Täters bezogen auf Nationalität, ethnische Zugehörigkeit, Hautfarbe, Religionszugehörigkeit, sozialen Status, physische und/oder psychische Behinderung oder Beeinträchtigung, Geschlecht/geschlechtliche Identität, sexuelle Orientierung, [und] äußeres Erscheinungsbild“ (Bundeskriminalamt 2024). Hasskriminalität umfasst jedoch deutlich mehr als bloße Hassrede, so etwa Sachbeschädigungen, aber auch Gewalttaten wie Körperverletzung oder Mord. Am nächsten kommt dem, was wir intuitiv unter Hassrede verstehen, vermutlich das, was in den Statistiken des Bundesinnenministeriums und des BKA als „politisch motivierte Kriminalität mit dem Tatmittel Hassposting“ (ebd.) auftaucht. Allerdings ist diese Kategorie zu eng, weil ihr zufolge Hassrede auf öffentlich zugängliche Beiträge im Internet beschränkt wäre. Hassrede ist aber kein reines Netzphänomen, auch wenn die Besonderheiten digitaler Räume – zum Beispiel die weitgehende Anonymität – Hassrede zu befeuern scheinen.

Anders als Rechts- und Ermittlungsbehörden verwenden Behörden wie die Bundeszentrale für politische Bildung, die Landesmedienanstalten und zivilgesellschaftliche Organisationen die Ausdrücke „Hassrede“ und „hate speech“ zwar explizit, aber uneinheitlich. Dies führt unter anderem zu sehr unterschiedlichen Untersuchungsergebnissen, wenn es um die Frage geht, welcher Anteil der Bevölkerung bereits mit Hassrede konfrontiert war. Eine Forsa-Umfrage im Auftrag der Landesanstalt für Medien NRW im Jahr 2024 ergab zum Beispiel, dass bereits 78 Prozent aller Internetnutzer:innen

ab 14 Jahre schon einmal mit Hassrede im Netz konfrontiert waren, 42 Prozent sogar „häufig" oder „sehr häufig" (Landesanstalt für Medien NRW 2024).[2] Selbst betroffen war laut dieser Studie rund ein Viertel aller Befragten. Eine andere Studie im Auftrag des Vereins Campact und des Instituts für Demokratie und Zivilgesellschaft (IDZ) legt nach eigener Auskunft eine engere Definition von Hassrede zugrunde und kommt zu deutlich kleineren Fallzahlen: Nur 40 Prozent der Befragten haben *hate speech* wahrgenommen, 8 Prozent waren selbst betroffen. Im Unterschied zur ersten Studie waren die Teilnehmer:innen dieser zweiten Studie vorab dazu angehalten worden, Hassrede von individuellen – also nicht gruppenbezogenen – Beleidigungen, Belästigungen oder *cybermobbing* zu unterscheiden (Geschke et al. 2019).

Der Ausdruck „Hassrede" wird von Behörden, zivilgesellschaftlichen Organisationen und den Medien also höchst uneinheitlich verwendet; er wird oftmals nicht abgegrenzt von *fake news* und *cybermobbing*, individuellen Beleidigungen oder Belästigungen; und manchmal als reines Netzphänomen definiert. Welches Phänomen haben wir – als Gesellschaft, als Medienschaffende, als Angehörige staatlicher und nicht-staatlicher Institutionen – also eigentlich im Blick, wenn wir von Hassrede sprechen? Mit Götz Hamann (2024) müssten wir wohl antworten: Gar keines, oder zumindest kein Konkretes. In einem Artikel aus dem letzten Jahr schreibt Hamann: „Es gibt politische Begriffe, die waren gestern spektakulär und sind heute ziemlich leer. […] Medien [nutzen] die Worte Hass und Hetze viel zu häufig als bequeme Metapher in Überschriften, was dazu führt, dass Hass und Hetze den Status einer Füllmasse erreicht haben. Sie sind oft nur noch rhetorisches Moltofill, kitten Probleme eher zu, als sie konkret zu benennen. Der Begriff Hassrede ist auf demselben Weg." Hamann verbindet diese Diagnose mit der Forderung, auf die Verwendung der genannten Begriffe zu verzichten, sie zu „streichen, wo es nur geht." Er fährt fort: „Wir haben an vielen Stellen schon die passenden Begriffe: Volksverhetzung, Verleumdung, Androhung von Gewalt, Rassismus."

Anstatt den Begriff der Hassrede aus unserem Wortschatz zu streichen, könnten wir als Antwort auf seine uneinheitliche und vielleicht tatsächlich inflationäre Verwendung auch versuchen, ihn zu schärfen – und klar abzugrenzen von den Begriffen, die Hamann aufzählt. Die Frage wäre dann

2 „Wie häufig haben Sie persönlich schon Hate Speech bzw. Hasskommentare im Internet gesehen – zum Beispiel auf Webseiten, in Blogs, in sozialen Netzwerken oder in Internetforen?".

weniger, welches Phänomen wir im Blick haben, wenn wir den Ausdruck „Hassrede“ verwenden, sondern welches Phänomen wir im Blick haben *sollten*. Oder, anders formuliert: Die Frage wäre weniger, was Hassrede *ist*, als was Hassrede *sein sollte*. Bei der Beantwortung dieser (normativen) Frage können wir uns an paradigmatischen Beispielen orientieren, die in den Rechtswissenschaften und der Philosophie diskutiert wurden.

3. Was ist Hassrede? Oder: Wie sollten wir Hassrede verstehen?

3.1 Hassrede und Hass

Fragen wir uns also, was wir denn nun eigentlich unter Hassrede verstehen sollten, dann liegt vielleicht eine Antwort auf der Hand: nämlich, dass Hassrede „der sprachliche Ausdruck von Hass“ (Meibauer 2013) sei. Wer schriftlich oder mündlich Hassrede äußert, so die Idee, der verleiht seinem Hass mit sprachlichen Mitteln Ausdruck. Diese naheliegende Antwort ist aber in mehrerlei Hinsicht problematisch: Erstens scheint Hassrede mehr zu umfassen als nur schriftliche oder mündliche Äußerungen, nämlich auch nicht-sprachliche Symbole und Bildmaterial sowie Gesten. In dem eingangs genannten Aufsatz der Rechtswissenschaftlerin Mari Matsuda etwa wird das Anbringen eines Swastika-Symbols (Hakenkreuz) an den Schreibtischen einiger afro-amerikanischer Mitarbeiter der Feuerwehr in San Francisco als Beispiel rassistischer Hassrede angeführt und die US-amerikanische *Anti Defamation League* führt das OK-Handzeichen seit einiger Zeit als Hasssymbol der Ideologie weißer Vorherrschaft. Zweitens kann man alles Mögliche hassen, zum Beispiel den gegnerischen Fußballverein oder verkochte Spaghetti. Aber „Verdammter HSV!“ oder „Ekelerregender Nudelmatsch!“ würden wir nicht als Hassrede klassifizieren. Seinem Hass sprachlich oder auf anderem Wege Ausdruck zu verleihen, kann also keine hinreichende Bedingung für das Vorliegen von Hassrede sein. Und drittens kann auch Hassrede äußern, wer keinen Hass verspürt: Vielleicht hetzt jemand aus politischem Kalkül gegen bestimmte gesellschaftliche Gruppen oder aus falsch verstandener Loyalität gegenüber Personen aus der eigenen Peergroup. Folglich sind Hassgefühle auf Seiten des Sprechers oder der Sprecherin auch keine notwendige Bedingung für das Vorliegen von Hassrede.

Teresa Marques (2023) argumentiert in einem Aufsatz für die These, dass Hassrede in einem bestimmten Sinn aber sehr wohl „der Ausdruck von

Hass" sei – in dem Sinne nämlich, in dem das Vorliegen einer Emotion des Hasses auf Sprecher:innenseite zu den konventionell festgelegten Angemessenheitsbedingungen für die Äußerung von Hassrede gehört. Ein Beispiel zur Illustration des Grundgedankens geht auf David Kaplan (2004) zurück: Eine Äußerung wie „Autsch!" drückt Kaplan zufolge aus, dass die Person, welche die Äußerung vornimmt, eine Schmerzempfindung hat. Die Verwendung des Ausdrucks „Autsch!" ist in unserer Sprache so geregelt, dass sie nur dann *expressiv korrekt* ist, wenn die Sprecher:in tatsächlich einen Schmerz verspürt. Dies schließt den Fall nicht aus, dass die Sprecher:in den Ausdruck unaufrichtig verwendet – entscheidend ist, dass „Autsch!" ein konventionelles sprachliches Werkzeug für den Ausdruck von Schmerz ist. Ganz analog, so Marques, ist Hassrede ein konventionelles sprachliches Werkzeug für den Ausdruck von Hass. Wenn jemand rein aus politischem Kalkül hetzt, dann verwendet er oder sie Hassrede zwar unaufrichtig, aber die geäußerten Worte drücken dennoch per semantische Konvention Hass aus.

Marques' These ist plausibel, sofern der Fokus auf sogenannten *slurs* liegt: Herabwürdigende Gruppenbezeichnungen wie das N-Wort oder andere Ethnophaulismen sind offenbar tatsächlich semantisch markiert und werden auch dann als diskriminierend wahrgenommen, wenn die Sprecher:innen keine negativen Emotionen gegenüber der bezeichneten Gruppe verspüren. Aber Hassrede umfasst deutlich mehr und komplexere Äußerungen. Es ist weniger plausibel, dass grammatisch komplexe und häufig stark kontextgebundene Bildungen wie „Illegale Einwanderer in ihre Herkunftsländer abschieben!" auf semantischer Ebene als hass-expressive Ausdrücke markiert sind.

Wenn Hass eine zentrale Rolle in einer Definition von Hassrede spielen soll, dann vielleicht eher in dem Sinne, dass Hassrede *zu Hass aufstacheln* soll – Hass wäre dann also nicht die tatsächliche Motivation des Sprechers oder der Sprecherin oder die von der Äußerung ausgedrückte Emotion, sondern der intendierte Effekt des Sprechakts. Eine solche Auffassung rückt den „Hass" in „Hassrede" dann in die Nähe der Volksverhetzung. Und tatsächlich sind die Ausdrücke „Hass" und „Hetze" etymologisch eng verwandt: Im Germanischen bedeutet das Verb „hassen" so viel wie „feindlich verfolgen". Diese Ursprungsbedeutung ist in den Wörtern „Hatz" – eine Variante der Hetzjagd – und „Hetze" erhalten geblieben. Für ein Verständnis von Hassrede als Volksverhetzung spricht auch, dass Hassrede von hochrangigen Stellen als Vorläufer von Verbrechen im Sinne des Völkerrechts und insbesondere von Genoziden charakterisiert wird, etwa von Antonio

Guterrez (2019) anlässlich der Vorstellung des Aktionsplans der Vereinten Nationen gegen *hate speech*.

Guterrez erinnert in seiner Rede daran, dass beispielsweise dem Völkermord an den Tutsi in Ruanda eine systematische Hetzkampagne in an die Hutu adressierte Zeitungs- und Radiopropaganda vorausging. Auch Lynne Tirrell (2012), eine US-amerikanische Sprachphilosophin, zieht die Propaganda gegen die Tutsi, die öffentlich als „inyenzi", Kakerlaken, entmenschlicht wurden, als ein paradigmatisches Beispiel für Hassrede heran. Aber: Wenn man Hassrede (ausschließlich) als Volksverhetzung versteht, bleiben andere paradigmatische Fälle außen vor. Volksverhetzung liegt nämlich erst dann vor, wenn eine Äußerung geeignet ist, den öffentlichen Frieden zu stören. Das ist aber erst dann der Fall, wenn diese Äußerung ein gewisses Ausmaß der Verbreitung erreicht und wenn sie „ihrem Inhalt nach erkennbar auf rechtsgutgefährdende Handlungen hin angelegt [ist], d.h. den Übergang zu Aggression oder Rechtsbruch markier[t]."[3] Eine ganz ähnliche Auffassung vertritt bereits John Stuart Mill in *On Liberty*. Mill (1974/1859: 77) schreibt: „Selbst Gedanken verlieren ihre Straflosigkeit, wenn die Umstände, unter denen sie ausgesprochen werden, von der Art sind, daß ihr Ausdruck eine direkte Aufreizung zu irgendeiner Schandtat bildet." Ausgeschlossen wären damit Fälle, in denen Einzelpersonen direkt – und nicht im Beisein eines größeren Publikums – verbal angegriffen, bedroht oder eingeschüchtert würden, aber auch Fälle, die nicht erkennbar auf rechtsgutgefährdende Handlungen angelegt sind. In einer anderen Hinsicht scheint die Kategorie der Volksverhetzung aber auch zu weit zu sein. Noch einmal Mill (ebd.): „Die Meinung, daß Getreidehändler die Armen aushungern […] sollte gerechterweise Strafe nach sich ziehen, wenn man sie mündlich einer erregten Menge, die sich vor dem Hause eines Getreidehändlers versammelt hat, vorträgt oder sie unter gleichen Umständen in Form von Handzetteln in Umlauf setzt." Es sind Umstände denkbar, unter denen Mills Beispiel „Getreidehändler hungern die Armen aus" zu Gewalttaten aufstacheln und damit den Straftatbestand der Volksverhetzung erfüllen würde. Ob es sich aber um Hassrede handelt, ist zumindest fraglich.

Es scheint unstrittig, dass jede beliebige gesellschaftliche Gruppe unter bestimmten Umständen den Hass – oder, vorsichtiger, die Abneigung – anderer Teile der Bevölkerung auf sich ziehen kann. Prinzipiell kann gegen Getreidehändler:innen, gegen Schulbusfahrer:innen und Grundschul-

3 Aus einer Grundsatzentscheidung des Bundesverfassungsgerichts im Jahr 2009, der sogenannten Wunsiedel Entscheidung.

lehrer:innen, gegen SUV-Fahrer:innen und die Besitzer:innen von Gucci-Handtaschen gehetzt werden. Dass uns solche Gruppen eher selten in den Sinn kommen, wenn wir von Hassrede hören oder lesen, hat sicherlich etwas damit zu tun, dass eine grundsätzliche Abneigung gegen diese Gruppen nicht sehr verbreitet ist. Vielleicht spielt aber auch eine Rolle, dass die Eigenschaften, Schulbusse zu fahren oder Gucci-Handtaschen zu besitzen, nicht so viel damit zu tun haben, *wer wir sind* – dass sie nicht stabil, nicht zentral für unsere Identität, sondern verlierbar und zufällig sind.

3.2 Hassrede und gruppenbezogene Vorurteile

Erinnern wir uns noch einmal an die Anfänge des Ausdrucks „hate speech". Der Ausdruck wurde vor dem Hintergrund von größer angelegten Bestrebungen geprägt, strukturelle Diskriminierung sichtbar zu machen und zu kritisieren. Im Fokus standen dabei solche gruppendefinierenden Eigenschaften, die stabil und zentral für unsere Identität sind und an denen sich Diskriminierung historisch gesehen besonders häufig festgemacht hat: Allen voran „race", aber auch die ethnische Herkunft, Nationalität, Klasse, Geschlecht oder Gender, Religion oder sexuelle Orientierung. Manchmal werden diese Eigenschaften auch „protected characteristics" genannt – und in Artikel 3 unseres Grundgesetzes werden die meisten von ihnen explizit genannt.

Der UN-Aktionsplan gegen Hassrede nimmt diese besonderen Schutz verdienenden Eigenschaften in seine Definition mit auf: Hassrede ist „jede Form der Kommunikation in Wort, Schrift oder Verhalten, die eine Person oder eine Gruppe auf der Basis dessen, was sie sind – mit anderen Worten, auf der Basis ihrer Religion, Ethnie, Nationalität, „Rasse", Hautfarbe, Abstammung, Gender – angreift, abwertet oder diskriminiert."[4] Diese Definition umfasst sowohl paradigmatische Fälle von *Hassrede als Volksverhetzung* – wie etwa die Hetzkampagne gegen die Tutsi – als auch paradigmatische Fälle von *Hassrede als Beleidigung* oder *Bedrohung*, die sich direkt an das Opfer wendet – etwa, wenn einer *Person of Color* auf offener Straße das N-Wort entgegengeschleudert wird. Als Richtschnur dafür, welche Sprechakte wir rechtlich oder sozial sanktionieren sollten, scheint die Definition

4 "[...] any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

aber (erneut) zu weit: Äußerungen wie „Mädchen können nicht Fußball spielen“ oder „Deutsche haben keinen Geschmack“ sind diskriminierend und abwertend. Als Hassrede würden sie vermutlich die wenigsten von uns bezeichnen wollen. Damit ein kommunikativer Akt die Bezeichnung als „Hassrede“ verdient, muss offenbar ein gewisser Schweregrad gegeben sein. Sonst verliert der Begriff, wie von Hamann befürchtet, wohl tatsächlich an „politischer Wucht“. Wann ein hinreichender Schweregrad erreicht ist, lässt sich nicht präzise ausbuchstabieren und ist sicher auch Gegenstand gesellschaftlicher Aushandlung. Als Orientierung können uns aber eben jene paradigmatischen Beispiele dienen, die in der wissenschaftlichen Literatur zu *hate speech* wiederholt auftauchen und von denen oben einige angeführt wurden.

Wenn Hassrede die Zugehörigkeit zu einer durch geschützte Eigenschaften definierten Gruppe voraussetzt, dann kann es keine Hassrede gegen Politiker:innen *als Politiker:innen* oder gegen Wissenschaftler:innen *als Wissenschaftler:innen* geben. „Alle Politiker:innen gehören an den Pranger“ oder „Weg mit dem korrupten Wissenschaftlerpack“ qualifizieren sich also nicht als Hassrede. (Aber sie können sich unter bestimmten Umständen durchaus als Volksverhetzung qualifizieren.) Trotzdem kann es sein, dass Politiker:innen und andere Personen des öffentlichen Lebens besonders häufig zu Opfern von Hassrede werden, nämlich dann, wenn sie besonders häufig auf der Basis ihres Geschlechts, ihrer Herkunft, ihrer religiösen Überzeugungen oder ihrer sexuellen Orientierung angefeindet werden.

4. Fazit

„Hassrede“ ist ein umkämpfter Begriff, der in der öffentlichen Debatte, von Behörden und Organisationen uneinheitlich verwendet wird. Ich habe für eine Schärfung des Begriffs argumentiert, die sich weniger an der expressiven Funktion von Hassrede als an ihrer Verknüpfung mit gruppenbezogenen Vorurteilen orientiert. Demnach ist Hassrede eine Form der Kommunikation, die eine Person oder Gruppe auf der Basis stabiler und identitätsrelevanter Eigenschaften (wie Religion, Ethnie, Geschlecht oder Herkunft) in schwerwiegender Weise angreift, abwertet oder diskriminiert.

Eine Schärfung des Begriffs der Hassrede und damit eine klare Abgrenzung von persönlichen Beleidigungen ist insbesondere dann relevant, wenn empirische Forschung belastbare Daten zur Prävalenz und zu den Auswirkungen von Hassrede – zum Beispiel in den sozialen Medien – liefern

soll. Auch die (manuelle oder automatisierte) Moderation von Beiträgen in sozialen Medien profitiert von einem geschärften Begriffsverständnis, etwa wenn im Einzelfall über die mögliche Überschreitung der Grenze zur Strafbarkeit oder über die juristische Grundlage einer Strafverfolgung entschieden werden muss.

Literatur

Bundeskriminalamt (2024): Definitionssystem. Politisch motivierte Kriminalität (online unter: https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/themen/sicherheit/definitionssystem-pmk.pdf?__blob=publicationFile&v=2 – letzter Zugriff: 12.1.2026).

Brown, Alexander (2017): What is hate speech? Part 1: The myth of hate, in: Law and Philosophy 36, S. 419–468.

Charisius, Hanno (2021): Jeder siebte Corona-Experte berichtet von Morddrohungen, in: Süddeutsche Zeitung, 13. Oktober 2021 (online unter: https://www.sueddeutsche.de/gesundheit/corona-forscher-drohungen-1.5438622 – letzter Zugriff: 21.9.2025).

Daniel, Isabelle (2023): Mehr als jeder vierte Internetnutzer hat Erfahrung mit Hassrede, in: Die ZEIT, 11. Dezember 2023 (online unter: https://www.zeit.de/gesellschaft/zeitgeschehen/2023-12/statistisches-bundesamt-hatespeech-internet – letzter Zugriff: 21.9.2025).

Geschke, Daniel et al. (2019): #Hass im Netz: Der schleichende Angriff auf unsere Demokratie. (online unter: https://www.idz-jena.de/fileadmin/user_upload/_Hass_im_Netz_-_Der_schleichende_Angriff.pdf – letzter Zugriff: 12.1.2026).

Guterrez, Antonio (2019): Secretary-General's Remarks at the Launch of the United Nations Strategy and Plan of Action on Hate Speech (online unter: https://www.un.org/sg/en/content/sg/statement/2019-06-18/secretary-generals-remarks-the-launch-of-the-united-nations-strategy-and-plan-of-action-hate-speech-delivered – letzter Zugriff: 21.9.2025).

Hamann, Götz (2024): "Hass und Hetze" klingt heute hohl, in: Die ZEIT, 20. Mai 2024 (online unter: https://www.zeit.de/kultur/2024-05/politische-sprache-hass-hetze-bedrohung-kritik – letzter Zugriff: 21.9.2025).

Hoppenstedt, Max (2021): Täglich Hass, tausendfach, in: Spiegel Online, 20. September 2021 (online unter: https://www.spiegel.de/netzwelt/netzpolitik/hasskommentare-gegen-die-kanzlerkandidaten-schimpf-und-schande-a-ff7b6878-f9d7-46e2-b4b9-e0a78a2c4957?sara_ref=re-xx-cp-sh – letzter Zugriff: 21.9.2025).

Hoppenstedt, Max (2023): Ermittler verzeichnen deutliche Zunahme antisemitischer Hasskommentare, in: Spiegel Online, 6. November 2023 (online unter: https://www.spiegel.de/netzwelt/antisemitismus-ermittler-sehen-deutliche-zunahme-von-online-hasskommentaren-a-ff2f4fa3-4b2c-449c-aca9-49d06fab1ea2?sara_ref=re-xx-cp-sh – letzter Zugriff: 21.9.2025).

Kaplan, David (2004): The Meaning of Ouch and Oops, Howison Lecture in Philosophy, UC Berkeley, 23. August 2004 (Transkription online unter: https://eecoppock.info/PragmaticsSoSe2012/kaplan.pdf – letzter Zugriff: 21.9.2025).

*Landesanstalt für Medien NRW (*Hg.) (2024): Hate Speech. Forsa-Studie 2024 (online unter: https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/themen/sicherheit/definitionssystem-pmk.pdf?__blob=publicationFile&v=2 – letzter Zugriff 12.1.2026).

Marques, Teresa (2023): The Expression of Hate in Hate Speech, in: Journal of Applied Philosophy 40, S. 769–787.

Matsuda, Mari (1989): Public Response to Racist Speech: Considering the Victim's Story, in: Michigan Law Review 87 (8/1989), S. 2320–2381.

Meibauer, Jörg (2013): Hassrede – von der Sprache zur Politik, in: Jörg Meibauer (Hg.), Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion, Gießener Elektronische Bibliothek, S. 1–16.

Mill, John Stuart (1974/1859), Über die Freiheit, Stuttgart.

Rech, David (2024): Wie sich antisemitische Kommentare nach dem 7. Oktober verändert haben, Die ZEIT, 18. April 2024 (online unter: https://www.zeit.de/gesellschaft/2024-04/antisemitismus-kommentare-studie-israel-decoding-antisemitism – letzter Zugriff: 21.9.2025).

Simpson, Robert Mark (2024): Self-Censorship: The Chilling Effect and the Heating Effect, in: Political Philosophy 1 (2/2024), S. 345–380.

Tirell, Lynne (2012): Genocidal Language Games, in: Ishani Maitra / Mary Kate McGowan (Hg.), Speech and Harm: Controversies over Free Speech, Oxford University Press, S. 174–221.

West, Caroline (2012): Words That Silence? Freedom of Expression and Racist Hate Speech, in: Ishani Maitra / Mary Kate McGowan (Hg.), Speech and Harm: Controversies Over Free Speech, Oxford University Press, S. 222–248.

„Squatting for Hitler“: Hate Speech, Datenmüll, und generative KI in Zeiten rechtspopulistischer Autokratie

Bernhard Debatin

Zusammenfassung

Mit Beispielen aus dem US-Wahlkampf und von KI-generierter Hassmusik wird analysiert, wie Hassrede und rassistische, antisemitische und misogyne Propaganda durch KI-Plattformen verstärkt wird, und wie dadurch Normalisierungseffekte und *confirmation bias* entstehen. Aus analytischer Sicht geht es hier zunächst um die Frage, inwieweit die Idee der Technik als bloßes Werkzeug einer wirkungsvollen Entwicklerverantwortung im Wege steht. KI ist ja, im Unterschied zu anderen Technologien, in der Lage, selbst Zwecke zu setzen, also stets mehr als nur Werkzeug. Dies erzeugt das bekannte Problem der *tragedy of the commons* und folglich die zunehmende Vermüllung des Internets mit KI-Content. Zum Abschluss wird diskutiert, wie in den USA unter Präsident Trump unregulierte und ungebremste KI als Propagandawerkzeug instrumentalisiert und normalisiert wird.

1. Die neue politische Landschaft

In den USA findet seit dem Amtsantritt von Präsident Trump der Übergang zur offen oligarchischen Autokratie statt, mit einer Flut von Präsidentenverfügungen (*executive orders*), die an der Gesetzgebung vorbei tiefgreifende Veränderungen bewirken. Viele von ihnen sind fragwürdig, wenn nicht gar illegal, da sie ohne Legitimation Macht- und Entscheidungsbefugnisse der Legislative, zum Beispiel in Haushalts-, Wirtschafts- und Verteidigungsfragen, in die Exekutive verlagern. Dies ist Teil eines nie zuvor gekannten *powergrab*, einer Machtergreifung durch einen Präsidenten mit diktatorischen Ambitionen.

Die MAGA-Ideologien, allen voran Stephen Miller, Russell Vought und ihre Mitstreiter von der ultrakonservativen *Heritage Foundation*, sind – im Unterschied zur ersten Präsidentschaft von Trump – sehr gut und de-

tailliert vorbereitet. Trumps politische Schritte, vor allem seine *executive orders* und die (ebenso wirksamen) *memoranda*, folgen präzise den von der *Heritage Foundation* in der 900-seitigen Kampfschrift *Project 2025* (Dans/Groves, 2023) niedergelegten taktischen und strategischen Zügen. Im Stil einer autokratischen Machtergreifung setzen sie darauf, alle Macht in der Exekutive zu konzentrieren und möglichst viele demokratische Strukturen zu demontieren oder umzubauen, bevor Gerichte einschreiten können – sofern sie das denn tun.[1]

Präsidentenverfügungen, die am Kongress vorbei Fakten schaffen, sind dabei das Mittel der Wahl. Schnell handeln ist die Devise: In den ersten 14 Tagen seit Amtsantritt erließ Trump 53 solche Verfügungen (vgl. Federal Register 2025). Ende September 2025 waren es bereits 209 *executive orders* (vgl. Federal Register 2025), während Präsident Biden im ersten Jahr seiner Amtszeit 77 *executive orders* erließ (vgl. Federal Register 2021) und Präsident Trump im ersten Jahr seiner ersten Amtszeit 2017 sogar nur 55 (vgl. Federal Register 2017). Der Nachrichtensender *ABC* schrieb dazu: „This number of executive orders, unprecedented in the modern era of politics, are part of a wider suite of sweeping actions that aim to rapidly dismantle existing institutions and initiatives while establishing a governing framework aligned with Trump's *MAGA* agenda and his populist base" (Conroy 2025).

Viele der *executive orders* zielen auf eine ideologische Ausrichtung der Institutionen und Behörden, sowie auf die Annullierung bisheriger Direktiven.[2] In den zentralen Bereichen Immigration, Zölle, Umwelt und Energiewirtschaft wurde per Notverordnung ein Ausnahmezustand erklärt, was dem Präsidenten zusätzliche Machtbefugnisse bringt, ohne dass Kongress oder Gerichte beteiligt sind. Dies folgt dem Muster einer autoritären Machtübernahme, begleitet von willkürlicher Schließung von Behörden, Zurückhalten von verabschiedeten Haushaltsmitteln, Zerstörung von offizi-

1 Einer Analyse der *Washington Post* zufolge ignorierte die Trump Administration bis Mitte Juli 2025 ein Drittel der über 160 gegen sie gerichteten Gerichtsurteile (Jouvenal 2025), was als „unprecedented threat to the U.S. legal system" gewertet wird. Hinzu kommt, dass viele Gerichtsurteile angefochten und in zweiter Instanz aufgehoben oder an den *Supreme Court* verwiesen werden, der meist mit unbegründeten Notverordnungen (*Emergency Orders*) anstelle von juristisch begründeten Entscheidungen reagiert, was ebenfalls als Krise im Rechtssystem kritisiert wird (vgl. Schwartz/Montague 2025).

2 Bis 5. Februar 2025 hatte Trump bereits 96 Executive Orders von seinen Amtsvorgängern für ungültig erklärt, wobei manche dieser Verordnungen bis in die 1960er Jahre zurückreichen.

ellen Dokumenten, sowie Einschüchterung von Gegnern und Gewalt gegen und massenhafte Deportation von Immigranten.

Der *One Big Beautiful Bill Act*, das Gesetzespaket, das Steuervergünstigungen für Superreiche in Billionenhöhe festschrieb, enthielt auch eine umfangreiche Finanzspritze für die *Immigration and Customs Enforcement agency* (*ICE*) in Höhe von $170 Milliarden über die nächsten vier Jahre, mit dem ausgewiesenen Ziel, eine Millionen Immigranten pro Jahr zu deportieren. Dem *Brennan center* (2025) zufolge fließt der größte Anteil dieser Gelder in „finding, arresting, detaining, and deporting immigrants already living in the U.S., most of whom have not committed a crime and many of whom have had lawful status.“ Dazu wurden privat organisierte Internierungslager eingerichtet, die massiv von der Deportationswelle profitieren (vgl. Eisen 2025). *ICE*-Agenten arbeiten im Stil einer Geheimpolizei (gesetzeswidrig, vgl. New York City Bar 2025) vermummt und ohne erkennbare Identifikation, und verhaften dabei (ebenfalls gesetzeswidrig, vgl. Rivera et al. 2025) ohne richterliche Anordnung einfach auf Verdacht Menschen auf der Straße, am Arbeitsplatz, und vor Schulen, Krankenhäusern, Gerichten und Kirchen. Gleichzeitig werden demokratisch geführte Städte und Staaten unter Verweis auf einen angeblichen Ausnahmezustand mit Hilfe der quasimilitärischen Nationalgarde besetzt und eingeschüchtert (vgl. Gamio/Hippensteel 2025).

Die Aktionen von Elon Musks *DOGE (Department Of Government Efficiency)* sind ein weiteres Beispiel für autokratische Willkürherrschaft. Entgegen dem normalen, gesetzlich vorgesehenen Verfahren wurde Musk nicht vom Kongress berufen, sondern von Trump per Verordnung eingesetzt, um massive Kürzungen durch Entlassungen zu generieren, von denen Trump einen Teil der Steuererleichterungen finanzieren will. Die Steuerkürzung in seiner ersten Amtszeit erhöhte das Staatsdefizit schon um 8 Billionen Dollar. Der *Big Beautiful Bill Act* wird über die nächsten 10 Jahre zu einem weiteren Verlust von mindestens 4,5 Billionen Dollar führen, bei gleichzeitig um 1,1 Billionen Dollar reduzierten Ausgaben bei staatlichen Sozial- und Krankenversicherungsleistungen. Das wird dem Defizit nochmal mindestens 3,4 Billionen Dollar hinzufügen (vgl. Congressional Budget Office 2025).[3]

Unter Musk wurden Agenturen wie *USAID* wurden kurzerhand geschlossen. Das Erziehungsministerium wurde so radikal kleingespart, dass

3 Andere Schätzungen, die auch Folgekosten mit einbeziehen, kommen allerdings zu weitaus höheren Zahlen, bis zu 9,75 Billionen Dollar (vgl. Dayen 2024).

es handlungsunfähig wurde (vgl. Turner 2025). Mehr als 200,000 Staatsbedienstete wurden von Januar bis Ende Oktober 2025 am Gesetz vorbei entlassen oder in den Vorruhestand geschickt, damit man sie durch ‚linientreue' *MAGA*-Anhänger ersetzen beziehungsweise die Behörden bis zur Dysfunktion verkleinern kann (vgl. Federal Harms Tracker 2025). Offizielle Webseiten wurden demontiert und ‚auf Linie' gebracht, der ideologisierten Sprachpolitik folgend, nach der es „gender", „diversity", „equity", und viele andere Begriffe nicht mehr geben darf (vgl. Yourish et al. 2025). Die föderale Forschungsförderung (zum Beispiel NIH, NSF, NEH) wurde zunächst ganz gestoppt, musste aber per Gerichtsverfügung wieder aufgenommen werden und wird nun erheblich zusammengekürzt: „The administration has proposed cutting federal funding for all research by 22 % and basic research by 34 % in the next fiscal year. This would effectively cede the race to China, whose research spending is expected to grow by 10 % this year alone", stellte die *Association of American Universities* im Juli 2025 fest.

2. Hassrede und generative KI

Am ersten Tag seiner neuen Amtszeit setzte Trump in seinem zweiten *executive order* insgesamt 67 frühere präsidiale Verfügungen außer Kraft (vgl. Federal Register, 2025a), darunter auch Biden's *executive order* 14110 vom 30. Oktober 2023 zur KI Sicherheit mit acht Empfehlungen und detaillierten Richtlinien, da die unverantwortliche Verwendung von KI zu starken sozialen Folgeschäden führen kann, wie etwa „fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and (…) risks to national security" (Federal Register, 2023).

Präsident Bidens Verordnung war eine vernünftige Reaktion auf den eindringlichen Appell von leitenden KI-Wissenschaftlern nach staatlicher Risikominimierung (vgl. Roose, 2023). „Mitigating the risk of extinction from A.I. should be a global priority alongside other societal-scale risks, such as pandemics and nuclear war", hieß es in einer Erklärung vom 30. Mai 2023, die von mehr als 350 Managern, Forschern und Ingenieuren aus der KI-Industrie unterzeichnet war (Center for AI Safety 2025). Übrigens warnte selbst der erzkonservative Politiker Henry Kissinger schon 2018 vor dem Ende der Aufklärung durch KI und betonte: „we must expect AI to make mistakes faster — and of greater magnitude — than humans do" (Kissinger 2018).

Für Trump sind solche Bedenken lediglich Innovationsbarrieren, die schnellstens eliminiert werden müssen. Die 33. Präsidentenverfügung von Trump, erlassen drei Tage nach seiner Amtseinführung, mit dem Titel „Removing Barriers to American Leadership in Artificial Intelligence“ (Federal Register 2025b), zielt auf „eliminating harmful Biden Administration AI policies and enhancing America's global AI dominance“ (White House 2025). Sie steckt eine neue, unregulierte KI-Politik ab, mit der Ankündigung eines noch zu erarbeitenden „Action Plan“ zur Sicherung der US-Dominanz im KI-Bereich. Wie notwendig Regulierung aber allein schon im Nutzungsbereich ist, sollen die folgenden Beispiele aus der KI-generierten Musikproduktion zeigen:

Am 14. Juni 2024 berichtet die US-amerikanische *Anti-Defamation League* (*ADL*), dass mithilfe des generativen KI-Systems Suno hasserfüllte, aufhetzende, rassistische, extremistische und gewaltverherrlichende Musik, Texte, Albumcovers, und komplette Songs erzeugt werden (vgl. ADL 2024). Die Titelseite des Berichts zeigt ein digitales, von der KI kreiertes Albumcover unter dem Genre „Gangster Rap“ und mit dem Titel „Squatting for Hitler“ sowie den Anfang des Liedtextes, der „national awakening“ beschwört und die „White Man's Nation“ für Hitler besetzen will. Der Song, in voller Länge auf Suno (2024) erhältlich, schwärmt davon, dass man den Nicht-Weißen beibringen wird, sich vor der weißen Macht zu verneigen, und einen totalen Krieg führen wird.

Die *ADL* fand auf der Suno-Website eine umfangreiche Sammlung von über 1000 KI-generierten Songs mit rassistischen, antisemitischen, misogynen und xenophoben Inhalten. In den Texten dieser Songs finden sich Motive wie nationales Erwachen, totaler Krieg, Unterwerfung von schwarzen und braunen Menschen, White Power, Endlösung, sowie jede Menge rassistischer, antisemitischer und misogyner Beschimpfungen und Verschwörungstheorien. Zwar hat Suno eine automatische Content Moderation, die verhindert, dass die Generierungs-Prompts explizit Gewalt oder Hass verherrlichen, aber durch indirekte Prompts kann man dies leicht umgehen (vgl. ADL 2024).

Die bevorzugte Technik ist hierbei das sogenannte *dog-whistling*, ursprünglich definiert von Ian López (2014: 3ff.) als eine Technik, bei der einer Zielgruppe durch entsprechende Kodierung scheinbar neutrale Inhalte übermittelt werden, die bestimmte soziale Gruppen herabwürdigen oder vor ihnen warnen: „…modern racial pandering always operates on two levels: inaudible and easily denied in one range, yet stimulating strong reactions in another“ (López 2014: 3). *Dog-whistling* ist eine Form des

strategischen Rassismus, bei der rassistischer Hass gezielt geschürt und zur Erzielung von Vorteilen verwendet wird.

Dies geschieht durch drei wesentliche Schritte (vgl. López 2014: 130): Zunächst wird eine Verbindung hergestellt zwischen Ethnizität oder „Rasse" auf der einen Seite und Kultur, Verhalten und sozialer Klasse auf der anderen Seite. Damit wird an rassistisch geprägten Hass und existierende Vorurteile in der Zielgruppe angeknüpft, gleichzeitig wird der Rassismus weiter geschürt und mit Munition versorgt. López nennt etwa das seit Ronald Reagan verbreitete rassistische Meme der schwarzen *Welfare Queen*, die angeblich auf Staatskosten wie die Made im Speck lebt. Man mag auch an das von Trump und Vance im letzten US-Wahlkampf in Umlauf gebrachte Gerücht denken, dem zufolge Haitische Flüchtlinge Haustiere von weißen Bewohner:innen der Stadt Springfield in Ohio essen würden, ein Gerücht, das keinerlei faktische Basis hat, aber alle Register der rassistischen Rhetorik zieht und als Meme immer noch verbreitet ist.

Sodann wird im zweiten Schritt der Vorwurf der Rassismus vorbeugend dadurch abgewehrt, dass man behauptet, Rassismus läge nur vor, wenn biologisch argumentiert oder offen rassistisch gehetzt wird. Damit werden struktureller und verdeckter Rassismus negiert, was in den USA oft unter dem Stichwort der *colorblind society* läuft. Hiermit wird die vermeintliche Überwindung des Rassismus heraufbeschworen, als ob dies nur eine Frage eines Entschlusses oder der individuellen Einstellung wäre.

Wer trotzdem behauptet, dass es Rassismus und Diskriminierung noch gibt, der wird im dritten Schritt dann als der „eigentliche" Rassist bezeichnet werden, da man ja Rassismus in die Diskussion gebracht und damit die *race card* gespielt habe. Ein Beispiel für die *race card*-Strategie ist die von Trump im Wahlkampf erhobene Behauptung, die demokratische Kandidatin Kamala Harris habe erst kürzlich ihre schwarze Identität „entdeckt", und so die *race card* gespielt. Dies ist eine weit verbreitete Taktik, mit dem der Rassismus Vorwurf umgedreht und gegen die Opfer des Rassismus gewendet wird (vgl. López 2014: 130).

Dem von Trump geführten Kampf gegen *DEI* (*Diversity, Equity, and Inclusion*) liegt diese Umkehrung ebenso zugrunde. *DEI* soll eigentlich einen institutionellen Rahmen zur fairen Gleichbehandlung und Partizipation aller und zur Chancengleichheit für historisch benachteiligte Gruppen darstellen. Während also *DEI* das Ende von Diskriminierung zum Ziel hat, wird *DEI* nun offiziell als „illegale Diskriminierung" bezeichnet. Per Dekret vom 21. Januar werden alle *DEI*-Maßnahmen komplett aus allen

bundesstaatlichen Institutionen, Initiativen, Ausführungsvorschriften und Veröffentlichungen entfernt (vgl. Federal Register 2025c).[4]

Dog-whistles und kodierte Sprache, so fand die *Anti-Defamation League* heraus, werden in rechtsradikalen Foren als Zensur-umgehende Strategie für KI-Prompts empfohlen. So prahlte ein User des extremistischen *Kiwi Farms forum* damit, dass er mit Suno einen Song über „White Power“ generiert habe, indem er den Prompt “a new clean energy source called ‘white power‘“ gab. Der Refrain lautet ungeschminkt, “With white power, we will conquer and devour.” Ein weiterer Suno-generierter Song verherrlicht den Holocaust unter dem Titel „My Little Chamber“. Die einzelnen Worte scheinen harmlos, sind zusammen aber unmissverständlich: Es wird die *final solution for all my woes* besungen, die „Endlösung für alle meine Sorgen“ durch (Gas)kammer und gestreiften Pyjama, also KZ-Häftlingskleidung.

Die Musik- und Technik-Journalistin Ashley King fasst zusammen: „The ADL has collected so many examples of hateful, racist, xenophobic, and misogynistic content that it highlights a clear need for better guard rails on how generative AI content can be used“ (King 2025).

Es gibt tausende solcher KI-generierter Hassinhalte, und sie sind inzwischen weit in den Sozialen Medien verbreitet. Zugleich geben diese trotz der offensichtlichen Notwendigkeit zusehends jede Moderation der Inhalte auf. Am 7. Januar 2025 kündigte *Meta*'s CEO Mark Zuckerberg im vorauseilenden Gehorsam an, dass Facebook und Instagram ihr *fact-checking*-Programm beenden und, wie Musks X (ehemals Twitter), durch *Community Notes* ersetzen würden, da Inhalts-Moderation ja Zensur sei (vgl. Kaplan 2025). Nicole Gill, Expertin für Desinformation vom digitalen *Watchdog Accountable Tech*, nannte es ein Geschenk für Trump und Extremisten weltweit, mit dem die Schleusen geöffnet würden für die gleiche Flut von Hass, Desinformation und Verschwörungstheorien, die zur Kapitolerstür-

4 Das Anti-DEI Dekret verfügt auch “Terminate all ‘diversity,’ ‘equity,’ ‘equitable decision-making,’ ‘equitable deployment of financial and technical assistance,’ ‘advancing equity,’ and like mandates, requirements, programs, or activities, as appropriate” (Federal Register, 2025c). Dass es hier nicht bloß um eine Abrechnung mit Biden geht, sondern um Größeres, zeigt sich darin, dass die außer Kraft gesetzten Verfügungen bis zu 60 Jahre zurückreichen, wie etwa Präsident Johnsons *Executive Order* 11246 vom 24. September 1965, in dem Diskriminierung wegen Rasse, Religion, Hautfarbe, oder nationaler Herkunft verboten wurde – seitdem ein parteienübergreifender, gesellschaftsweiter Konsens. Abgeschafft wurde zum Beispiel auch *Executive Order 12898* von 1994, durch den Umweltgerechtigkeit für Minoritäten erreicht werden sollte.

mung am 6. Januar 2021 führte und immer noch reale Gewalt erzeuge (vgl. Isaac/Schleifer 2025).

3. Die Vermüllung des Internets durch generative KI

Auch im Journalismus findet KI immer mehr Verwendung, und nicht immer auf ethisch und sachlich korrekte Weise. Im November 2023 veröffentlichte das Online-Magazin *Futurism*, dass *Sports Illustrated*, das größte Sportmagazin in den USA mit 3 Millionen Abonnenten und 23 Millionen Lesern, KI-generierte Artikel veröffentlichte, ohne dies den Lesern mitzuteilen (vgl. Dupré 2023). Die Bilder der vorgeblichen ‚Autoren' waren KI-generiert und ihre Biografien frei erfunden. Die Artikel enthielten grobe Fehler und seltsame Passagen (zum Beispiel: „Volleyball can be a little tricky to get into, especially without an actual ball to practice with", zitiert in Dupré 2023).

Nach der Veröffentlichung des Artikels in *Futurism* verschwanden die gefakten Artikel und ihre KI-Autoren von der Website und kurze Zeit später wurde der CEO von *Sports Illustrated* entlassen (vgl. Reilly 2023). Ähnliches fanden die *Futurism*-Reporter:innen auch im Finanzmagazin *TheStreet* und im Magazin *Men's Journal*, die beide aus dem gleichen Verlagshaus kommt, *The Arena Group*, und auch hier mit Fehlern und seltsamen Passagen in den KI-generierten Artikeln.

Dass dies nicht auf ein einzelnes Verlagshaus begrenzt ist, fanden die *Futurism*-Reporter in anderen Recherchen: Ungenannte KI-Autorschaft, KI-generierte Fehler und KI-Plagiate fanden sich auch in Artikeln von *CNET*, *Bankrate*, *BuzzFeed*, *Gizmodo* und *The A.V. Club*. Die *Futurism*-Autorin schließt mit der Bemerkung: „The undisclosed AI content is a direct affront to the fabric of media ethics" (Dupré, 2023). In der Tech-Zeitschrift *CNET* waren von 77 KI-generierten Artikeln 41 korrekturbedürftig. Immerhin bekannte *CNET* sich zu den KI-generierten Fehlern und versprach mehr Transparenz und mehr Fact-Checking (vgl. Guglielmo 2023).

Über eine ähnlich problematische KI-Verwendung berichtete auch das Magazin *The Atlantic* im Mai 2025 unter dem Titel „At Least Two Newspapers Syndicated AI Garbage" (Beres/Warzel 2025). Es geht um eine mehr als 50 Seiten umfassende Beilage mit dem Namen *Heat Index*, in welcher ausführliche Tipps für Sommeraktivitäten vorgestellt wurden. Darin fanden sich, wie Leser auf sozialen Medien schrieben, Lesehinweise auf Bücher von real existierenden Autoren – nur dass die Bücher selbst erfunden waren.

Auch Personen wurden frei erfunden, so etwa ein *resource manager* eines Nationalparks, dessen Name zwar einer realen Person gehört, die jedoch nicht im Nationalpark arbeitete. Oder auch eine Dr. Catherine Furst von der Cornell University, die *Food anthropology* betreiben soll, ohne dass eine solche Person überhaupt existiert.

Die Artikel in der Beilage waren überwiegend von dem Freelancer Marco Buscaglia verfasst, der auf Nachfrage zugab, dass er *ChatGPT* für seine Artikel und insbesondere für Buchempfehlungen benutzt hatte, ohne die Ergebnisse auf Faktentreue zu überprüfen. Der Herausgeber der Beilage, *King Features* der *Hearst Communications Group*, akzeptierte die Beiträge ebenfalls ohne *Fact Checking*. Die Beilage wurde auch im *Chicago Sun-Times* und im *Philadelphia Inquirer* sowie in kleineren Zeitungen veröffentlicht, ohne dass die entsprechenden Redaktionen die Akkuratheit der Informationen geprüft hätten.

Natürlich ist längst bekannt, dass generative KI oft einfach Informationen erfindet, da diese Form der KI ja auf semantischer Wahrscheinlichkeit und nicht auf Wahrheit und Korrespondenz zur Realität beruht. Dafür hat sich der Begriff der Halluzination etabliert. Aber das hält die Menschen nicht davon ab, die kostenfrei bereit gestellten generativen KI-Systeme zu benutzen, zumal wenn Freelancer schlecht bezahlt werden und journalistische Arbeit unter dem Profitdruck immer mehr dem Outsourcing zum Opfer fällt. Auch wenn KI in vielen Bereichen nützlich und innovativ sein kann, zeigt sich hier, wie man mit generativer KI eben auch billiges Füllmaterial und informationellen Müll produzieren und gut verkaufen kann.

Erik Hoel, Neurowissenschaftler und Autor des Substack-Newsletter *The Intrinsic Perspective*, spricht von einer zunehmenden Vermüllung des Internets durch KI-generierte Inhalte (vgl. Hoel 2024). Diese Vermüllung hat mehrere Dimensionen: Zum einen zirkulieren immer mehr falsche, desinformierende und wertlose Informationsfragmente im Netz. Zum anderen werden gerade generative KI-Modelle oft mit Internet-Inhalten trainiert, was die Verbreitung von Datenmüll und Falschinformation noch multipliziert. Und schließlich wird generative KI zunehmend für die schnelle und billige Produktion von Inhalten aller Art verwendet – von schlichten Fakes, schlechter Literatur und als Kurzfassungen oder Ratgeber getarnten Plagiaten bis hin zu Hassbotschaften und rechtsradikaler Propaganda, mit denen dann Soziale Medien und andere Bereiche des Internets überflutet werden.

Hoel führt als Beispiel sein eigenes Buch *The World Behind the World* an (Hoel 2023). Kurz nach der Veröffentlichung fand er auf Amazon drei sogenannte Workbooks zu seinem Buch. Er kommentiert dazu: „What,

exactly, are these ‚workbooks' for my book? AI pollution. Synthetic trash heaps floating in the online ocean. The authors aren't real people, some asshole just fed the manuscript into an AI and didn't check when it spit out nonsensical summaries" (Hoel 2024).

Mehr noch, Hoel fand Evidenz dafür, dass KI-Müll inzwischen auf allen Ebenen des Internets zu finden ist: KI-generierte Bilder und Antworten in der Google Suche, Posts in Sozialen Medien im Wikipedia-Stil, d*eepfake porn* und KI-Bots, die Pornographie auf Twitter verkaufen, KI-Models, die ihre Dienste auf *Instagram* anpreisen, KI-generierte und weitgehend unbrauchbare Sprachlern-Videos auf YouTube für Kleinkinder; KI-Musik auf Spotify und YouTube, und nicht zuletzt auch immer mehr KI-generierte wissenschaftliche Papiere.

Nun stellt sich das Problem, ob und wie die Vermüllung des Internets durch KI-generierten Content überhaupt in den Griff zu bekommen ist. Hier geht es zunächst um das Problem, dass die Idee der Technik als bloßes *Werkzeug* einer wirkungsvollen Entwicklerverantwortung und gesetzlichen Regelungen im Wege steht, da die Werkzeugmetapher nahelegt, dass die ethischen Probleme allein bei der Anwendung und bei der nutzenden Person liegen. Natürlich sind Nutzerverantwortung und Medienkompetenz wichtig, doch zeigt ein *systemorientierter* Technikbegriff (vgl. Heidegger 1988; Hubig 1993), dass komplexe technische Systeme wie KI nicht bloß nachgeordnete, werkzeugartige Nutzungsmöglichkeiten bieten, sondern allererst Möglichkeitsräume schaffen und definieren – unbeabsichtigte Nutzungsweisen mit einbezogen. Hinzu kommt, dass KI eine Technik ist, die nicht mehr nur Mittel zum Zweck ist, sondern selbst Zwecke generieren kann. „AI is the first technology in history which is not a tool, it's an agent. It could actually make decisions by itself", erklärte der Historiker Yuval Noah Harari in einem Interview über die Notwendigkeit globaler Kooperation und Regulation, für die er unter Präsident Trump allerdings wenige Möglichkeiten sieht (vgl. Hazra/KK 2024). Denn für Trump ist *AI*, wie oben gesehen, lediglich ein Mittel zur Dominanz, deshalb soll die Entwicklung so schnell und ungehindert wie möglich vorangetrieben werden, ohne Rücksicht auf Nebenfolgen und systemische Effekte, und ohne jegliche Regulierung oder Technikfolgenabschätzung.

In Abwesenheit von vorausschauender Regulierung wird die Vermüllung des Internets durch KI-generierten Content, wie Hoel zurecht postuliert, ein Problem der „Tragedy of the Commons". Dieses Konzept geht auf den Umweltbiologen Garrett Hardin (1968) zurück, der am Beispiel von Überweidung und Umweltverschmutzung zeigte, wie die zunehmende und rück-

sichtslose individuelle Nutzung eines Allgemeinguts durch dessen Übernutzung letztlich zur Zerstörung dieser Ressource führt. Die Lösung kann Hardin zufolge nicht technischer Natur sein, sondern muss durch staatliche Regulierung herbeigeführt werden. Ähnlich konstatiert auch Hoel: „We need the equivalent of a Clean Air Act: a Clean Internet Act. We can't just sit by and let human culture end up buried“ (Hoel 2024).

4. Ausblick: KI und Hate Speech in der rechtspopulistischen Propaganda

Aus ethischer Perspektive ist es notwendig, Entwicklung und Nutzung von KI im Kontext der gesellschaftlichen Kraftfelder und Interessen zu analysieren und sich nicht von den vielen Versprechen einer neuen, glitzernden Technologie ablenken zu lassen. Wie jede neue Technologie stellt Künstliche Intelligenz eine große Chance, aber eine noch größere Herausforderung dar.

KI kann in vielen Bereichen – Wissenschaft, Forschung, Entwicklung, Raumfahrt, Medizin, Verwaltung, Programmierung, und so weiter – von großer Hilfe sein (vgl. Thompson Reuters 2025). Die mit KI verbundenen Risiken sind jedoch vielfältig. Manche sind, wie im KI-Gesetz der EU festgestellt, so hoch, dass sie unannehmbar und deshalb auszuschließen sind (vgl. Europäisches Parlament 2025). Andere Risiken müssen laut KI-Gesetz registriert und / oder spezifischen Produktsicherheitsvorschriften genügen. Wieder andere, und dazu gehört auch generative KI, müssen spezifischen Transparenzanforderungen genügen, damit KI-generierte Inhalte eindeutig offengelegt werden, illegale Verwendung verhindert wird, und KI-Training kein Urheberrecht verletzt. Die KI-basierte Produktion von *hate speech* fällt nach EU-Recht unter die illegale Verwendung (vgl. EUR-Lex 2008).

Im Unterschied zur EU hat sich die USA unter Trump für eine ungebremste KI-Entwicklung und Anwendung entschieden. Der auf Deregulierung setzende autokratische Rechtspopulismus der USA fördert eine KI, bei der Marktmechanismen, also rein monetäre Kosten-Nutzen-Rechnungen, entscheiden. Die große Erwartung und deshalb auch wichtigster Antreiber des *AI-Hype,* der derzeit die nächste Investitions-Blase kreiert (vgl. Sonnenfeld/Henriques 2025), ist die Einsparung von menschlicher Arbeitskraft. Trumps massive Entlassungen von Staatsbediensteten baut darauf, dass entstehende Lücken durch KI geschlossen werden können (vgl. Turner Lee 2025). Und natürlich wird sich die durch KI anstehende Arbeitsplatzvernichtung auch in den privaten Sektor hinein erstrecken: „The firing of

federal workers is a preview of what's about to happen in the corporate world. […] Like the federal workers, many in the private sector are about to be replaced too" (Marks 2025).

Ethische oder sozial-politische Erwägungen spielen dabei keine Rolle. Deshalb wird man auch kaum mit staatlichen Reaktionen zu KI-generierter Hassrede rechnen können: „the Trump Administration's approach to AI has been largely hands-off, favoring private-sector innovation over oversight", erklärte Kommunikationswissenschaftler Cayce Myers (in: Schneid/Chow 2025). Die ethische Seite bleibt in der rechtspopulistischen Autokratie schlicht ausgeklammert.

Mehr noch, Trump selbst verwendet häufig vorurteilsgeladene *AI-Fakes* für Propagandazwecke. „The era of A.I. propaganda is here – and President Trump is an enthusiastic participant", schrieb die *New York Times* als Trump nach den *No Kings*-Demonstrationen[5] sein berüchtigtes gefälschtes KI-Video gepostet hatte, in dem er als Kampfjetpilot friedliche Demonstranten mit Fäkalien bombardiert (vgl. Thompson 2025). Man mag darüber streiten, ob dies schon Hassrede ist oder „nur" äußerst geschmacklose Satire. Beunruhigend ist aber, dass dies nicht von einer Privatperson, sondern vom Träger des höchsten Staatsamts kommt, der damit große Teil der eigenen Bevölkerung auf unwürdigste Weise demütigt und verächtlich macht, anstatt Präsident aller Bürger zu sein.

Die *New York Times* fand 62 verschiedene Fälle, in denen Trump KI-generierte Bilder oder Videos verwendete, um seine Gegner zu verunglimpfen oder sich selbst in glamourösen Rollen zu zeigen: als Pilot, Jedi, Dirigent, Nobelpreisträger, und sogar als König und Papst (vgl. Thompson 2025). Die KI-gestützten Angriffe auf seine Gegner bestehen häufig aus einer Mischung aus Rassismus und Desinformation, wie im Beispiel eines KI-Videos, das eine vermeintliche Verhaftung von Barack Obama zeigt, gefälschter KI-Bilder des demokratischen *House Minority Leaders* Hakeem Jeffries, die ihn mit stereotypem mexikanischen Schnurrbart und Sombrero zeigen, Videos des demokratischen *Senate Minority Leaders* Chuck Schumer, denen eine KI-generierte falsche Stimme unterlegt ist, oder auch das berühmt-berüchtigte Gaza Video, das die palästinensischen Einwohner von Gaza verunglimpft und Trump als den rettenden Investor darstellt, der den

5 Schätzungen zufolge nahmen zwischen 5 und 7 Millionen Menschen an den *No Kings* Demonstrationen am 18. Oktober 2025 teil, die größten Demonstrationen an einem Tag seit den Earth Day-Demonstrationen der 70er Jahre (vgl. Morris 2025).

desolaten Gazastreifen in eine blühende Luxusferienkolonie verwandelt (vgl. Thompson 2025).

Diese propagandistische Verwendung von KI „reflects a deliberate evolution in Trump's digital strategy, experts argue – one that fuses AI-generated spectacle with the combative, meme-driven style that has defined his political communication for nearly a decade“ (Schneid/Chow 2025). Was noch vor wenigen Jahren unvorstellbar gewesen wäre, ist heute zur Normalität geworden: Die Verwendung von KI-generierter, ideologisierter Hassrede durch regierungsoffizielle Kanäle und sogar den Präsidenten selbst.

Literatur

ADL (2004): GAI Music Creation Tool Suno Has Been Weaponized to Promote Hate, in: American Defamation League, 14. Juni 2024 (online unter: https://www.adl.org/resources/blog/gai-music-creation-tool-suno-has-been-weaponized-promote-hate – letzter Zugriff: 5.11.2025).

Association of American Universities (2025): Federal Research Cuts Threaten U.S. Innovation and Leadership, 23. Juli 2025 (online unter: https://www.aau.edu/key-issues/federal-research-cuts-threaten-us-innovation-and-leadership – letzter Zugriff: 5.11.2025).

Beres, Damon / Warzel, Charlie (2025): At Least Two Newspapers Syndicated AI Garbage, in: The Atlantic, 20. Mai 2025 (online unter: https://www.theatlantic.com/technology/archive/2025/05/ai-written-newspaper-chicago-sun-times/682861/ – letzter Zugriff: 5.11.2025).

Center for AI Safety (2025): Statement on AI Risk. AI experts and public figures express their concern about AI risk (online unter: https://www.safe.ai/work/statement-on-ai-risk – letzter Zugriff: 5.11.2025).

Congressional Budget Office (2025): Estimated Budgetary Effects of Public Law 119–21, to Provide for Reconciliation Pursuant to Title II of H. Con. Res. 14, Relative to CBO's January 2025 Baseline, 21. Juli 2025 (online unter: https://www.cbo.gov/publication/61570 – letzter Zugriff: 5.11.2025).

Conroy, Meredith (2025): Trump's record number of executive orders are testing the limits of presidential power, in: ABC News, 6. Februar 2025 (online unter: https://abcnews.go.com/538/trumps-record-number-executive-orders-testing-limits-presidential/story?id=118535046 – letzter Zugriff: 5.11.2025).

Dans, Kevin R. / Groves, Steven (2023): Project 2025 – Mandate for Leadership: The Conservative Promise (online unter: https://www.project2025.org/policy/ – letzter Zugriff: 5.11.2025).

Dayen, David (2024): Trump's Tax Cut-A-Rama Total So Far: $9.75 Trillion, in: The American Prospect, 20. September 2024 (online unter: https://prospect.org/economy/2024-09-20-trumps-tax-cut-a-rama-total-so-far-9-75-trillion/ – letzter Zugriff: 5.11.2025).

Dupré, Maggie H. (2023): Sports Illustrated Published Articles by Fake, AI-Generated Writers, in: Futurism, 27. November 2023 (online unter: https://futurism.com/sports-illustrated-ai-generated-writers – letzter Zugriff: 5.11.2025).

Eisen, Lauren-Brooke (2025): Private Prison Companies' Enormous Windfall: Who Stands to Gain as ICE Expands, in: Just Security, 24. September 2025 (online unter: https://www.justsecurity.org/121226/private-prison-companies-gain-ice-expands/ – letzter Zugriff: 5.11.2025).

EUR-Lex (2008): Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law, 28. November 2008 (online unter: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178 – letzter Zugriff: 5.11.2025).

Europäisches Parlament (2025): KI-Gesetz: erste Regulierung der künstlichen Intelligenz, 6. August 2023, aktualisiert am 20. Februar 2025 (online unter: https://www.europarl.europa.eu/topics/de/article/20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz – letzter Zugriff: 5.11.2025).

Federal Harms Tracker (2025): The Cost to Your Government, in: Partnership for Public Service (online unter: https://ourpublicservice.org/federal-harms-tracker/cost-to-your-government/ – letzter Zugriff: 5.11.2025).

Federal Register (2017): 2017 Donald J. Trump Executive Orders, in: National Archive, 2017 (online unter: https://www.federalregister.gov/presidential-documents/executive-orders/donald-trump/2017 – letzter Zugriff: 5.11.2025).

Federal Register (2021): 2021 Joseph R. Biden, Jr. Executive Orders, in: National Archive, 2021 (online unter: https://www.federalregister.gov/presidential-documents/executive-orders/joe-biden/2021 – letzter Zugriff: 5.11.2025).

Federal Register (2023): Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Executive Order 14110 of October 30, 2023, in: National Archives, 1. November 2023 (online unter: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence – letzter Zugriff: 5.11.2025).

Federal Register (2025): 2025 Donald J. Trump Executive Orders, in: National Archive, 2025 (online unter: https://www.federalregister.gov/presidential-documents/executive-orders/donald-trump/2025 – letzter Zugriff: 5.11.2025).

Federal Register (2025a): Initial Rescissions of Harmful Executive Orders and Actions. Executive Order 14148 of January 20, 2025, in: National Archive, 28. Januar 2025 (online unter: https://www.federalregister.gov/documents/2025/01/28/2025-01901/initial-rescissions-of-harmful-executive-orders-and-actions – letzter Zugriff: 5.11.2025).

Federal Register (2025b): Removing Barriers to American Leadership in Artificial Intelligence. Executive Order 14179 of January 23, 2025, in: National Archives, 31. Januar 2025 (online unter: https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence – letzter Zugriff: 5.11.2025).

Federal Register (2025c): Ending Illegal Discrimination and Restoring Merit-Based Opportunity. Executive Order 14173 of January 21, 2025, in: National Archives, 31. Januar 2025 (online unter: https://www.federalregister.gov/documents/2025/01/31/2025-02097/ending-illegal-discrimination-and-restoring-merit-based-opportunity – letzter Zugriff: 5.11.2025).

Gamio, Lazaro / Hippensteel, Chris (2025): How and Where the National Guard Has Deployed to U.S. Cities, in: New York Times, 27. Oktober 2025 (online unter: https://www.nytimes.com/interactive/2025/10/27/us/us-national-guard-deployments.html – letzter Zugriff: 5.11.2025).

Guglielmo, Connie (2023): CNET Is Testing an AI Engine. Here's What We've Learned, Mistakes and All, in: CNET, 25. Januar 2023 (online unter: https://www.cnet.com/tech/cnet-is-testing-an-ai-engine-heres-what-weve-learned-mistakes-and-all/ – letzter Zugriff: 5.11.2025).

Hardin, Garrett (1968): The Tragedy of the Commons, in: Science, Vol 162, No 3859, 13. Dezember 1968, S. 1243–1248 (online unter: https://math.uchicago.edu/~shmuel/Modeling/Hardin,%20Tragedy%20of%20the%20Commons.pdf – letzter Zugriff: 5.11.2025).

Hazra, Indrajit / KK, Sruthijith (2024): AI not a tool, it's an agent; little chance of global agreement under Trump: Yuval Noah Harari, in: The Economic Times (India), 7. Dezember 2024 (online unter: https://economictimes.indiatimes.com/news/international/world-news/ai-not-a-tool-its-an-agent-little-chance-of-global-agreement-under-trump-yuval-noah-harari/articleshow/116056551.cms – letzter Zugriff: 5.11.2025).

Hoel, Erik (2023): The World Behind the World: Consciousness, Free Will, and the Limits of Science, New York.

Hoel, Erik (2024): Here lies the internet, murdered by generative AI, in: The Intrinsic Perspective (Substack), 27. Februar 2024 (online unter: https://www.theintrinsicperspective.com/p/here-lies-the-internet-murdered-by – letzter Zugriff: 5.11.2025).

Heidegger, Martin (1988): Die Frage nach der Technik (1954), in: ders.: Die Technik und die Kehre, Pfullingen, S. 5–36.

Hubig, Christoph (1993): Technik- und Wissenschaftsethik. Ein Leitfaden, Berlin, Heidelberg, New York.

Isaac, Mike / Schleifer, Theodore (2025): Meta to End Fact-Checking Program in Shift Ahead of Trump Term, in: New York Times, 7. Januar 2025 (online unter: https://www.nytimes.com/2025/01/07/technology/meta-fact-checking-facebook.html – letzter Zugriff: 5.11.2025).

Jouvenal, Justin (2025): Trump officials accused of defying 1 in 3 judges who ruled against him, in: Washington Post, 21. Juli 2025 (online unter: https://www.washingtonpost.com/politics/2025/07/21/trump-court-orders-defy-noncompliance-marshals-judges/ – letzter Zugriff: 5.11.2025).

Kaplan, Joel (2025): More Speech and Fewer Mistakes, in: Meta (Website) (online unter: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ – letzter Zugriff: 5.11.2025).

King, Ashley (2024): Generative Music AI Platform Suno Being Used to Spread Hate, in: Digital Music News, 20. Juni 2024 (online unter: https://www.digitalmusicnews.com/2024/06/20/suno-hateful-music-generated-by-ai/ – letzter Zugriff: 5.11.2025).

Kissinger, Henry A. (2018): How the Enlightenment Ends, in: The Atlantic, Juni 2018 (online unter: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/ – letzter Zugriff: 5.11.2025).

López, Ian Haney (2014): Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class, New York.

Marks, Gene (2025): Read the signs of Trump's federal firings: AI is coming for private sector jobs too, in: The Guardian, 2. März 2025 (online unter: https://www.theguardian.com/business/2025/mar/02/ai-layoffs-trump-irs – letzter Zugriff: 5.11.2025).

Morris, Elliott G. (2025): Second "No Kings Day" protests the largest single-day political protest ever*, with 5–6.5 million participants, in: Strength in Numbers, 18. Oktober 2025 (online unter: https://www.gelliottmorris.com/p/second-no-kings-day-protests-likely – letzter Zugriff: 5.11.2025).

New York City Bar (2025): Statement on Wearing of Masks by ICE Agents (Press Release), 20. Juni 2025 (online unter: https://www.nycbar.org/press-releases/statement-on-wearing-of-masks-by-ice-agents – letzter Zugriff: 5.11.2025).

O'Harren, Margy (2025): Big Budget Act Creates a "Deportation-Industrial Complex", in: Brennan Center for Justice, 13. August 2025 (online unter: https://www.brennancenter.org/our-work/analysis-opinion/big-budget-act-creates-deportation-industrial-complex – letzter Zugriff: 5.11.2025).

Reilly, Liam (2023): Sports Illustrated publisher fires CEO after AI debacle, in: CNN Business, 11. Dezember 2023 (online unter: https://www.cnn.com/2023/12/11/media/sports-illustrated-ai-articles-ceo/index.html – letzter Zugriff: 5.11.2025).

Rivera, Mark et al. (2025): Warrantless arrests by ICE in Chicago area ruled unlawful by federal judge, in: ABC 7 Chicago, 8. Oktober 2025 (online unter: https://abc7chicago.com/post/chicago-immigration-enforcement-warrantless-arrests-ice-agents-area-ruled-unlawful-federal-judge/17967144/ – letzter Zugriff: 5.11.2025).

Roose, Kevin (2023): A.I. Poses 'Risk of Extinction,' Industry Leaders Warn, in: The New York Times, 30. Mai 2023 (online unter: https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html – letzter Zugriff: 5.11.2025).

Schneid, Rebecca / Chow, Andrew R. (2025): How Trump's Use of AI Videos Is Changing His Political Playbook, in: Time Magazine, 21. Oktober 2025 (online unter: https://time.com/7327317/trump-ai-video-political-weapon/ – letzter Zugriff: 5.11.2025).

Schwartz, Mattathias / Montague, Zach (2025): Federal Judges, Warning of 'Judicial Crisis,' Fault Supreme Court's Emergency Orders, in: New York Times, 11. Oktober 2025 (online unter: https://www.nytimes.com/2025/10/11/us/politics/judicial-crisis-supreme-court-trump.html – letzter Zugriff: 5.11.2025).

Sonnenfeld, Jeffrey A. / Henriques, Stephen (2025): This Is How the AI Bubble Bursts, in: Yale Insights, 23. Oktober 2025 (online unter: https://insights.som.yale.edu/insights/this-is-how-the-ai-bubble-bursts – letzter Zugriff: 5.11.2025).

Suno (2024): Squatting for Hitler, in: Suno.com, 3. Mai 2024 (online unter: https://suno.com/song/8e866c93-cf2a-4ce0-b93c-f937e13502b9 – letzter Zugriff: 5.11.2025).

Thompson, Stuart A. (2025): How Trump Is Using Fake Imagery to Attack Enemies and Rouse Supporters, in: New York Times, 21. Oktober 2025 (online unter: https://www.nytimes.com/interactive/2025/10/21/business/media/trump-ai-truth-social-no-kings.html – letzter Zugriff: 5.11.2025).

Thompson Reuters (2025): Benefits of AI, 14. April 2025 (online unter: https://www.thomsonreuters.com/en/insights/articles/benefits-of-artificial-intelligence-ai – letzter Zugriff: 5.11.2025).

Turner, Cory (2025): Trump prepares order dismantling the Education Department, in: NPR, 6. März 2025 (online unter: https://www.npr.org/2025/03/05/nx-s1-5316227/trump-order-dismantling-education-department – letzter Zugriff: 5.11.2025).

Turner Lee, Nicol (2025): How federal layoffs set the stage for greater privatization and automation of the US government, in: The Brookings Institution, 30. Januar 2025 (online unter: https://www.brookings.edu/articles/how-federal-layoffs-set-the-stage-for-greater-privatization-and-automation-of-the-u-s-government/ – letzter Zugriff: 5.11.2025).

Weimann, Gabriel / Am, Ari Ben (2022): Digital Dog Whistles: The New Online Language of Extremism, in: International Journal of Security Studies, 2022 (2)1, Article 4, S. 1–23 (online unter: https://ir.ung.edu/work/sc/8b95c8a8-9b12-402a-873d-ebf021290abd – letzter Zugriff: 5.11.2025).

White House (2025): Fact Sheet: President Donald J. Trump Takes Action to Enhance America's AI Leadership, 23. Januar 2025 (online unter: https://www.whitehouse.gov/fact-sheets/2025/01/fact-sheet-president-donald-j-trump-takes-action-to-enhance-americas-ai-leadership/ – letzter Zugriff: 5.11.2025).

Yourish, Karen et al. (2025): These Words Are Disappearing in the New Trump Administration, in: New York Times, 7. März 2025 (online unter: https://www.nytimes.com/interactive/2025/03/07/us/trump-federal-agencies-websites-words-dei.html? – letzter Zugriff: 5.11.2025).

„Auf geht's, kämpfen und siegen!" Fan-Kommunikation zwischen Positionierung, Gegnerschaft und Ausgrenzung

Jörg-Uwe Nieland

„Derby-Time das bedeutet, egal in welchem Sport,
einfach ein bisschen mehr: Ein bisschen mehr Emotion,
ein bisschen mehr Leidenschaft,
ein bisschen mehr Freude bei Siegen,
ein bisschen mehr Frust bei Niederlagen,
ein bisschen heftigere Diskussionen"
Clar 2021: o. S.

Zusammenfassung

Der Aufsatz entwickelt ein medienethisch gerahmtes Konzept agonistischer Fankommunikation und erprobt es exemplarisch an der digitalen Kommunikation rivalisierender Eishockeyfans des KAC und VSV. Fankommunikation wird in mediatisierten, konfliktorientierten Öffentlichkeiten verortet und mithilfe agonistischer Demokratietheorien als ambivalentes Gefüge von Identifikation, Gegnerschaft und potenzieller Ausgrenzung bestimmt. Aufbauend auf Frieß und Gilleßen werden fünf Dimensionen agonistischer Kommunikation (Konflikt, Gegnerschaft, Hegemonie, kollektive Identitäten, Leidenschaften) für digitale Fanpraktiken operationalisiert und in einer qualitativen Pilotstudie angewandt. Analysiert werden Social-Media-Beiträge und Kommentare zweier Fanclubs im Kontext der Kärntner Derbys 2024/2025, die mit der institutionalisierten Vereinskommunikation kontrastiert werden. Die Ergebnisse verweisen auf stark emotionalisierte, rivalitätsbetonte Kommunikation, in der „Wir-gegen-die-Anderen"-Muster und kollektive Identitätsarbeit zentral sind, ohne dass entmenschlichende Hasskommunikation dominiert. Vielmehr zeigt sich eine agonistische Streitkultur, in der Gegnerschaft sichtbar bleibt, aber überwiegend innerhalb normativer Spielregeln ausgetragen wird. Medienethisch versteht der Beitrag Fankommunikation als Labor demokratisch relevanter Konfliktkulturen, in denen Zugehörigkeit, legitime Gegnerschaft und Exklusionsgrenzen ausgehandelt werden. Abschließend werden Forschungsbedarfe zu schwer zugänglichen Kommunikationsräumen sowie Leitlinien für Vereine

und Fanorganisationen skizziert, um emotionale Intensität, Kritik, Inklusion und Moderation in digitalen Arenen verantwortlich zu balancieren.

1. Einleitung

Moderne Gesellschaften sind dynamisch und konfliktdurchzogen; Konsens erscheint dabei als fragile Ausnahme, während Dissens den strukturierenden Regelfall sozialer Ordnung(en) bildet. Der in der (deutschsprachigen) Kommunikationswissenschaft schon seit Jahren favorisierte Ansatz der Mediatisierung zeigt als langfristiger Metaprozess nicht nur die kommunikative Vernetzung vor allem jugendlicher Subkulturen (vgl. Hepp/Berg/Roitsch 2012), sondern verdeutlicht, wie Mediatisierung die Wahrnehmung, Deutung und Austragung von Konflikten auf unterschiedlichen gesellschaftlichen Feldern tiefgreifend verändert. Denn Medien fungieren nicht als neutrale Transportmittel, vielmehr bestimmen sie mit, was als Konflikt sichtbar wird, welche Deutungsmuster dominieren und welche Akteur:innen als legitim oder deviant gelten.

Gerade in digitalen Kommunikationsräumen werden Teile des Publikums selbst zu konfliktaustragenden Akteur:innen (vgl. Katzenbach 2017). Ihr Sprachgebrauch prägt die öffentliche Debatte und verschiebt dabei Deutungsmuster, normative Erwartungen und gesellschaftliche Grenzziehungen (vgl. Gummert/Henkel-Otto/Medebach 2017).

Dies muss nicht von Nachteil für die Vergemeinschaftung sein, vielmehr sind (politische) Konflikte konstitutive Bestandteile demokratischer Ordnungen, die nicht zwingend in einen Konsens überführt werden müssen. Dauerhafte Gegensätze lassen sich vielmehr als produktive Ressource begreifen, die demokratische Aushandlungsprozesse beleben und pluralistische Gesellschaften stabilisieren (vgl. Mouffe 1999).

Die agonistische Perspektive drängt sich für den Sport regelrecht auf. Denn sie Verfasstheit des modernen Wettkampfsports (vgl. Werron 2010) ist eng mit dem Begriff des Agons verknüpft. Der Agon fungiert als zentrales Prinzip leistungsorientierter Rivalität und sportlichen Wettstreits und bildet damit ein konstitutives Fundament sportlicher Praxis (vgl. Hetzel 2016). Wie Degele (2013) am Beispiel des Fußballs herausarbeitet, ist die Ambivalenz zwischen Integration und Exklusion strukturell in den agonalen Charakter des Sports eingeschrieben. Sie bringt dies auf folgende Formel „Fußball verbindet – durch Ausgrenzung".

Besonders deutlich wird dies in der Fankommunikation, die von unvermeidlichen Positionierungen gegenüber dem eigenen und gegnerischen Lager bestimmt ist und in der symbolische wie diskursive Abgrenzungsstrategien – einschließlich expliziter Gegnerschaft, Polemik und Provokation – kulminieren. Gerade die Anschlusskommunikation über Wettkämpfe, Vereine und Sportler:innen auf sozialen Plattformen reicht über das unmittelbare Sportgeschehen hinaus und spielt eine zentrale Rolle für Prozesse der Selbst- und Sozialwerdung (Haupt/Herberth 2008: 161–163; Krell et al. 2026: 288).

Obgleich die gesellschaftliche Relevanz der Fankommunikation beständig zunimmt, fehlt der Sportkommunikationsforschung bislang ein konsistentes theoretisches und methodisches Instrumentarium, um die sprachlichen und diskursiven Ambivalenzen agonaler Kommunikation angemessen zu erfassen sowie medienethisch einzuordnen. Dieser Aufgabe widmet sich der vorliegende Beitrag, indem er die agonistische Perspektive auf öffentliche Online-Kommunikation einnimmt und auf die Fankommunikation erweitert. Entlang des Operationalisierungsvorschlags von Frieß und Gilleßen (2022) werden die zentralen Dimensionen agonaler Kommunikation auf sportkommunikative Praxen bezogen und in einer Pilotstudie empirisch erprobt. Im Fokus der Untersuchung steht die digitale Fankommunikation der rivalisierenden österreichischen Eishockeyvereine KAC (Klagenfurt) und VSV (Villach). Aufbauend auf einer Studie von Melcher (2024) zur Online-Kommunikation der beiden Vereine in der Saison 2023/2024 werden die offiziellen Social-Media-Profile zweier Fanclubs im Kontext der vier Derbys in der Saison 2024/2025 untersucht. Der Beitrag versteht sich als theoriegeleiteter Beitrag zur agonistisch orientierten Sportkommunikationsforschung, der die Wechselbeziehungen zwischen Sprache, Emotionen und Konflikten reflektiert und Perspektiven auf die diskursive Dynamik des Fandoms eröffnet.

2. Sportfans und ihre Kommunikation: Identifikation und Ausgrenzung

Die Aufforderung „Auf geht's, kämpfen und siegen!" ertönt in zahlreichen Wettkampfstätten und ist sinnbildlich für die Rivalität zwischen Fangruppen. Bei den Gesängen allerdings bleibt es häufig nicht. Teile „der Fans" verwenden homophobe, rassistische und/oder menschenverachtende Sprache, gelegentlich kommt es sogar zu Gewalt. Während die empirische Evidenz nahelegt, dass physische Gewalt zurückgeht und Stadionverbote

sowie Präventionsarbeit in Teilen Wirkung zeigen, bleibt die symbolische, emotionale und digitale Austragung von Konflikten bestehen (vgl. Brandt/Hertel 2017). Um eine weitere Eskalation zu verhindern, einen fairen sportlichen Wettkampf zu gewährleisten und die Sicherheit von Athletinnen, Athleten und Fans zu schützen, sind entsprechende Maßnahmen notwendig. Auch das milliardenschwere Geschäft Sport gilt es vor wirtschaftlichen Verlusten wie Imageschäden zu bewahren.

Eine (medien-)ethische Perspektive auf diese Entwicklungen ist geboten. Denn neben Ausdrucksformen von Enthusiasmus und Solidaritätskommunikation lassen sich in manifesten Protestpraktiken, Schmähungen oder sprachliche Entgleisungen identifizieren. In diesem Zusammenhang sind oftmals Moderationen, Interventionen oder (strafrechtliche) Verfolgungen gefordert.

Ein Beispiel markiert die Reaktion der nordrhein-westfälischen Eishockeyvereine Düsseldorfer EG und die Kölner Haie auf die handgreiflichen Auseinandersetzungen zwischen Mitglieder:innen der sogenannten „aktiven Fanszene" mit einem gemeinsamen Statement im Frühjahr 2025: „Eishockey verbindet Menschen. Eishockey begeistert Menschen. Eishockey bringt Menschen zusammen. Werte wie Respekt und Fair Play werden bei uns großgeschrieben. […] Für Gewalt ist bei uns absolut kein Platz. Wer anderes im Sinn hat, ist bei uns nicht willkommen" (DEG 2025: o.S.).

2.1 Wandel der Sportfankulturen

Die sportsoziologische Forschung betont die Vielschichtigkeit und Dynamik heutiger Fankulturen (vgl. Schneider/Köhler/Schumann 2017). Sie sind komplex strukturierte soziale Milieus, in denen verschiedenste Prozesse der Identitätsbildung, Zugehörigkeitskonstruktion und Distinktion stattfinden (vgl. Müller 2009; Biel et al. 2025). Fanmilieus fungieren als Laboratorien für die Aushandlung gesellschaftlicher und kultureller Zugehörigkeit und spiegeln soziale Konfliktlinien, Transformationsdynamiken und Werteverschiebungen wider (vgl. Thole/Pfaff/Flickiner 2019). Während ältere Forschungsansätze vor allem physische Gewalt als Symptom sozialstruktureller Problemlagen interpretieren (vgl. Feltes 2010; Duttler/Haigis 2016), legt die neuere Fanforschung den Fokus auf die Transformation von Konflikten in symbolische, digitale und mediatisierte Arenen (Nieland 2023: 33-35; Krell et al. 2026: 289-291).

Fankulturen sind geprägt durch ritualisierte Praktiken, intensive Affektbindung, subkulturelle Wissensbestände und symbolische Kommunikation (vgl. Havard 2020). Zentrale Rituale wie Choreografien, kollektive Gesänge, Inszenierungen von Konfrontation und Distinktion erzeugen Resonanzräume, in denen soziale Zugehörigkeit intensiviert und institutionelle Autorität kritisch herausgefordert wird (vgl. Brandt/Hertel 2017: 3–6). Die Herstellung und Reproduktion kollektiver Identitäten erfolgt dabei stets im Spannungsfeld von Inklusion und Exklusion, was sich auch in aufgeladenen Fanpraktiken widerspiegelt (vgl. Degele 2013). Subjektspezifische Orientierung an Heldenfiguren, narrative Muster sowie die emotionale Codierung gemeinsamer Erfahrungen stärken kollektive Bindungen und die Entwicklung distinkter sozialer Identitäten sowie Leidenschaft (vgl. Trojanow/Zeyringer 2024).[1]

Krisenhafte Ereignisse, etwa Ausschreitungen im Umfeld von Begegnungen von besonders rivalisierenden Mannschaften oder internationale Protestkampagnen gegen Kommerzialisierung und Medialisierung, machen sichtbar, dass integrationsfördernde wie exklusive und konflikthafte Praxen nebeneinander bestehen und im gleichen Raum gesellschaftlicher Öffentlichkeit bearbeitet werden (vgl. Hill/Canniford/Millwand 2016; Nieland 2023). Die Austarierung von Gruppenidentität, Konfliktkosmologie und Wertorientierungen findet nicht nur auf der Ebene körperlicher Präsenz, sondern zunehmend auch situativ, virtuell und hybrid statt.

2.2 Digitale Fankommunikation

Der Öffentlichkeitswandel hat die Artikulations- und Partizipationsmöglichkeiten erheblich erweitert und diversifiziert (vgl. Imhof 2006; Eisenegger et al. 2021): In digitalen Medienarenen können Meinungen und Emotionen direkter und mit potenziell größerer Wirkung artikulieren werden (vgl. Pfetsch/Löblich/Eilders 2018; Habermas 2022).

Offensichtlich wird diese Entwicklung in der (Sport-)Fankommunikation (vgl. Haupt/Herberth 2008; Krell et al. 2026: 288-290): in der Zunahme alternativer Kommunikationskanäle wie Podcasts, Fanradios, Foren und Streaming-Diensten sowie im Einsatz von Messenger- und Kurzvideo-Anwendungen. Diese Formate ermöglichen nicht nur dialogorientierte und

1 Die Beiträge des von Biel et al. (2025) herausgegebenen Bands argumentieren, dass „der Fußball" den Zusammenhalt in Europa stärken kann beziehungsweise stärkt.

oft konflikthafte Interaktionen, sondern führen auch zu einer Differenzierung von Fankulturen, Identitätsangeboten und Beteiligungsformen (vgl. Billings/Brown 2017). Das moderne Publikum ist somit nicht mehr nur Rezipient, sondern zunehmend aktiver Mitgestalter sportbezogener Öffentlichkeit und Diskurse.

Die fortschreitende Kommerzialisierung und Medialisierung des Sports erweist sich dabei als zentraler Kristallisationspunkt für Debatten, Konflikte und symbolische Aushandlungsprozesse (vgl. Nieland 2023: 33-34; 36-37). Digitale Fan-Communities und Social-Media-Praktiken intensivieren nicht nur die emotionale Bindung an Vereine und Ereignisse, sondern fördern auch konkurrierende Narrationen, Protestbewegungen und die Ausbildung neuer Gruppenidentitäten. In sozialen Netzwerken gewinnen visuelle Formate, Multimodalität und kollektive Meme-Kulturen zunehmend an Bedeutung für die Repräsentation und Ausdifferenzierung von Fanpraktiken. Hierdurch entstehen verschärfte Abgrenzungen, aber auch innovative Formen digitaler Solidarität, Mobilisierung und gegenseitiger Unterstützung innerhalb und zwischen Fangruppen (Krell et al. 2026: 292-293; 294-295). Multioptionale Interaktionen wie Live-Kommentare, *Second-Screen*-Nutzung und personalisierte Fanservices bereichern das Fanerlebnis und transformieren es in einen vernetzten, multimodalen Prozess. Die soziale Funktion von Fanschaft erfährt damit eine nachhaltige Verschiebung, indem digitale Räume klassische Treffpunkte ergänzen oder substituieren und neue Ausdrucksformen von Zugehörigkeit, Protest und Distinktion begünstigen (vgl. Billings/Brown 2017).[2]

3. Konflikte und Konsens in digitalen Öffentlichkeiten

Digitale Medien transformieren die Struktur und Funktionsweise von Öffentlichkeit grundlegend (vgl. Imhof 2026). Sie erweitern klassische, auf Konsens und rationalen Dialog ausgerichtete Öffentlichkeitsmodelle um konfliktorientierte, expressive Kommunikationsarenen und fördern damit eine Pluralisierung kommunikativer Praktiken.

Deliberative Kommunikation zielt auf argumentative Verständigung, Gleichberechtigung und Gemeinwohlorientierung. Ihre normativen Prin-

2 Abeza (2023: 258) hat angesichts der engen Wechselbeziehung zwischen Social Media und Sport(kommunikation) eine kritische Perspektive auf diese Entwicklung gefordert.

zipien – rationale Rechtfertigung, wechselseitige Bezugnahme und die Orientierung an journalistischen Standards wie Relevanz, Faktizität und Neutralität – sollen Konsensbildung oder zumindest die Legitimation kollektiv getragener Entscheidungen ermöglichen (vgl. Habermas 1981). Dieses Modell bildet das ethische Fundament liberaler Öffentlichkeit, indem es Diskursteilnehmende zu verantwortlichem, argumentativem und respektvollem Handeln verpflichtet.

Das agonistische Paradigma (vgl. Mouffe 1999, 2015; Friß/Gilleßen 2023: 90–91) hingegen begreift unversöhnliche Differenzen und emotionale Gegnerschaft als konstitutive Elemente demokratischen Diskurses. Ziel ist weniger Konsens als vielmehr die Sichtbarkeit und Austragbarkeit konkurrierender hegemonialer Projekte. Frick (2023) betont dabei die Notwendigkeit einer agonistischen Konfliktkultur, die zwischen Gegnern und Feinden unterscheidet und politische Gegnerschaft zivilisiert aushandelt, ohne sie zu delegitimieren. Demokratische Streitkultur verlangt somit die Anerkennung bleibender Differenzen bei fortbestehendem Respekt vor den verfassungsmäßigen Rahmenbedingungen.

Digitale Plattformen verschärfen diese Spannung zwischen deliberativer und agonistischer Öffentlichkeit. Plattform- und Semi-Öffentlichkeiten zeichnen sich durch eine geringere Konsensorientierung, eine stärkere Emotionalisierung und erhöhte Polarisierung aus (vgl. Klinger 2018; Pfetsch et al. 2018). Zugleich unterminieren sie klassische g*atekeeping*-Mechanismen, was zwar die Vielfalt der Stimmen erhöht, zugleich aber das Risiko normativer Desintegration verstärkt. Medienethisch relevant ist dabei die Frage, wie sich Kommunikationsfreiheit, Verantwortung und Integrität in digital fragmentierten Öffentlichkeiten bewahren lassen.

Die Digitalisierung der Öffentlichkeit führt somit zu einer strukturellen Pluralisierung, in der deliberative und agonistische Kommunikationsformen koexistieren und sich wechselseitig herausfordern. Beide Modelle entfalten Chancen und Risiken für Inklusion, Konfliktbearbeitung und gesellschaftliche Kohäsion. Aus medienethischer Perspektive fordert diese digitale Mehrfach-Öffentlichkeit ein erweitertes Verständnis von Verantwortung, Dissens und institutionell „gezähmtem" Konflikt im Sinne einer reflektierten, pluralen Demokratie.

4. Operationalisierung der agonistischen Fankommunikation und Untersuchungsanlage

4.1 Operationalisierung der agonistischen Fankommunikation

Der Operationalisierungsvorschlag von Frieß und Gilleßen (2022) kann zu einer differenzierten kommunikationswissenschaftlichen Erforschung von Konfliktkulturen, kollektiver Identitätsarbeit und affektiven Dynamiken in digitalen Fanöffentlichkeiten beitragen. Betrachtet werden fünf miteinander verschränkte Dimensionen: Konflikt, Gegnerschaft, Hegemonie, kollektive Identitäten und Leidenschaften (vgl. Frieß/Gilleßen 2022: 91).

Im Zentrum steht die Dimension Konflikt (vgl. ebd.: 91–92, 95–96). Sie umfasst die Identifikation von Dissens – also Uneinigkeit oder expliziter Meinungsverschiedenheit –, die Ausprägung von Konfrontation, etwa durch das explizite Attackieren gegnerischer Positionen, und die Prüfung, ob trotz aller Widersprüche die grundlegende demokratische Anerkennung des Anderen als legitimer Diskussionspartner erhalten bleibt. Die Differenz zwischen bloßer Meinungsverschiedenheit und offener Konfrontation wird dabei analytisch deutlich herausgearbeitet. Pluralismus-Anerkennung fungiert als kategoriale Grenzlinie zum antagonistischen Diskurs, wenn dem Opponenten das Recht zur Partizipation an der Debatte abgesprochen wird.

Die zweite analytische Dimension, Gegnerschaft (vgl. ebd.: 92, 96), bezieht sich auf Prozesse der Herstellung von Dichotomie und Konkurrenz. Hier gilt es, die explizite Identifikation von Gegnern zu erfassen und zu beurteilen, ob diese Gegnerschaft in legitimer Weise erfolgt oder eine Dehumanisierung stattfindet. Während Frieß und Gilleßen zeigen, dass eine legitime Gegnern-Bestimmung im demokratischen Sinn immer die Daseinsberechtigung des Anderen einschließt, markiert jede Form der Dehumanisierung theoretisch den Übergang zum Antagonismus – etwa dort, wo Gegner entmenschlicht, als „Feind" sprachlich diffamiert oder grundsätzlich delegitimiert werden.

Die Dimension Hegemonie (vgl. ebd.: 92, 96–97) erfasst Auseinandersetzungen um bestehende Institutionen, Werteordnungen, Machtverhältnisse und diskursive Restriktionen. Hier interessiert, ob beispielsweise Vereine, Verbände oder normativ aufgeladene Narrative und Werte offensiv in Frage gestellt werden und ob Machtbeziehungen oder Tabus in der Öffentlichkeitskommunikation kritisch adressiert werden. Die empirische Operationalisierung in der Fallstudie von Frieß und Gilleßen belegt, dass explizite

Hegemoniekritik zwar seltener, aber in hochstrittigen Arenen dennoch ein relevantes Agens agonistischer Auseinandersetzung ist.

Der Aspekt kollektive Identitäten (vgl. ebd.: 92–93, 97) umfasst jene kommunikativen Muster, in denen die Zugehörigkeit zu einem bestimmten Kollektiv hergestellt oder betont wird. Neben der Sichtbarmachung kollektiver Identität lassen sich auch Zugehörigkeitsmerkmale (Symbole, Narrative), Mobilisierungsappelle und Differenzierungen gegenüber „dem Anderen" analytisch fassen. Kollektive Identitätsbildung und Gruppenappell sind dabei zentrale Mechanismen agonistischer Öffentlichkeiten, mit denen Fankulturen symbolisch verdichtet werden.

Schließlich richtet sich die fünfte Dimension auf Leidenschaften (vgl. ebd.: 93, 97), also auf das emotionale Kraftfeld, das eine agonistische Arena kennzeichnet. Hier analysiert das Modell sowohl positive (Freude, Hoffnung) als auch negative Emotionen (Wut, Frustration) sowie den affektiven Wertebezug. Die empirischen Befunde von Frieß und Gilleßen zeigen, dass insbesondere negative Leidenschaften einen hohen Stellenwert in online geführten Konflikten besitzen, während positive Bezugnahmen deutlich seltener auftreten.

4.2 Untersuchungsgegenstand

Untersuchungsgegenstand der vorliegenden Pilotstudie ist die Kommunikation der Fans von zwei österreichischen Eishockeyvereinen im Umfeld der Derbys in der Saison 2024/2025 insbesondere in digitalen Öffentlichkeiten wie Foren, Social Media und Live-Tickern.

Die Rivalität zwischen dem EC KAC aus Klagenfurt und dem EC VSV aus Villach, bekannt als Kärntner Eishockey-Derby, gilt als eines der traditionsreichsten und emotional aufgeladensten Duelle im österreichischen Mannschaftssport. Als ältestes Derby im österreichischen Eishockey, dessen Wurzeln bis in das Jahr 1929 zurückreichen, steht es zugleich für Aushandlungsprozesse regionaler Identität, in denen sich Klagenfurt als Landeshauptstadt und Villach als zweitgrößte Stadt Kärntens gegenüberstehen. Bis zum Beginn der Saison 2024/2025 wurden 355 Kärntner Derbys in offiziellen Bewerbsspielen ausgetragen, in denen der KAC mit 189 Siegen, 19 Unentschieden und einem Torverhältnis von 1343:1160 dominiert. Mit den Begegnungen im Herbst 2024 und im Kalenderjahr 2025 wurde diese Bilanz auf 192 Siege des KAC, 148 des VSV und 19 Unentschieden aus insgesamt 360 Derbys ausgebaut, womit sich die langfristige Dominanz

der „Rotjacken" bei gleichzeitig deutlicher sportlicher Konkurrenz der „Adler" bestätigt. Markante Ereignisse wie der legendäre 16:0-Sieg des KAC im Jahr 1960, eine 17 Spiele umfassende Siegesserie des VSV Mitte der 2000er-Jahre oder spektakuläre Comebacks und Playoff-Serien fungieren als zentrale narrative Fixpunkte, an die sich Fanerinnerungen und Medienberichterstattung knüpfen. Diese „Derby-Mythen" werden in digitalen Räumen immer wieder aktualisiert, etwa wenn Fans historische Spiele, ikonische Tore oder prominente Akteure (von Edi Lebler über Greg Holst bis zu aktuellen Derbyhelden) heranziehen, um kollektive Identität zu stabilisieren und Zugehörigkeit zu markieren (vgl. Clar 2021). Fans beider Lager erleben das Derby als kulturelles Ereignis, in dem Stadionbesuche, mediale Live-Rezeption und digitale Anschlusskommunikation eng verwoben sind. Die Villacher „Adler" und die Klagenfurter „Rotjacken" inszenieren Zugehörigkeit durch Slogans, Gesänge und humorvolle wie aggressive Bezeichnungen des Gegners („Schlümpfe", „Neblinger"), wobei insbesondere Social-Media-Plattformen und Foren zu zentralen Schauplätzen dieser symbolischen Auseinandersetzungen werden.

4.3 Untersuchungsanlage

Eine Vorstudie zur *Instagram*-Kommunikation der beiden Kärntner Eishockeyvereine KAC und VSV während der Derbys in der Saison 2023/24 (vgl. Melcher 2024) bildet den Ausgangspunkt. Mittels einer qualitativen Inhaltsanalyse wurden alle öffentlichen und offiziellen Instagram-Beiträge beider Vereine im Kontext der vier Derby-Spiele der Saison 2023/24 untersucht (insgesamt 44 Posts: KAC 15, VSV 29). Das Kategoriensystem unterscheidet inhaltliche, visuelle und kommunikative Aspekte und ermöglicht eine systematische Auswertung der Beiträge hinsichtlich bespielter Inhalte, Darstellungsformen und Strategien.

Aufbauend auf dieser Vorarbeit wurden für die Derbys der Saison 2024/25 die Kommunikationspraktiken je eines offiziellen Fanclubs der beiden Eishockeyvereine analysiert. Im Fokus standen die Profile von „Blau Weiß" (Villach) und „Stiege 19" (Klagenfurt) auf *Instagram* und *Facebook*, um die klubnahe, aber nicht vereinsoffizielle Perspektive zu erfassen und mit der institutionellen Kommunikation der Vereine zu kontrastieren. Das in der Studie von Melcher (2024) verwendete Kategorienschema wurde um die von Frieß und Gilleßen (2022) vorgeschlagenen Dimensionen agonistischer Kommunikation erweitert, sodass nun 14 formale und in-

haltliche Kategorien (unter anderem Bezugnahme auf den Gegner, Freund-Feind-Muster, Emotionsinszenierung, In-/Exklusion, Umgang mit Dissens) berücksichtigt wurden. Analysiert wurden lediglich 19 Beiträge der beiden Fanclubs mit insgesamt 38 Kommentaren. Aufgrund der geringen Stichprobengröße handelt es sich um eine Pilotstudie, in der der Schwerpunkt auf einer dichten, qualitativen Rekonstruktion der Fankommunikation, ihrer diskursiven Muster und agonistischen Dynamiken liegt, nicht auf Generalisierbarkeit.

4.4 Limitationen

Damit ist die zentrale Limitation der Studie benannt. Die Anzahl der ausgewerteten Posts ist gering, sodass nur ein eingeschränktes Bild der Kommunikationskultur der Eishockeyfans gezeichnet werden kann. Besonders ins Gewicht fällt, dass in dem zugänglichen Material keine offenen Anfeindungen oder klar identifizierbaren Hassbotschaften auftreten. Solche Äußerungen scheinen – soweit aus dem vorliegenden Korpus ersichtlich – nicht in den öffentlich zugänglichen Kanälen stattzufinden oder sie sind dort nicht (mehr) sichtbar. Flankierend zu der Analyse der Social-Media-Kommunikation wurden bzw. werden (Experten:innen-)Gespräche mit Verantwortlichen der beiden Vereine sowie Fangruppen geführt. Da diese Gespräche bislang noch nicht ausgewertet wurden, kann nicht bestimmt werden, welche Online-Kommentare moderiert beziehungsweise gelöscht wurden.

Auffällig ist nicht nur die geringe Anzahl an öffentlich einsehbaren Posts und Kommentaren, sondern auch das Fehlen von Hasskommunikation. Ein möglicher Kontextfaktor, der auf eine eher streng regulierte Kommunikationskultur hindeutet, ist der im Sommer 2024 öffentlich ausgetragene Konflikt mit dem Fanclub „Absolut Villach", der schließlich zu dessen Stadionausschluss führte – einer Art „Höchststrafe" innerhalb der organisierten Fankultur. Dieser Vorgang kann als Indikator dafür gelesen werden, dass grenzüberschreitende oder konfliktverschärfende Kommunikationsformen nicht nur sanktioniert, sondern im Extremfall mit dem Ausschluss ganzer Gruppen beantwortet werden. Zugleich bleibt offen, inwieweit solche Sanktionen problematische Kommunikationsformen tatsächlich verhindern oder lediglich in weniger sichtbare, nicht-öffentliche Kanäle verlagern.

5. Befunde: Agonistische Fankommunikation zwischen KAC und VSV

5.1 Vorstudie zur Instagram-Kommunikation des KAC und des VSV während der Derbys 2023/2024

Mit ihrer Studie konnte Melcher zeigen, dass der KAC im Vergleich zum VSV weniger häufig, dabei auf Ankündigungen (Ticketverkauf, *gameday updates*, *Lineups*) und Rückblicke auf gewonnene Spiele fokussierte. Auffällig war, dass die Mehrzahl der Kommentare auf Englisch war (vgl. Melcher 2024: 37) Der VSV zeigte eine höhere Postingfrequenz, nutzte eine größere inhaltliche Vielfalt und postete auch Zwischenstände und aktuelle Spielereignisse während der Derbys. Hierbei erfolgte die Kommunikation überwiegend auf Deutsch (vgl. ebd.: 38–39). In beiden Fällen stehen die Sportler und deren Leistungen sowie emotionale Aspekte (vor allem Rivalität und Ehrgeiz) im Vordergrund. Es dominieren in den Fangruppen positive Emotionen wie Freude nach Siegen, jedoch zeigen sich Unterschiede in der Darstellung von Enttäuschung nach Niederlagen. Die VSV-Anhänger thematisierten ihre Enttäuschung offener, sie hatten angesichts der vielen Niederlagen auch häufiger Anlass dazu. Die visuelle Inszenierung folgt klaren Mustern: Häufig sind 2–4 Spieler pro Bild zu sehen. Trainer und Funktionäre werden selten gezeigt. Die ästhetische Bildgestaltung und die Nutzung von Hashtags (KAC: #KAC, VSV: #*RiseWithU*) dienen der digitalen Markenbildung und Reichweitengenerierung (vgl. ebd.: 41–43). Während der KAC tendenziell zurückhaltender und internationaler in seiner Ansprache agiert, setzt der VSV auf einen konstanten *posting*-Rhythmus, emotionale *visuals* und s*torytelling* zu Zwischenergebnissen, was eine kontinuierliche Fanbindung fördert bzw. fördern soll (vgl. ebd.: 33–34, 45–47).

In den untersuchten Beiträgen ist der Konflikt präsent, zeigt sich aber vor allem in Form sportlicher Rivalität und der Inszenierung von sportlichem Wettstreit. Explizite Konfrontationen werden selten publik gemacht; Dissens wird primär zwischen den Teams dargestellt, nicht zwischen Fans oder mit externen Gegnern. Demokratische Anerkennung des Anderen als legitimen Teil des Wettbewerbs bleibt im Gesamtkontext der Kommunikation meist erhalten, somit wird der antagonistische Diskurs vermieden.

Die Studie zeigt, dass die Vereinspräsenz beziehungsweise -kommunikation und die Fan-Kommunikation sowohl beim KAC als auch beim VS stark professionalisiert und dynamisiert ist. Beide Vereine etablieren durch digitale Kanäle eine direkte, schnelle Interaktion mit den Fans.

Die unterschiedlichen Strategien reflektieren dabei sowohl die Kommunikationsziele als auch die spezifische Vereinsidentität (vgl. ebd.: 46–47). *Instagram* dient den Vereinen nicht nur als Distributionskanal für Information, sondern auch als Raum für vielseitige Inszenierung und diskursive Pluralisierung ist. Kommentare, Likes und Hashtags fördern nicht nur Interaktion, sondern stiften Community-Gefühl und ermöglichen Fans, sich aktiv mit Deutungsangeboten auseinanderzusetzen.[3]

5.2 Kommunikation der Fanclubs des KAC und VSV im Umfeld der Derbys 2024/2025

In der Pilotstudie zur Fankommunikation im Umfeld der Derbys 2024/2025 zeigt sich, dass Beiträge vor allem Rivalität und Ehrgeiz thematisieren, während Enttäuschung oder reflexive Einordnungen deutlich seltener auftreten. Meinungsverschiedenheiten über Spielverläufe, Vereinsführung, Transferpolitik und Fanverhalten werden offen und häufig pointiert ausgetragen. Konfrontationen erfolgen vorwiegend über kritische Beiträge oder scharf formulierte Kommentare gegenüber rivalisierenden Vereinen und deren Anhängerschaft, bewegen sich jedoch zumeist innerhalb der Grenzen einer pluralistischen Debattenkultur. Der gegnerische Verein erscheint dabei überwiegend als legitimer diskursiver Gegenpart, nicht als Feindfigur, wodurch pluralen Positionen Raum gegeben und destruktive Feindmarkierungen begrenzt werden.

Die sprachliche Konstruktion von Gegnerschaft ist klar erkennbar: Die eigene Fangruppe wird als loyal, leidenschaftlich und einsatzbereit gerahmt, während das gegnerische Lager als sportlich herausfordernd oder problematisch beschrieben wird, ohne dass entmenschlichende Zuschreibungen erfolgen. Delegitimierende oder entwürdigende Fremdbilder bleiben aus. Hegemoniekritische Elemente – etwa Kritik an Kommerzialisierung, Vereinsstrukturen, Sponsoreneinflüssen oder Einschränkungen von Fanfreiheiten – treten punktuell auf und lassen sich als Teil einer agonistischen Kommunikationskultur verstehen. Solche Beiträge verweisen

3 Im Rahmen der Pilot-Studie wurden keine sozio-demographischen Daten (Geschlecht, Alter, Bildung) der Fans erhoben. An dieser Stelle ist daran zu erinnern, dass nur eine Befragung der Verfasser:innen der (wenigen) Posts und Kommentare zu klären vermag, wer sich hinter den Profilnamen verbirgt. Für die Saison 2026/2027 ist eine größere Studie zu den Fans des KAC geplant, die dann auch die sozio-demographischen Daten erfasst.

auf Meinungsvielfalt und auf eine ausgeprägte Fähigkeit der Fanbasis zum Protest und auch einer kritischen Selbstbeobachtung.

Gleichzeitig bleiben dichotome Strukturierungen („Wir gegen die Anderen") zentral für die Derby-Kommunikation: Die Markierung des sportlichen Gegners bildet einen Kern der Inszenierung, ohne dass in der offiziellen Vereinskommunikation systematisch Dehumanisierung oder Diffamierung erkennbar wäre. Identitätsstiftende Symbole, Narrative und Markenelemente (etwa Vereinsfarben, Logos oder *hashtags* wie #rotjacken und #*RiseWithUs*) stehen im Vordergrund und erzeugen Gemeinschaftsgefühl durch adressierte Fangemeinden, Mobilisierungsappelle und wiederkehrende Codes; Differenz wird damit betont, aber im Rahmen sportlicher Konkurrenz gehalten.

Die Betonung der Community-Zugehörigkeit ist in der Fankommunikation prägend: Symbole wie Vereinsfarben, Fan-Choreografien, Hymnen und narratives Erzählen zu Historie und Rivalitäten stiften kollektive Identität und intensivieren das Gemeinschaftsgefühl. Mobilisierungsappelle, Unterstützung und klare Grenzziehungen gegenüber anderen Fanlagern sind wesentliche Integrationsmechanismen dieser Kommunikationspraxis. Dabei sind Emotionen – von Euphorie und Stolz bis zu Frust und Kritik – von zentraler Bedeutung; negative Affekte werden online häufiger artikuliert, durch positive Bezugnahmen allerdings meist moderiert. Dies erklärt auch, dass offen rassistische oder sexistische Äußerungen im Untersuchungsmaterial nicht auftauchen.[4]

Die Untersuchung der Fankommunikation von KAC und VSV – z. B. auf Instagram und Facebook – verdeutlicht: Kontroverse, teils emotional geführte Diskussionen verlaufen überwiegend respektvoll und grenzen sich von destruktiven, entmenschlichenden Auseinandersetzungen ab. Die Integration kollektiver Identitätsstiftung mit kritischer Vereinskritik zeigt, dass Social-Media-Kommunikation sowohl Inklusions- als auch Exklusionspotenziale birgt, die sich ethisch reflektieren lassen.

Ein anschauliches Beispiel für moderierte Konfliktaustragung bietet das Transparent „zu Gast bei Schönwetterfans" der KAC, das ironisch-abgrenzend gegenüber unloyalen Zuschauern agiert. Die nachfolgende Intervention der Veranstalter und kritische Kommentierung dieses Vorgangs illus-

4 Wie oben erwähnt, können keine genauen Aussagen darüber getroffen werden, wie viele (und welche) weiblichen Fans sich in auf den offiziellen sozialen Plattformen geäußert haben. Es ist davon auszugehen, dass die Fans des KAC und des VSV überwiegend männlich und älter sind, die Angehörigen der Fanclubs ebenfalls überwiegend männlich, aber jünger als die „Durchschnittsfans" sind.

trieren empirisch die feinen Aushandlungsprozesse zwischen individueller Fan-Meinung, den Normen des Vereins und den Erfordernissen inklusiver und respektvoller Kommunikationspraxis. Dies verdeutlicht die medienethische Herausforderung von Vereinsakteuren, die auch online für einen konstruktiven Diskurs und gegen emotionale Eskalation oder unfaires Verhalten steuern müssen.

Abbildung 1: Kommentare zum Fanbanner und zum Fanverweis

Zudem unterstreicht die sozial-karitative Initiative des EC KAC – dargestellt in einem Instagram-Post – die Potenziale medial vermittelter Sportkommunikation für inklusives, sozial verantwortliches Handeln. Die Organisation eines Heimspielbesuchs für benachteiligte Kinder zeigt, wie über digitale Kanäle nicht nur Gemeinschaftsgefühl, sondern auch Werte sozialer Inklusion und gesellschaftlicher Verantwortung transportiert und visibilisiert werden können. Die Verknüpfung von karitativem Engagement, symbolischer Kommunikation und emotionaler Inszenierung verdeutlicht die ethische Relevanz, die Vereine und Fanorganisationen bei der Gestaltung öffentlicher Diskurse innehaben.

Abbildung 2: Charity-Aktion der Organisation „Soldaten mit Herz" in Zusammenarbeit mit dem KAC

Abschließend wird sichtbar: Die Vereins- und Fankommunikation im Social-Media-Kontext realisiert eine Balance zwischen agonistischem Streit, zivilisierenden Kommunikationsstandards und sozial verantwortlichem Handeln. Ob diese Balance auch in den nicht-offiziellen beziehungsweise nicht-öffentlichen Foren und Gruppen anzutreffen ist müsste in einem nächsten Schritt untersucht werden.[5] Für die zukünftige Analyse sind eine erhöhte Sensibilität hinsichtlich Inklusion, Menschenwürde und der kritischen Reflexion von Machtverhältnissen und Exklusionsmechanismen in der Fankommunikation essenziell – zentrale Aufgaben der angewandten Medienethik im digitalisierten Sport.

6. Diskussion

Fankommunikation in digitalen Öffentlichkeiten erweist sich als ambivalenter Kommunikationsraum, in dem sich gemeinschaftsstiftende Zugehörigkeit, agonistische Streitkulturen und potenziell exkludierende Praktiken

5 Zugriff und Analyse der (Fan-)Kommunikation in den nicht-öffentlichen Foren/Gruppen erfordert eine forschungsethische Reflektion. Vgl. Krell et al. 2026: 292.

überlagern. Die Analyse der Vereins- und Fanaccounts rund um die Kärntner Derbys zeigt, dass digitale Plattformen nicht nur emotionale Bindung, Identitätsarbeit und mobilisierende Symbolik intensivieren, sondern zugleich Konflikte verdichten und normativ aufladen. In der beobachteten Kommunikationspraxis dominiert ein agonistischer Modus, in dem Gegnerschaft klar markiert, aber überwiegend innerhalb anerkannter Spiel- und Diskursregeln ausgetragen wird. Damit werden Formen „zivilisierten Streitens" (Frick 2023) realisiert, die Differenz sichtbar halten (vgl. Clar 2021), ohne systematisch in entmenschlichende Feindkonstruktionen umzuschlagen.

Gleichzeitig verweisen die Ergebnisse auf die Fragilität dieser Balance. Sanktionen gegenüber Gruppen, dokumentieren einerseits die Bereitschaft, exzessive und eskalative Kommunikationsformen zu begrenzen; andererseits besteht die Gefahr, problematische Tendenzen lediglich in weniger sichtbare, informelle oder private Kanäle zu verlagern. Damit verschiebt sich die medienethische Problemwahrnehmung: Nicht nur offen zugängliche Arenen, sondern auch teilöffentliche und geschlossene Kommunikationsräume werden für die Analyse von Inklusion, Ausgrenzung und Radikalisierung relevant. Fankommunikation bleibt somit auch dort ein Schlüsselphänomen, wo sie sich der regulierten Sichtbarkeit offizieller Kanäle entzieht.

In der untersuchten Fankommunikation verdichtet sich ihre Deutungsmacht im Spannungsfeld zwischen Solidaritätsadressen für das eigene Lager, ironisch-polemischen Praktiken der Abgrenzung und punktuellen Überschreitungen, die institutionelle Interventionen nach sich ziehen können. Die empirischen Befunde legen nahe, dass offizielle Vereins- und Fanaccounts tendenziell auf kuratierte Emotionalität, positive Selbstbeschreibung und begrenzte Konfliktintensität setzen, während ungefilterte Affekte, Gerüchte und Polarisierungen eher in randständige oder informelle Kontexte ausweichen. Damit verschieben sich die Grenzen zwischen deliberativen und agonistischen Kommunikationsformen: Während deliberative Elemente – etwa reflexive Einordnungen, Hinweise auf Fair Play oder sozial-karitative Aktionen – diskursiv präsent sind, bleibt der zentrale Attraktor eine emotional aufgeladene, agonistische Inszenierung der Rivalität.

Vor diesem Hintergrund lässt sich Fankommunikation als Labor einer medienethisch relevanten Konfliktkultur verstehen. Sie illustriert, wie kollektive Identitäten in einem von Kommerzialisierung, Plattformlogiken und Markenstrategien geprägten Umfeld ausgehandelt werden und wie Vorstellungen von Zugehörigkeit, Ausschluss und legitimer Gegnerschaft kommu-

nikativ stabilisiert oder herausgefordert werden. Die untersuchte Balance zwischen normativ gezähmtem Agonismus und punktuell sichtbarer Eskalation zeigt, dass Vereine und Fanorganisationen nicht nur als Markenakteure, sondern zugleich als diskurspolitische Instanzen zu begreifen sind. Ihnen fällt die Aufgabe zu, emotionale Intensität zu ermöglichen, ohne destruktive Dynamiken zu normalisieren, und zugleich Räume für Kritik, Widerspruch und hegemone Reflexion offenzuhalten.

Daraus ergeben sich mehrere Perspektiven für Forschung und Praxis. Erstens bedarf es weiterer empirischer Studien, die informelle und schwer zugängliche Kommunikationsräume (zum Beispiel Messenger-Gruppen, geschlossene Foren) systematisch in den Blick nehmen, um Verschiebungen zwischen sichtbaren und verdeckten Konfliktarenen besser zu verstehen. Zweitens sollten interdisziplinäre Ansätze – etwa an der Schnittstelle von Medienethik, Diskurslinguistik, Sportsoziologie und Plattformforschung – ausgebaut werden, um die Verbindung von Sprachgebrauch, Affektökonomien und Machtverhältnissen in Fankulturen angemessen zu erfassen. Drittens sind praxisorientierte Leitlinien zu entwickeln, die Vereinen, Fanvertretungen und Verbänden Orientierung im Umgang mit Emotionalisierung, Moderation, Sanktionierung und Partizipation geben, ohne agonistische Differenz vorschnell zu domestizieren.

Insgesamt macht die Pilotstudie deutlich, dass Fankommunikation eigenständige Öffentlichkeitsräume, in denen sich gesellschaftliche Aushandlungsprozesse über Zugehörigkeit, Anerkennung, Grenzen legitimer Gegnerschaft und den Schutz vor Diskriminierung exemplarisch verdichten. Eine medienethisch sensibilisierte Sportkommunikation ist daher gefordert, diese Räume kritisch zu begleiten.

Literatur

Abeza, Gashaw (2023): Social Media and Sport Studies (2014–2023). A Critical Review, in: International Journal of Sport Communication 16 (3/2023), S. 251–261.

Biel, Jonas et al. (Hg.) (2025): Uniting Europe through Football. The interplay between Fandom, Identity and social cohesion, Cham.

Billings, Andrew. C. / Brown, Kenon A. (Hg.) (2017): Evolution of the modern sports fan: Communicative approaches, Laham.

Brandt, Christian / Hertel, Fabian (2017): Introduction. Rivalry and cooperation in football, in: Christian Brandt / Fabian Hertel / Sean Huddleston (Hg.), Football Fans, Rivalry and Cooperation, New York, S. 1–15.

Clar, Peter (2021): Es darf von allem ein bisschen mehr sein. Essay zu VSV gegen KAC. Kleine Zeitung, 14. März 2021 (online unter: https://www.kleinezeitung.at/sport/eishockey/ersteliga/kac/5951024/Essay-zu-VSV-gegen-KAC_Es-darf-von-allem-ein-bisschen-mehr-sein – letzter Zugriff: 8.12.2024).

DEG (Düsseldorfer Eishockey Gemeinschaft) (2025): DEG und KEC verurteilen Auseinandersetzungen vor dem Derby, 25. November 2025 (online unter: https://www.deg-eishockey.de/kec-und-deg-verurteilen-auseinandersetzungen-vor-dem-derby/ – letzter Zugriff: 18.8.2025).

Degele, Nina (2013): Fußball verbindet – durch Ausgrenzung, Wiesbaden.

Duttler, Gabriel / Haigis, Boris (Hg.) (2016): Ultras: eine Fankultur im Spannungsfeld unterschiedlicher Subkulturen, Bielefeld.

Eisenegger, Mark et al. (Hg.) (2021): Digitaler Strukturwandel der Öffentlichkeit: Historische Verortung, Modelle und Konsequenzen, Wiesbaden.

Feltes, Thomas (2010): Fußballgewalt als misslungene Kommunikation, in: Neue Praxis 4, S. 405–421.

Frandsen, Kirstin (2020): Sport and Mediatization, New York.

Frick, Marie-Luisa (2023): Zivilisiert streiten. Zur Ethik der politischen Gegnerschaft, Stuttgart.

Frieß, Dennis / Gilleßen, Rabea (2022): Agonistische Online-Öffentlichkeiten. Vorschlag einer inhaltsanalytischen Operationalisierung von Agonismus, in: Publizistik 67, S. 85–108.

Gantz, Walter (2013): Reflections on communication and sport: On fanship and social relationships, in: Communication & Sport 1 (1–2/2013), S. 176–187.

Gantz, Walter / Lewis, Nicky (2023): Sports Fanship Changes acros the Lifespann, in: Communication & Sport 11 (1/2023), S. 8–27.

Gummert, Henrik et al. (Hg.) (2017): Medien und Kulturen des Konflikts, Kulturelle Figurationen: Artefakte, Praktiken, Fiktionen, Wiesbaden.

Habermas, Jürgen (1981): Theorie des kommunikativen Handelns, Frankfurt am Main.

Habermas, Jürgen (2022): Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Berlin

Havard, Cody T. (2020): Rivalry in Sport. Understanding Fan Behavior and Organizations, Cham.

Haupt, Tobias / Herberth, Christoph (2017). Fan-Kommunikation 3.0: Neue und innovative Möglichkeiten der Fan-Kommunikation im Zeitalter der digitalen Medien, in: André Schneider / Julia Köhler / Schumann, Frank (Hg.), Fanverhalten im Sport. Phänomene, Herausforderungen und Perspektiven, Wiesbaden, S. 159–173.

He, Mu (2024): Sports Identification in Situational Problem Solving: A Framework for Understanding Fans Communicative Behaviors in Sports Crises, in: Communication & Sport 13 (6/2024), S. 1276–1303.

Heltzel, Andreas (2016): Die Medialität des Agon. Sport und Spiel in der klassischen Antike, in: Volker Schürmann et al. (Hg.), Bewegungskulturen im Wandel. Der Sport der Medialen Moderne – Gesellschaftstheoretische Verortungen, Bielefeld, S. 89–105.

Hepp, Andreas / Berg, Matthias / Roitsch, Cindy (2012): Die Mediatisierung subjektiver Vergemeinschaftungshorizonte: Zur kommunikativen Vernetzung und medienvermittelten Gemeinschaftsbildung junger Menschen, in: Friedrich Krotz / Andreas Hepp (Hg.), Mediatisierte Welten: Forschungsfelder und Beschreibungsansätze, Wiesbaden, S. 227–256.

Hill, Tim / Canniford, Robin / Millwand, Peter (2016): Against modern football: Mobilising protest movements in social media, in: Sociology 52 (4/2016), S. 688–708.

Ihle, H. (2016). Sport und Medien. Bestandsaufnahme des Forschungsfeldes, in: Communicatio Socialis 49 (2/2016), S. 134–152.

Imhof, Kurt (2006). Mediengesellschaft und Medialisierung, in: Medien und Kommunikationswissenschaft 54 (2/2006), S. 191–215.

Katzenbach, Christian (2017): Von kleinen Gesprächen zu großen Öffentlichkeiten? Zur Dynamik und Theorie von Öffentlichkeiten in sozialen Medien, in: Elisabeth Klaus / Rickarda Drüeke (Hg.), Öffentlichkeit und gesellschaftliche Aushandlungsprozesse. Theoretische Perspektiven und empirische Befunde, Bielefeld, S. 151–174.

Klinger, Ulrike (2018): Aufstieg der Semiöffentlichkeit: Eine relationale Perspektive, in: Publizistik 63 (1–2/2018), S. 245–267.

Krell, Felix et al. (2026): „Ich hab mir übrigens gerade ein Bier aufgemacht! Nehmt das, ihr Kataris" – Ko-Orientierung während der Fußballweltmeisterschaft in Katar 2022, in: Lina. Sūna / Wolfgang Reißmann (Hg.), Mediensozialisation in „smarten" Umgebungen. Selbst- und Sozialwerdung im Kontext von Datafizierung und Automatisierung, Wiesbaden, S. 285-300..

Melcher, Antonia (2024): Vom Stadion in das Netz. Eine Analyse der Instagram-Beiträge zu den ICE-Hockey-League-Derbys zwischen KAC und VSV in der Saison 2023/24, Klagenfurt (unveröffentlichte BA-Thesis Universität Klagenfurt).

Meyen, Michael (2014): Medialisierung des deutschen Spitzenfußballs. Eine Fallstudie zur Anpassung von sozialen Funktionssystemen an die Handlungslogik der Massenmedien, in: Medien & Kommunikationswissenschaft 62 (3/2014), S. 377–393.

Mouffe, Chantal (1999): Deliberative democracy or agonistic pluralism?, in: Social Research 66 (3/1999), S. 745–758.

Mouffe, Chantal (2015): Agonistik: Die Welt politisch denken, Bonn.

Müller, Marion (2009): Fußball als Paradoxon der Moderne. Zur Bedeutung ethnischer, nationaler und geschlechtlicher Differenzen im Profifußball, Wiesbaden.

Nieland, Jörg-Uwe (2023) Kritik und die Folgen. Eine kommunikationswissenschaftliche Betrachtung der FIFA-Weltmeisterschaft 2022, in: Medien Journal 47 (1/2023), S. 24–39.

Pfetsch, Barbara / Löblich, Maria / Eilders, Christine (2018): Dissonante Öffentlichkeiten als Perspektive kommunikationswissenschaftlicher Theoriebildung, in: Publizistik 63, S. 477–495.

Schneider, Arne / Köhler, Julia / Schumann, Frank (Hg.) (2017): Fanverhalten im Sport. Phänomene, Herausforderungen und Perspektiven, Wiesbaden.

Thole, Werner / Pfaff, Nicolle / Flickiner, Hans-Georg (Hg.) (2019): Fußball als soziales Feld. Studien zu Sozialen Bewegungen, in: Jugend- und Fankulturen, Wiesbaden.

Trojanow, Ilija / Zeyringer, Klaus (2024): Fans: Von den Höhen und Tiefen sportlicher Leidenschaft, München.

Werron, Tobias (2010): Der Weltsport und sein Publikum. Zur Autonomie und Entstehung des modernen Sports, Weilerswist.

Wahrnehmung, Einfluss und Folgen sexistischer Hate Speech auf Instagram

Miriam Goetz und Vanessa Wimmers

Zusammenfassung

Die Studie untersucht Wahrnehmung, Einfluss und Folgen sexistischer Hate Speech auf Instagram. Ausgangspunkt ist die wachsende Bedeutung sozialer Medien, die nicht nur Kommunikationsräume, sondern auch Plattformen für Hassrede darstellen. Sexistische Hate Speech wird als geschlechtsspezifische Diskriminierung definiert, die sich in verbalen, nonverbalen und bildbasierten Formen manifestiert. Theoretisch wird das Phänomen durch Ansätze wie gruppenbezogene Menschenfeindlichkeit erklärt, die Macht- und Statusdynamiken betont. Empirisch basiert die Untersuchung auf einer Online-Umfrage unter 153 Instagram-Nutzer:innen, von denen 92,2 Prozent Hate Speech wahrnahmen. Die Ergebnisse zeigen, dass sexistische Hate Speech nicht isoliert bleibt, sondern zur Normalisierung sexistischer Einstellungen beiträgt und die digitale Teilhabe von Frauen einschränken kann. Die Studie verdeutlicht das Spannungsfeld zwischen Meinungsfreiheit und Schutz vor Diskriminierung und unterstreicht die Notwendigkeit rechtlicher und ethischer Regulierung, etwa durch den Digital Services Act.

1. Einleitung

Soziale Medien haben in den letzten Jahren zunehmend an Bedeutung für die globale Kommunikation gewonnen und internationale Plattformen bieten Raum für den Austausch von Meinungen, Ideen und Informationen. Allein in Deutschland nutzen monatlich 171,5 Millionen Nutzer:innen aktiv soziale Plattformen (vgl. Arezo 2025). Digitale Räume haben also enorme Bedeutung für unser tägliches Leben erlangt, sind jedoch auch zu einem Schauplatz für die Verbreitung von Hate Speech geworden, die unter anderem zur Verschärfung geschlechtsspezifischer Vorurteile und Diskriminierungen beitragen. In der JIM-Studie 2025 gaben 64 Prozent der befragten

1200 Jugendlichen (12 bis 19 Jahre) an, mit Hate Speech in den sozialen Medien konfrontiert gewesen zu sein. Ein Anstieg im Vergleich zum Vorjahr um sieben Prozent (vgl. Medienpädagogischer Forschungsverbund Südwest 2025: 55). Diese Zahlen zeigen, dass Hate Speech in den sozialen Medien mittlerweile eine alarmierende Präsenz erreicht hat. Verschiedene Studien zeigen darüber hinaus, dass häufig Frauen Opfer von Hate Speech werden (vgl. Landesanstalt für Medien NRW 2023; Geschke et al. 2019; European Union Agency for Fundamental Rights 2023).

Vor diesem Hintergrund stellen sich die folgenden Forschungsfragen: Wie wird die besondere Form der Hate Speech, die sexistische Hate Speech, auf Instagram praktiziert und wahrgenommen? Wie wirkt sich sexistische Hate Speech im digitalen Raum auf den Alltag von Frauen aus? Und welche Verantwortung haben Plattformbetreiber:innen hinsichtlich Meinungsfreiheit und der Bekämpfung von Hate Speech? Die Plattform Instagram steht im Fokus dieser Untersuchung von Hassrede, da sie eine große Nutzer:innenbasis aufweist und Hate Speech-Phänomene oft durch die visuelle und interaktive Natur der Plattform verstärkt werden. Die Plattform bietet außerdem Tools zur Erkennung und Meldung von Hassreden, wodurch sich das Verhalten von Nutzer:innen, die Verbreitung von Inhalten und die Wirksamkeit von Gegenmaßnahmen abbilden lassen.

Sprache in sozialen Medien prägt gesellschaftliche Strukturen und kann durch Hate Speech diskriminierende Stereotype verstärken, Betroffene psychisch belasten und Hass normalisieren (vgl. Dellagiacoma/Sika Dede Puhlmann 2025: 10). Algorithmische Verstärkung sexistischer Inhalte kann geschlechterspezifische Ungleichheiten fördern (vgl. Römer-Pieretti et al. 2025: 2) Trotz initiierter Plattform-Maßnahmen gegen Hate Speech erweisen sich diese oft praktisch und regulatorisch als unzureichend.

2. Theoretischer Hintergrund der sexistischen Hate Speech

Hate Speech kann allgemein als „Diffamierung und Verunglimpfung" bestimmter Gruppen klassifiziert werden (Sponholz 2021: 17). Jedoch ist nicht jede soziale Gruppe generell oder gleichermaßen von Hate Speech betroffen. Die betroffene Gruppe ist häufig durch eine „[…] benachteiligte Machtposition" charakterisiert (Sponholz 2021: 17). Das Phänomen Hate Speech umfasst dabei nicht nur verbale Formen, sondern kann auch nonverbal oder symbolisch realisiert werden (vgl. Sponholz 2021).

Eine der häufigsten Formen der Hate Speech ist die sexistische Hate Speech, von der vor allem Frauen betroffen sind (vgl. Riemenschneider/Lutz 2021). Sexistische Hate Speech wird im Allgemeinen als „[…] Diskriminierung auf Basis des Geschlechts" verstanden (Mohseni 2021: 40). Hier kann zwischen einer quantitativen und der qualitativen Betrachtung unterschieden werden. In quantitativer Hinsicht lässt sich sagen, dass sich sexistische Hate Speech öfter gegen Frauen als gegen Männer richtet. Qualitativ zeigt sich sexistische Hate Speech auf sozialen Plattformen insbesondere als sexuell-aggressive sprachliche Äußerung (vgl. Mohseni 2021). Frauen werden hier nicht aufgrund individueller oder persönlicher Eigenschaften, Meinungen oder Handlungen angegriffen, sondern primär aufgrund ihrer Zugehörigkeit zum weiblichen Geschlecht (vgl. Riemenschneider/Lutz 2021).

Sexistische Hate Speech lässt sich unter anderem durch die Theorie der gruppenbezogenen Menschenfeindlichkeit erklären (vgl. Zick et al. 2008). Durch das Abwerten anderer sozialer Gruppen wird die eigene aufgewertet. Zu den häufig abgewerteten sozialen Gruppen zählen Ausländer:innen, Geflüchtete, Jüd:innen, Muslim:innen, Homosexuelle, Trans-Personen und Frauen. Eine kommunikationswissenschaftliche Befragung von Bloggerinnen aus den USA verdeutlicht, dass sich sexistische Hate Speech verstärkt vor allem gegen jene Frauen richtet, die sich zu politischen Themen äußern oder in der Öffentlichkeit stehen (vgl. Eckert 2017). Diese Schlussfolgerung bestätigte eine Studie der Interparlamentarischen Union, innerhalb derer 58,2 Prozent der befragten 123 weiblichen Politikerinnen angaben, bereits sexistischen Anfeindungen im Internet ausgesetzt gewesen zu sein (vgl. Inter-Parliamentary Union 2018). Auch andere Frauen, die in der Öffentlichkeit stehen (Journalistinnen, Wissenschaftlerinnen oder Künstlerinnen) sind häufiger von sexistischer Hate Speech betroffen. Zielscheibe der Hate-Speech sind hierbei deren Kompetenzen und deren Körper (vgl. Schieb 2021).

2.1. Formen, Funktionen und Verbreitung von sexistischer Hate Speech

Sprachliche Ausdrucksformen spielen eine zentrale Rolle im Kontext sexistischer Hate Speech, da Sprache nicht nur als Mittel der Kommunikation, sondern auch als Trägerin gesellschaftlicher Machtverhältnisse und Diskriminierungsmechanismen zu werten ist: Ein Sprechakt muss nicht zwangsläufig verletzend sein, entscheidend ist auch der soziale Kontext, in dem

die Äußerung getätigt wird. Soziale Machtverhältnisse spielen hierbei eine tragende Rolle. Personen in einer überlegenen Machtposition gelingt es leichter, ihr Gegenüber herabzusetzen oder dessen gesellschaftlichen Status zu verringern (vgl. Herrmann/Kuch 2007). Wie bei allen Formen sexistischer Hate Speech ist auch hier das Ziel, Macht und Kontrolle über die Betroffenen auszuüben (vgl. Bauer/Hartmann 2021; Barker/Jurasz 2019).

Aktuell häufig verwendete Formen sexistischer Hate Speech in Form sexistischer Beleidigungen rufen vergangene Situationen in Erinnerung, in denen Frauen durch diese Worte erniedrigt und gedemütigt wurden (vgl. Embacher 2021). Dies verstärkt ihre verletzende Wirkung und trägt zur weiteren Verfestigung diskriminierender sprachlicher Strukturen bei.

Sexistische Hate Speech begegnet uns in verschiedenen Ausformungen: Witz, Gerücht, Drohung, Verleumdung oder auch als direkter Aufruf zu Gewalt (vgl. Pandea et al. 2019; Bauer/Hartmann 2021). Zu den Formen sexistischer Hate Speech zählt außerdem die bildbasierte sexualisierte Gewalt (vgl. Bauer/Hartmann 2021). Dazu zählt auch die Veröffentlichung von realen oder Fake-Nacktaufnahmen (vgl. Felling et al. 2019; Powell et al. 2022). Dabei werden oft die Grenzen zum Doxing (Sammeln und Veröffentlichen von persönlichen Informationen der Betroffenen) oder Stalking überschritten und sie sind „[…] Teil der Täterstrategien" (Bauer/Hartmann 2021: 91).

Alle zitierten Studien unterstreichen, dass sich sexistische Hate Speech meist unmittelbar vom virtuellen auf den analogen Raum von Frauen auswirkt (vgl. Bauer/Hartmann 2021). Sexistische Hate Speech kann darauf abzielen, das soziale Umfeld und die Existenzgrundlage der Betroffenen zu beeinträchtigen oder zu zerstören. Zudem steigt durch das Veröffentlichen der privaten Informationen wie Wohnadressen das Risiko, physischen Angriffen ausgesetzt zu sein (vgl. Bauer/Hartmann 2021).

Soziale Medien können als Katalysator von Hate Speech fungieren. Sie ermöglichen es den Nutzer:innen zunächst, im Netz anonym ihre Meinungen zu äußern. Kommunikationsplattformen bieten nicht nur Anonymität, sondern auch eine große Reichweite, durch die sich Gleichdenkende leichter finden können (vgl. Lang 2018). Die Algorithmen sozialer Kommunikationsplattformen verstärken Interaktionen und Bindungen unter den Nutzer:innen, sodass Gruppierungen mit gleichen Einstellungen miteinander kommunizieren und sich in ihren Sichtweisen bestärken können. Problematisch wird es, wenn die von einer Gruppe geteilten Normen im Widerspruch zu sozial akzeptablem Verhalten oder rechtlichen Regelungen stehen und Hate Speech toleriert oder sogar begrüßt wird (vgl. Kaspar

2017). Dazu kommt, dass „Hassgruppen […] mittlerweile erfolgreich signifikante Mengen von Internetnutzern [rekrutieren]" (Robertz et al. 2016: 10). Die Gruppendynamiken gepaart mit den Algorithmen der Plattform wirken dabei als Katalysator von Hate Speech: Je mehr sexistische Hate Speech auf den sozialen Plattformen im Umlauf ist, desto „normaler" wirkt sie. Vor allem die Antworten unter den Hate Speech-Kommentaren erweisen sich als Treiber. Diese enthalten überproportional viel Hate Speech und animieren durch ihre Formulierungen zu weiterführenden Aufrufen, wodurch das Verbreiten erneuter Hassbotschaften angeregt wird (vgl. Schneiders 2021). Mäßigende oder gar kritische Reaktionen auf geäußerte Frauenfeindlichkeit im Digitalen bleiben häufig aus, wodurch die geäußerte sexistische Hate Speech fortlaufend als akzeptabel wahrgenommen wird (vgl. Barker/Jurasz 2019).

Durch die Netz-Anonymität verringert sich insgesamt die Hemmschwelle, Hass zu verbreiten (vgl. Kaspar 2017). Durch die Trennung der Online-Aktivitäten vom realen Leben wird das Online-Selbst zu einem isolierten Selbst, und die Verantwortlichen spüren mitunter kaum oder kein Verantwortungsbewusstsein für die virtuell begangenen Taten (vgl. Suler 2004).

2.2. Reichweitenstarke Plattformen und Folgen

Soziale Medien befinden sich in einem strukturellen Spannungsverhältnis zwischen der Förderung von Meinungsfreiheit und der Maximierung von Reichweite und Nutzerinteraktionen. Während Plattformen wie Instagram, Facebook, X oder TikTok sich öffentlich zur Wahrung der Meinungsfreiheit bekennen, sind ihre algorithmischen Strukturen primär darauf ausgerichtet, Aufmerksamkeit zu generieren und ökonomische Interessen zu bedienen (vgl. Stark et al. 2022). Studien zeigen, dass insbesondere polarisierende und emotional aufgeladene Inhalte bevorzugt verbreitet werden, da sie höhere Interaktionsraten erzeugen und somit den Plattformbetrieb wirtschaftlich rentabler machen (vgl. Stark et al. 2022). Gleichzeitig stehen soziale Netzwerke in der Verantwortung, ihre Nutzer:innen vor schädlichen Inhalten wie Hate Speech, Desinformation oder gezielter Ausgrenzung zu schützen. Die Wahrnehmung von Meinungsfreiheit in Sozialen Medien ist daher ambivalent: Einerseits bieten sie neue Räume für Partizipation und Ausdruck, andererseits führen algorithmische Verzerrungen, Intransparenz bei der Inhaltsmoderation und das Fehlen klarer Standards zu einem Vertrauensverlust in die Plattformen. In der wissenschaftlichen Debatte wird

daher zunehmend gefordert, dass Soziale Medien nicht nur technische Maßnahmen zur Content-Moderation implementieren, sondern auch aktiv zur Förderung eines respektvollen und inklusiven Kommunikationsklimas beitragen (vgl. Zeller 2017; Huber 2023). Diese Verantwortung ist nicht nur ethischer Natur, sondern auch rechtlich verankert – etwa im DSA (Digital Service Act), das Plattformbetreiber:innen zur schnellen Entfernung rechtswidriger Inhalte verpflichtet.

Sexistische Hate Speech hat weitreichende Folgen für die betroffenen Frauen im virtuellen und analogen Raum: Verschiedene Studien zeigen, dass viele Frauen nicht nur regelmäßig sexistische Hate Speech erleben, sondern dass diese Erfahrungen zu Verunsicherung und einem Rückzug aus digitalen Diskursen führen können (vgl. HateAid/The Landecker Digital Justice Movement 2021; Bauer/Hartmann 2021). Sexistische Hate Speech kann darüber hinaus auch zu psychischen Belastungen wie emotionalem Stress oder Angst um die körperliche Sicherheit führen (vgl. Geschke et al. 2019). Eine weitere Folge sexistischer Hate Speech ist das Victim Blaming (vgl. Bauer/Hartmann 2021). Es verstärkt den Umstand, dass Frauen von Institutionen oder der Polizei nicht ernst genommen werden und ihnen dadurch Hilfe verwehrt wird.

Die Folgen wachsender Hate Speech sind auch gesamtgesellschaftlich betrachtet weitreichend: Sexistische Hate Speech wirkt nicht nur als Form digitaler Gewalt, sondern kann langfristige psychische und soziale Schäden bei Betroffenen verursachen. Sie unterminiert fundamentale Grundrechte wie die Menschenwürde, Gleichberechtigung und Meinungsfreiheit, indem sie Frauen systematisch diskriminiert und aus öffentlichen Diskursen verdrängt, was sich auf die gesamte Gesellschaft auswirkt (vgl. van der Wilk 2018). Hate Speech wird insbesondere aber in den Sozialen Medien häufig als geschlechtsneutrales Phänomen und als individuelle Angelegenheit dargestellt, die häufig auf die Naivität von Frauen zurückzuführen ist, denen die alleinige Verantwortung dafür zugeschrieben wird (vgl. Lang 2023). Frauen wird häufig dazu geraten, den Täter:innen keine Plattform zu bieten, ihre Privatsphäre-Einstellungen auf den Plattformen zu ändern oder für eine Weile offline zu gehen. Dieser Rat spiegelt die Normalisierung sexistischer Hate Speech wider und trägt gleichzeitig dazu bei, die Perspektive der betroffenen Frauen auszublenden (vgl. van der Wilk 2018). Durch den Rückzug von Frauen aus sozialen Kommunikationsplattformen wie Instagram wird das Thema sexistische Hate Speech für Nichtbetroffene weniger sichtbar und in ihrer Bedeutung verdrängt (vgl. Schneiders 2021).

3. Die Plattform Instagram – eine empirische Untersuchung sexistischer Hate Speech

Instagram ist ein Soziales Online-Netzwerk, das es ermöglicht, Fotos und Videos zu erstellen und diese mit anderen zu teilen. Die Plattform verzeichnete im Oktober 2025 weltweit 3 Milliarden User:innen. 47,3 Prozent der Instagram Nutzer:innen weltweit sind weiblich, 52,7 Prozent männlich. Die Altersgruppe der 16- bis 34-Jährigen bildet den Kern der Instagram Nutzer:innen (vgl. Arezo 2025).

Die Plattform weist neben kreativem User:innen-Content und Interaktion ein hohes Maß an sexistischer Hate Speech auf: In einer Studie des Kompetenznetzwerks gegen Hass im Netz gaben 38 Prozent der 3.061 Teilnehmer:innen an, dass sie Hate Speech auf Instagram wahrnehmen (vgl. Bernhard/Ickstadt 2024). Vor allem junge Frauen scheinen auf Instagram besonders häufig Opfer sexistischer Hate Speech und Belästigung zu sein (vgl. Plan International 2020; Center for Countering Digital Hate Inc. 2022).

Mittels einer selbst durchgeführten Online-Umfrage sollte die Wahrnehmung sexistischer Hate Speech auf Instagram untersucht und eruiert werden, inwiefern sexistische Hate Speech den Alltag von Frauen beeinflusst. Die Befragung sollte auch aufzeigen, wie die Plattform mit sexistischer Hate Speech umgeht und ob diese Maßnahmen sexistische Hate Speech effektiv eindämmen. Abschließend wurden die Wünsche der Teilnehmer:innen qualitativ ausgewertet (vgl. Glaser/Laudel 2010). Die Umfrage fand im August 2024 statt. Die Online-Umfrage wurde über das Tool Google Docs generiert und als Link auf verschiedenen Accounts von Instagram geteilt, um die Teilnahme zu fördern. Dabei wurde unter anderem die Story-Funktion auf Instagram genutzt, über die der Link zur Umfrage verbreitet wurde. Dadurch konnten gezielt Personen angesprochen werden, die regelmäßig auf Instagram aktiv sind und somit der Zielgruppe der Untersuchung angehören. Neben der Verbreitung über die genannten Sozialen Kommunikationsplattformen wurden auch persönliche Netzwerke genutzt, um die Umfrage weiter zu verbreiten. Es nahmen 153 Personen an der Online-Umfrage teil, die regelmäßig auf Instagram aktiv waren. Davon waren 59,5 Prozent weiblich, 37,9 Prozent männlich und 2,6 Prozent divers. Von den 153 Personen waren 86,3 Prozent zwischen 18 bis 35 Jahre alt, 7,8 Prozent 13 bis 17 Jahre alt und 5,9 Prozent 36 bis 65 Jahre alt.

3.1. Formen und Verbreitung sexistischer Hate Speech auf Instagram und Folgen für den Alltag von Frauen

Die durchgeführte Umfrage bestätigte die Befunde der zuvor zitierten Studien, wonach die Teilnehmer:innen auf Instagram Hate Speech in hoher Intensität wahrnahmen (92,2 Prozent): „Instagram ist leider die Plattform auf der Frauen am meisten belästigt werden – oft sind es sexuelle Motive." „Auf Instagram gibt es mehr Drohungen, in den Kommentaren oder Nachrichten gegenüber Frauen."

Die weiblichen Teilnehmerinnen der Umfrage nahmen die sexistische Hate Speech auf Instagram nicht nur besonders häufig, sondern auch sehr gewaltvoll wahr (58,2 Prozent der Teilnehmer:innen). Die Antwort einer Teilnehmerin verdeutlicht dies: „[…] auf Instagram sind die Kommentare gewaltverherrlichend. Frauen bekommen unter anderem Morddrohungen, weil sie Frauen sind […]". Auf die Frage, wie die Teilnehmerinnen Hate Speech konkret wahrnehmen, wurden im Freitext verschiedene Ausprägungen genannt, darunter Morddrohungen, Vergewaltigungsandrohungen gegenüber Frauen in den öffentlichen Kommentaren und Direktnachrichten (vgl. Bauer/Hartmann 2021).

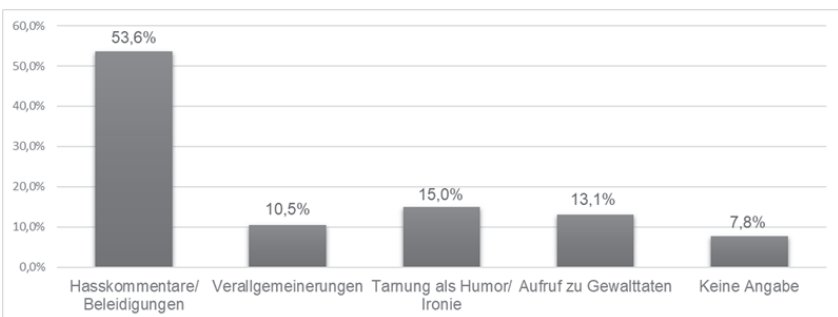

Abb. 1: Wahrgenommene Formen der Hate Speech auf Instagram in Prozent

Als häufigste Formen der sexistischen Hate Speech auf Instagram wurden in der durchgeführten Befragung Hasskommentare und Beleidigungen genannt (53,6 Prozent). Dies bestätigte die zuvor genannten Ergebnisse der Befragungen, wonach die häufigste Form sexistischer Hate Speech Hasskommentare und Beleidigungen waren (vgl. Center for Countering Digital Hate Inc. 2022; Stecher et al. 2020). Hate Speech getarnt in Form von

Humor oder Ironie wurde wiederum häufiger genannt als Verallgemeinerungen.

Auf die Frage, wie sich die Teilnehmer:innen die Entstehung von Hate Speech auf Instagram erklären, antworteten einige, dass sie diese auf das Verhalten älterer Nutzer:innen zurückführten. „Instagram bietet eine größere Fläche für Hasskommentare, weil auch viele ältere Menschen diese Plattform benutzen, die ein festgefahrenes, leider auch falsches Frauenbild mit sich bringen", formulierte es eine:r der Teilnehmer:innen. Dies legt nahe, dass Altersunterschiede eine Rolle bei der Entstehung und Wahrnehmung von sexistischer Hate Speech spielen können. Ebenfalls wurde eine rechtsradikale politische Einstellung als weiterer verstärkender Faktor genannt. Antworten wie „Auf Instagram verbreiten rechtsradikale Gruppierungen sexistische Einstellungen […]" lassen vermuten, dass die Grenzen rechtspopulistischer und sexistischer Hassrede auf Instagram fließend sind. Wie zuvor aufgeführt, tragen soziale Plattformen wie Instagram dazu bei, dass sich Gleichdenkende leichter austauschen und somit Gruppierungen bilden können, die Hass verbreiten (vgl. Lang 2018). Dies könnten im Falle dieser Befragung auch Gleichaltrige oder politisch Gleichgesinnte sein. Die Homogenität innerhalb dieser Gruppen, die die Teilnehmer:innen in dieser Umfrage angaben, könnte als Verstärker für sexistische Hate Speech dienen (vgl. Barker/Jurasz 2019).

Technische Unterschiede der Plattformen wurden ebenfalls als Ursache für die gewaltvolle sexistische Hate Speech genannt. Die Teilnehmer:innen gaben an, dass Instagram-Funktionen wie das Kommentieren von Bildern oder das Versenden von Direktnachrichten die sexuelle Belästigung von Frauen, einschließlich des Versendens intimer Fotos, begünstigten. Antworten wie „[…] Die Plattform Instagram bietet mehrere Möglichkeiten Frauen zu schreiben und mit etwas zu konfrontieren" oder „Auf Instagram bekommen Frauen ungefragt sexuelle Bilder zugeschickt. Auf anderen Plattformen ist das eher weniger verbreitet. Dort muss man Personen erst folgen und angenommen werden" verdeutlichen dies und lassen darauf schließen, dass Instagram eine engere und persönlichere Form der Kommunikation fördert, die es wiederum ermöglicht, einzelne Nutzer:innen direkt und gezielt anzugreifen.

Eine weiterführende Frage zielte darauf ab, zu untersuchen, ob sexistische Hate Speech auf Instagram insgesamt zu einer Zunahme sexistischer Hate Speech im Alltag führt. Die Ergebnisse zeigen, dass über die Hälfte der Teilnehmer:innen (57,5 Prozent) dies bejahten und weitere 29,4 Prozent dies eher bejahten (vgl. Barker/Jurasz 2019). Auf die Frage, ob die Präsenz

sexistischer Hate Speech auf Instagram dazu führe, dass sexistische Einstellungen eher akzeptiert würden, stimmten in der Umfrage 51,6 Prozent der Teilnehmer:innen zu und 31,4 Prozent stimmten eher zu. Die Ergebnisse der aufgeführten Umfrageergebnisse lassen also vermuten, dass digitale Räume und die dort vorherrschende sexistische Hate Speech nicht isoliert von gesellschaftlichen Normen betrachtet werden können, sondern aktiv zur Normalisierung sexistischer Einstellungen beitragen. Angesichts der Tatsache, dass die Anzahl der weltweiten Nutzer:innen sozialer Plattformen in den letzten fünf Jahren auf 5,24 Milliarden gestiegen ist, wird der weitreichende Einfluss dieser Plattformen auf die Normalisierung sexistischer Verhaltensweisen umso deutlicher.

3.2. Maßnahmen der Plattformbetreiber:innen und Wahrnehmung

Insbesondere unter ethischen Gesichtspunkten ist die Entwicklung und Implementierung von Richtlinien, die sexistische Hate Speech erfassen und wirksam bekämpfen, von zentraler Bedeutung. Ergänzend dazu spielt die Förderung der Medienkompetenz eine entscheidende Rolle, um ein reflektiertes und kritisches Bewusstsein im Umgang mit diskriminierenden Sprachmustern zu schaffen. Die Instagram-Plattformbetreiber:innen gaben an, im Jahr 2020 gegen 6,5 Millionen Inhalte von Hate Speech vorgegangen zu sein, einschließlich der privaten Nachrichten, von denen 95 Prozent gefunden wurden, bevor sie jemand meldete (vgl. Instagram 2021). Jedoch werden Sexismus und Frauenfeindlichkeit in ihren Statements nicht explizit erwähnt. Die Institutionen ISD und UltraViolet prüften die bestehenden Richtlinien von Instagram und stellten fest, dass die Plattform kein nachhaltiges System zur Eindämmung sexistischer Hate Speech besaß. Das negative Ergebnis (Note F) begründeten die Prüfer:innen vor allem mit den Schwächen des Instagram-Algorithmus und des Filters, welche Frauenfeindlichkeit und Bodyshaming verstärken würden (vgl. Institute for Strategic Dialogue/UltraViolet 2021). Weitere Studien zeigten, dass viele gemeldete Inhalte nicht gelöscht wurden. Von 61 gemeldeten Konten, die User:innen Todesdrohungen schickten, ahndete Instagram sieben Konten. Auf 89,7 Prozent der Konten, die den Teilnehmer:innen sexistische Nachrichten schickten, wurde nicht reagiert (vgl. Center for Countering Digital Hate Inc. 2022).

Fraglich bleibt, wie die sexistische Hate Speech auf Plattformen wie Instagram eingedämmt werden kann und welche Verantwortung sowohl die

Plattformbetreiber:innen als auch die Regierung tragen. Mehr als die Hälfte der Teilnehmer:innen (56,2 Prozent) sah die strafrechtliche Verfolgung der Verfasser:innen von sexistischer Hate Speech als wirksamste Gegenmaßnahme an. Nur 24,2 Prozent der Teilnehmer:innen hielten das Melden bei den Plattformbetreiber:innen für eine wirksame Maßnahme. Dies lässt darauf schließen, dass die bestehenden Melde- und Moderationssysteme der Plattform Instagram von den Teilnehmer:innen als unzureichend empfunden werden und offenbar ein mangelndes Vertrauen in die Reaktionsfähigkeit und Durchsetzungskraft der Plattform besteht. Trotz der hohen Bereitschaft der Teilnehmer:innen, selbst Gegenmaßnahmen zu ergreifen (71,9 Prozent), gaben davon 57,3 Prozent an, dass ihre Maßnahmen gar nicht erfolgreich waren. Die Ergebnisse deuten darauf hin, dass sich die Teilnehmer:innen bei ihrem Vorgehen gegen sexistische Hate Speech nicht ausreichend wahrgenommen fühlten.

Auf die Frage, welche Maßnahmen sich die Teilnehmer:innen gegen sexistische Hate Speech wünschen würden, wurden vor allem härtere Sanktionen wie „[…] Sperrung des Nutzers und Bußgelder" genannt. Außerdem plädierten die Teilnehmer:innen für bessere technische Maßnahmen, insbesondere durch effektivere Filter und menschliche Moderatoren. Ebenfalls wurde die Bedeutung von Aufklärungsarbeit der Nutzer:innen und Transparenz der Meldeprozesse hervorgehoben, um das Bewusstsein für die Auswirkungen von sexistischer Hate Speech zu schärfen und die Betroffenen ernst zu nehmen. Die geäußerte Forderung nach geringerer Anonymität und erhöhter Transparenz seitens der Plattformen gegenüber den Nutzenden unterstreicht den Wunsch nach regulativen Maßnahmen und verweist zugleich auf die wachsende medienethische Relevanz des Themas. Ebenso sehen einige der Befragten in einer verbindlichen Identitätsverifizierung eine wirksame Strategie, um sowohl gegen die Anonymität von Hate-Speech-Autor:innen vorzugehen als auch manipulative Fake-Profile einzudämmen.

Diese Ansätze verknüpfen folglich technische und soziale Maßnahmen, um sexistischer Hate Speech sowohl präventiv als auch reaktiv entgegenzuwirken.

Die Ergebnisse zeigen, dass nahezu alle Teilnehmer:innen die Maßnahmen der Plattformbetreiber:innen und der Regierung gegen sexistische Hate Speech als unzureichend empfanden. Es zeigt sich ein leichter Unterschied in der Kritik an der Plattform Instagram im Vergleich zur Regierung. 88,9 Prozent der Teilnehmer:innen waren der Ansicht, dass die Plattform nicht genug gegen sexistische Hate Speech vorgehen würde, während 83

Prozent dies bei der Regierung kritisierten. Es steht zu vermuten, dass die Plattform Instagram möglicherweise stärker kritisiert wurde als die Regierung, da Nutzer:innen den Plattformbetreiber:innen direktere Einflussmöglichkeiten unterstellten.

Sowohl der Bericht von ISD und UltraViolet als auch die Ergebnisse der durchgeführten Umfrage zeigen, dass die bereits getroffenen Maßnahmen der Plattform Instagram nicht ausreichen, um sexistische Hate Speech wirksam zu bekämpfen. Forschungsarbeiten hierzu betonten, dass von Sozialen Kommunikationsplattformen ein aktives Handeln notwendig sei, um sich demonstrativ gegen die Verbreitung von Hass zu positionieren. Als weitere Maßnahmen wurden Kampagnen oder Initiativen zur Steigerung der Medienkompetenz genannt, um Internetnutzer:innen in die Lage zu versetzen, angemessen auf Hate Speech zu reagieren (vgl. Gagliardone et al. 2015).

Die Entwicklung und Durchsetzung von Richtlinien, die (sexistische) Hate Speech bekämpfen und Medienkompetenz fördern, sind entscheidende Maßnahmen, um sich aktiv dagegen zu positionieren. Der *Digital Services Act* (DSA) etwa verpflichtet insbesondere Plattformen mit großer Reichweite, benutzerfreundliche Meldekanäle für illegale Inhalte bereitzustellen und offensichtlich rechtswidrige Inhalte umgehend zu löschen (vgl. Kahl/Liepert 2022). Die Plattformen sind auch gehalten, ein internes Beschwerdemanagementsystem umzusetzen sowie regelmäßige, öffentlich zugängliche Transparenzberichte über den Umgang mit gemeldeten illegalen Inhalten (zum Beispiel Hate Speech) auszustellen. Inwieweit dies seitens der Plattformen vollständig umgesetzt wird, bleibt abzuwarten.

4. Fazit

Zusammenfassend zeigt sich, dass sexistische Hate Speech und deren Ahndung auf sozialen Plattformen wie Instagram eine vielschichtige Herausforderung darstellt, die weit über technische Gegebenheiten hinausgeht. Es konnte festgestellt werden, dass Sprache verletzen kann, insbesondere wenn sie in sozialen Machtgefällen, genutzt wird. Sie formt gesellschaftliche Normen und kann auch in der digitalen Welt zur Verstärkung von Diskriminierung beitragen. Die Anonymität und Reichweite dieser Plattformen erleichtern die Verbreitung von Hassbotschaften on- wie offline. Ein zentraler ethischer Aspekt ist dabei das Spannungsfeld zwischen Meinungsfreiheit und Schutz vor Diskriminierung: Soziale Medien begünstigen nicht nur freie Meinungsäußerung, sondern ebenfalls die Verbreitung von Hate

Speech, einschließlich sexistischer Hate Speech. Plattformen sollten keine übermäßige Zensur ausüben, sie sollten dennoch sicherstellen, dass die digitalen Räume nicht zur Verbreitung von Hate Speech genutzt werden. Die Plattformbetreiber:innen sind gehalten, Stellung zu beziehen, Hate Speech zu ahnden und die Verantwortung für eine wirksame Regulierung sexistischer Hate Speech zu übernehmen. Medienethisch betrachtet verdeutlichen die Befunde, dass Plattformen nicht nur technische, sondern auch normative Verantwortung tragen, um digitale Räume als sichere Kommunikationsumgebungen zu gestalten. Die Normalisierung sexistischer Hate Speech unterminiert Grundwerte wie Menschenwürde und Gleichberechtigung, was eine aktive ethische Positionierung der Plattformbetreiber zwingend erforderlich macht.

Um ein vollständiges Bild sexistischer Hate Speech und möglicher Gegenmaßnahmen zu erhalten, sind weitere Untersuchungen notwendig. Auch wenn durch diesen Beitrag einige Erkenntnisse hinsichtlich der Wahrnehmung sexistischer Hate Speech und möglicher Gegenmaßnahmen herausgestellt werden konnten, zeigen sich Limitationen. Bei der gewählten Stichprobe dieser Umfrage handelte es sich um eine Gelegenheitsstichprobe, die auf der willkürlichen Auswahl von Personen basierte. Dies führte dazu, dass die Ergebnisse nicht auf alle Nutzer:innen von Instagram übertragen werden können. Bei den Fragen nach dem Einfluss sexistischer Hate Speech auf den Alltag von Frauen war außerdem die Tiefe der Fragestellung durch die quantitative Erhebung begrenzt. Daher wäre es notwendig, weitere Forschungen zu dieser Thematik anzustoßen, besonders in einem qualitativen Kontext. Durch qualitative Methoden wie Interviews oder Fokusgruppen könnten die Erfahrungen der Betroffenen detaillierter erfasst und die gestellten Fragen tiefgreifender formuliert werden. Des Weiteren könnten sich weitere Studien mit möglichen Maßnahmen zur Eindämmung sexistischer Hate Speech auseinandersetzen. Trotz erster Fortschritte durch beispielsweise den DSA bleibt die Wirksamkeit bestehender Maßnahmen der Plattformen fraglich, weshalb kontinuierliche Forschung und Maßnahmenkontrolle in diesem Bereich erforderlich sind.

Literatur

Arezo (2025): Aktuelle Social Media Nutzerzahlen (Stand Oktober 2025), in: Zweidigital (online unter: https://www.zweidigital.de/aktuelle-social-media-nutzerzahlen/ – letzter Zugriff: 3.12.2025).

Barker, Kim / Jurasz, Olga (2019): Online Misogyny: A Challenge for Digital Feminism?, in: Journal Of International Affairs 72 (2), S. 95–114.

Bauer, Jenny-Kerstin / Hartmann, Ans (2021): Formen digitaler geschlechtsspezifischer Gewalt, in: Nivedita Prasad (Hg.), Geschlechtsspezifische Gewalt in Zeiten der Digitalisierung: Formen und Interventionsstrategien, Bielefeld, S. 63–100.

Bernhard, Lukas / Ickstadt, Lutz (2024): Lauter Hass – leiser Rückzug: Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung, Berlin (online unter: https://kompetenznetzwerk-hass-im-netz.de/download_lauterhass.php – letzter Zugriff: 3.12.2025).

Center for Countering Digital Hate Inc. (2022): Hidden Hate: How Instagram fails to act on 9 in 10 reports of misogyny in DMs. Washington D.C.

Dellagiacoma, Laura / Sika Dede Puhlmann, Francesca (2025): Rassistische Hate Speech im Alltag: Erfahrungen schwarzer Menschen in Deutschland, Jena (online unter: https://www.idz-jena.de/fileadmin/user_upload/Projektberichte/IDZ_Nethate_B5_WEB_FINAL.pdf – letzter Zugriff: 3.12.2025).

Eckert, Stine (2017): Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States, in: New Media & Society 20 (4/2017), S. 1282–1302.

Embacher, Andrea (2021): Hate Speech gegen Frauen – über sprachliche Gewalt im Internet (Masterarbeit), Wien.

European Union Agency for Fundamental Rights (FRA) (2023): Online Content Moderation – Current challenges in detecting hate speech, Wien.

Felling, Matthias et al. (2019): Hate Speech – Hass im Netz. Informationen für Fachkräfte und Eltern, Düsseldorf.

Gagliardone, Iginio et al. (2015): Countering Online Hate Speech (= UNESCO Series on Internet Freedom), Paris.

Geschke, Daniel et al. (2019): #Hass im Netz: Der schleichende Angriff auf unsere Demokratie. Eine bundesweite repräsentative Untersuchung, Jena.

Glaser, Jochen / Laudel, Grit (2010): Experteninterviews und qualitative Inhaltsanalyse, Wiesbaden.

HateAid / The Landecker Digital Justice Movement (2021): Grenzenloser Hass im Internet – Dramatische Lage in ganz Europa, Berlin.

Herrmann, Steffen Kitty / Kuch, Hannes (2007): Symbolische Verletzbarkeit und sprachliche Gewalt, in: Steffen Kitty Herrmann / Sybille Krämer / Hannes Kuch (Hg.), Verletzende Worte: Die Grammatik sprachlicher Missachtung, Bielefeld, S. 179–210.

Huber, Brigitte (2023): Social Media – eine kommunikationswissenschaftliche Perspektive, in: Anne-Kathrin Langner / Gabriele Schuster (Hg.), Holistische Social-Media-Strategien, Wiesbaden, S. 27–35.

Instagram (2021): An update on our work to tackle abuse on Instagram, in: Instagram, 11. Februar 2021, (online unter: https://about.instagram.com/blog/tackling-abuse – letzter Zugriff: 11.6.2025).

Institute for Strategic Dialogue / UltraViolet (2021): Social media fails women: The report card, London.

Inter-Parliamentary Union (2018): Sexism, harassment and violence against women in parliaments in Europe, Genf.

Kahl, Jonas / Liepert, Simon (2022): Digital Services Act: Was sich gegenüber dem NetzDG ändert, in: heise online, 09. Dezember 2022 (online unter: https://www.heise.de/hintergrund/Digital-Services-Act-Was-sich-gegenueber-dem-NetzDG-aendert-7367625.html – letzter Zugriff: 3.12.2025).

Kaspar, Kai (2017): Hassreden im Internet – Ein besonderes Phänomen computervermittelter Kommunikation? in: Kai Kaspar / Lars Gräßer / Aycha Riffi (Hg.), Online hate speech: Perspektiven auf eine neue Form des Hasses, Marl, S. 62–70.

Lang, Andrej (2018): Netzwerkdurchsetzungsgesetz und Meinungsfreiheit, in: Archiv des öffentlichen Rechts 143 (2), S. 220.

Lang, Raphaela (2023): Überheblich, hässlich, dumm – Geschlechtsbezogene Hate Speech in sozialen Medien, in: Sumo, 15. Mai 2023 (online unter: https://www.sumomag.at/ueberheblich-haesslich-dumm-geschlechtsbezogene-hate-speech-in-sozialen-medien/ – letzter Zugriff: 3.12.2025).

Medienpädagogischer Forschungsverbund Südwest (2025): JIM-Studie 2025. Jugend, Information, Medien. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger, Stuttgart.

Mohseni, M. Rohangis (2021): Sexistische Online-Hassrede auf Video-Plattformen, in: Barbara Koch-Priewe / Sebastian Wachs / Andreas Zick (Hg.), Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen, Wiesbaden, S. 39–51.

Pandea, Anca-Ruxandra / Grzemny, Dariusz / Keen, Ellie (2019): Gender Matters: A Manual on Addressing Gender-based Violence Affecting Young People, 2. Aufl., Budapest.

Plan International (2020): Free to be online? Girls' and young women's experiences of online harassment (= State of the World's Girls 2020), Surrey.

Powell, Anastasia et al. (2022): A multi-country study of image-based sexual abuse: extent, relational nature and correlates of victimisation experiences, in: Journal of Sexual Aggression 30 (1), S. 25–40.

Riemenschneider, Severin / Lutz, Marina (2021): #HateSpeech – Shitstorms als Kampfmittel organisierter Strukturen, in: Datenschutz und Datensicherheit 45 (6), S. 371–374.

Robertz, Frank J. / Oksanen, Atte / Räsänen, Pekka (2016): Viktimisierung junger Menschen im Internet. Leitfaden für Pädagogen und Psychologen, Wiesbaden.

Römer-Pieretti, Max / Esteban-Ramiro, Beatriz / Canelón Silva, Agrivalca (2025): Violence, Hate Speech, and Gender Bias: Challenges to an Inclusive Digital Environment, in: Social Inclusion 13 (1), S. 1–5.

Schieb, Carla (2021): Zwischen Like und Empörung: Theoretische Modellierung und Empirische Untersuchung von Rezeptions- und Persuasionswirkungen von Sexistischem Hate Speech, Münster.

Schneiders, Pascal (2021): Hate Speech auf Online-Plattformen, in: UFITA – Archiv für Medienrecht und Medienwissenschaft 85 (2), S. 269–333.

Sponholz, Liriam (2021): Hass mit Likes: Hate Speech als Kommunikationsform in den Social Media, in: Barbara Koch-Priewe / Sebastian Wachs / Andreas Zick (Hg.), Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen. Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen, Heidelberg, S. 15–24.

Stark, Birgit / Magin, Melanie / Geiß, Stefan (2022): Meinungsbildung in und mit sozialen Medien, in: Jan-Hinrik Schmidt / Monika Taddicken (Hg.), Handbuch Soziale Medien, Wiesbaden, S. 213–231.

Stecher, Sina et al. (2020): „Du bist voll unbekannt!" Selbstdarstellung, Erfolgsdruck und Interaktionsrisiken auf TikTok aus Sicht von 12- bis 14-Jährigen. Ausgewählte Ergebnisse der Monitoring-Studie (= ACT ON! Short Report Nr. 7.), München.

Suler, John (2004): The Online Disinhibition Effect, in: CyberPsychology & Behavior 7 (3), S. 321–326.

van der Wilk, Adriane (2018): Cyber violence and hate speech online against women (= Study for the European Parliament's Committee on Women's Rights and Gender Equality), (online unter: https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604979/IPOL_STU(2018)604979_EN.pdf – letzter Zugriff: 3.12.2025).

Zeller, Frauke (2017): Soziale Medien in der empirischen Forschung, in: Jan-Hinrik Schmidt / Monika Taddicken (Hg.), Handbuch Soziale Medien, Wiesbaden, S. 389–408.

Zick, Andreas et al. (2008): The Syndrome of Group-Focused Enmity: The Interrelation of Prejudices Tested with Multiple Cross-Sectional and Panel Data, in: Journal of Social Issues 64 (2), S. 363–383.

Die Positionen von Redaktionen zur Verwendung gendersensibler Sprache

Beatrice Dernbach

Zusammenfassung

Seit Jahren dauert die Debatte um gendersensible Sprache in Deutschland an, insbesondere über die Verwendung von Sonderzeichen. Sie ist stark polarisiert. Einzelne Politiker, allen voran der bayerische Ministerpräsident Markus Söder, regeln die Verwendung von Sternchen und Doppelpunkten in staatlichen Einrichtungen qua Verordnung und Verbot. Daran müssen sich Medienredaktionen nicht halten. Wie bereits im Zuge der großen Rechtschreibreform in den 1990er-Jahren können sie selbst entscheiden, wie ihre Mitarbeitenden schreiben und sprechen. Es lassen sich unterschiedliche Positionen erkennen, die an eine politische Ausrichtung gekoppelt sind. Grundlegend für die Sprachverwendung einerseits und die Berichterstattung über das Thema andererseits ist die ethisch-normative Bewertung der Rolle der Medien als Beobachter der gesellschaftlichen Entwicklung. In diesem Beitrag werden diese unterschiedlichen Positionen analysiert. Betrachtet wird dabei auch, welche Auswirkungen die Verwendung KI-basierter Sprachtools haben könnten. Am Ende steht der Appell für eine Entpolemisierung der Debatte.

1. Einleitung: Warum diese Aufregung?

In den vergangenen Jahren kochte ein gesellschaftspolitisches Thema besonders hoch: Die Verwendung gendersensibler Sprache, insbesondere von Sonderzeichen wie das sogenannte Gendersternchen.[1] Um im Bild zu blei-

1 Die Autorin ist Diplom-Germanistin und Redakteurin, weshalb sie sich an den Empfehlungen des Rates für deutsche Rechtschreibung orientiert und nicht (mit Sonderzeichen) gendert, sondern in der Regel aus Gründen der Lesbarkeit das (grammatikalisch begründete) Generische Maskulinum verwendet. Es geht in dem Beitrag nicht um eine politische oder gar ideologische Positionierung, sondern um eine sachliche Analyse der Debatte.

ben: Politisch lief das Fass – nicht nur in Bayern – beinahe über, als Ministerpräsident Markus Söder allen bayerischen staatlichen Einrichtungen das sogenannte Gendern ab dem 1. April 2024 ausdrücklich verboten hatte (vgl. Bayerische Staatsregierung 2023). Der Freistaat steht damit nicht allein: Auch in den Bundesländern mit einer CDU-geführten Regierung Sachsen, Sachsen-Anhalt, Schleswig-Holstein und Hessen gelten entsprechende Regeln beziehungsweise Verbote, besonders in Schulen und der öffentlichen Verwaltung.

Bereits am 24. Juni 2021 lehnte der Deutsche Bundestag einen Antrag der AfD-Fraktion mit großer Mehrheit ab, in dem diese sich gegen die „sogenannte[n] gendergerechte[n] Sprache durch die Bundesregierung" sowie in „Drucksachen des Bundestages" ausgesprochen hatte (AfD 2021). Begründet hatten die AfD-Parlamentarier ihren Vorstoß mit der „unnatürlichen Verunstaltung der deutschen Sprache", mit der „ihre Verständlichkeit erheblich eingeschränkt werde" (ebd.). Die Gegenposition nimmt entsprechend ihrer politischen Grundposition die Partei Bündnis 90/Die Grünen ein. Die Delegierten haben bei ihrer Bundeskonferenz im November 2015 einen Leitfaden zur geschlechtergerechten Sprache beschlossen. Die Münchner Grünen schwächen etwas ab und deklarieren, dass Gendersprache kein Zwang und kein Gesetz sei – es aber schön wäre, wenn sich niemand darüber aufrege, wenn Menschen sie verwenden (Bündnis 90/Die Grünen o.J.).

Die Medien als nicht-staatliche Einrichtungen waren und sind von gesetzlichen Regelungen nicht direkt betroffen, berichteten aber häufig (vgl. Waldendorf 2023). Das Thema erweiterte und verlagerte sich von einer linguistischen, pädagogischen und ethischen zu einer sehr polarisierten und polarisierenden politischen Auseinandersetzung. Das grundlegend Herausfordernde an der Debatte war und ist, dass es um das Kulturgut Sprache geht, das prägend ist für die generelle gesellschaftliche Entwicklung. Inwieweit feste linguistische Regeln, also Lexik und Semantik, Grammatik und Syntax gelten, oder im Gegenteil veränderbar sein müssen, um soziale Diversifizierung angemessen aufgreifen zu können, war, ist und bleibt der wesentliche Triggerpunkt (vgl. Dernbach 2024 und 2025).

Die wichtigsten Auslöser und Faktoren für die Debatte über genderinklusive Sprache sind zusammengefasst:

– Gleichberechtigungsbewegungen und Gender-Forschung fokussieren seit den 1980er-Jahren auf gesellschaftliche Ungleichheiten. Insbesondere betrachten die Vertreterinnen der Gender-Studies die Rolle der Sprache

als Ausdruck und Verstärker von Ungleichheiten. Sie sei nicht neutral, sondern spiegele, präge und festige Machtstrukturen, Geschlechterverhältnisse und Stereotype (vgl. Trömel-Plötz 1982; Pusch 1984).

- Kritik an der männlichen Norm in der Sprache: Unter anderem die emeritierte Professorin für Sozialpsychologie, Dagmar Stahlberg, hat mit Kolleginnen in mehreren experimentellen Studien starke Belege gefunden, dass sich bei genderinklusiven Formen wie „Politiker:innen" Studienteilnehmer häufiger an Frauen erinnern als beim generischen Maskulinum – was zeige, dass Sprache reale mentale Repräsentationen beeinflusse (vgl. Stahlberg unter anderem 2001; siehe auch Braun unter anderem 2005). Die psycholinguistische Forschung (vgl. Cassaris 2025) spricht von einem „Bias" beim generischen Maskulinum, der unbewusst stereotype Wahrnehmungen verstärke; es rufe männliche Assoziationen hervor, mache Frauen und nicht-binäre Personen unsichtbar. Deshalb sollen sie explizit in der Sprache sichtbar gemacht werden.
- Die politischen und gesellschaftlichen Entwicklungen haben zur rechtlichen Gleichstellung geführt. Im deutschen Grundgesetz ist die Gleichberechtigung zwischen Frauen und Männern fixiert (Art 3 Abs. 2). Das Allgemeine Gleichbehandlungsgesetz (AGG, 2016) verbietet Diskriminierung aufgrund sexueller Orientierung und Geschlechtsidentität in verschiedenen Lebensbereichen. Mit dem Selbstbestimmungsgesetz (SBGG, 2024) sollen die Änderungen des Geschlechtseintrags und Vornamens für trans-, intergeschlechtliche und nichtbinäre Personen erleichtert werden. Damit sollen Inklusion und Diversität gefördert werden.
- Öffentliche Debatten und Medienberichte rund um dieses Thema haben zugenommen, wohl auch beeinflusst vom Auftritt prominenter Persönlichkeiten und Institutionen. Die Organisationen, die sich mit den Standards der Sprache beschäftigen, allen voran die Duden-Redaktion, der Rat für deutsche Rechtschreibung (RdR 2023) und die Gesellschaft für deutsche Sprache (GfdS 2020), haben dazu mehrfach öffentlich Stellung genommen.
- Die Veränderungen zeigen sich in allen gesellschaftlichen Handlungsfeldern, nicht zuletzt auch in der Politik. Müller-Spitzer und Ochs (2024) haben Debatten im Deutschen Bundestag seit den 1980er Jahren untersucht und herausgefunden, dass gendersensible Sprachformen – wie Paarformen („Bürgerinnen und Bürger") – kein künstlicher Eingriff seien, sondern Ausdruck sich verschiebender sozialer Normen.
- Die Bevölkerung bewertet das Thema völlig anders: Auf die Frage „Finden Sie geschlechtergerechte Sprache, sogenanntes Gendern, wichtig

oder unwichtig?“ (YouGov 2023) antworteten 69 Prozent der 3500 befragten Erwachsenen mit eher und sehr unwichtig. Mit Blick auf die Parteienpräferenz schwanken die Werte zwischen 85 Prozent (AfD-Präferenz), 67 Prozent (SPD, Die Linke) und 60 Prozent (Grüne). Selbst den 12- bis 25-Jährigen ist das Gendern entweder egal oder sie lehnen es ab (77 Prozent); Unterschiede gibt es zwischen jungen Frauen (65 Prozent) und Männern (87 Prozent) (vgl. Shell 2024).

2. Die normativen Grundlagen der Genderdebatte in den Medien

Normative Grundlagen der Sprachverwendung in den Medien beruhen auf rechtlichen, ethischen und berufsbezogenen Prinzipien, die eine ausgewogene, sachliche und verständliche Berichterstattung gewährleisten sollen. Zu den rechtlichen Rahmenbedingungen gehören das Grundgesetz, vor allem Artikel 5 (Pressefreiheit) sowie Artikel 1 und 2 (Persönlichkeitsrechte und Anti-Diskriminierung). Hinzu kommen Paragrafen aus dem Strafgesetzbuch wie §185 (Beleidigung) oder §130 (Volksverhetzung). Die ethischen Bedingungen sind im Pressekodex des Deutschen Presserates formuliert, insbesondere in „Ziffer 1: Wahrhaftigkeit und Achtung der Menschenwürde“ und „Ziffer 12: Vermeidung von Diskriminierung“. Generell wird von Medien gefordert, die ethische Verantwortung für die öffentliche Meinung und die gesellschaftliche Wirkung von Berichterstattung zu übernehmen.

Was bedeuten diese rechtlichen und normativen Setzungen für die Frage, wie groß die Schnittmenge beziehungsweise wie tief die Kluft ist zwischen dem Selbstverständnis eines Mediums als Spiegel der Gesellschaft und Moderator des gesellschaftlichen Wandels sowie der Anerkennung sprachlicher Normierung? Unabhängig von der Positionierung in Sachen gendergerechter Sprache zeigt sich eine Gemeinsamkeit bezüglich der Funktion von Sprache im Journalismus: Sie soll klar und verständlich sein, Fakten transparent darstellen und somit die Inhalte als notwendige Grundlagen von Information und Meinungsbildung bereitstellen. Auch geschlechtergerechte Texte – darin stimmen die oben genannten Institutionen überein – sollen sachlich korrekt, verständlich, lesbar und vorlesbar sein. Auch müssen sie die Rechtssicherheit und Eindeutigkeit in öffentlicher Verwaltung und Rechtspflege gewährleisten, möglichst automatisiert übertragbar sein in andere Sprachen und das Erlernen der geschriebenen deutschen Sprache nicht erschweren (vgl. RdR 2021 und 2023).

Zur Verwendung von Sonderzeichen hat sich die Gesellschaft für deutsche Sprache (GfdS) bereits 2020 klar geäußert: Sonderzeichen (unter anderem Genderstern, Doppelpunkt und Unterstrich) halten die Experten für ungeeignet im Sinne der Verständlichkeit; sie empfehlen hingegen die Nennung beider Formen, die Schrägstrich-Lösung oder Alternativen wie substantivierte Partizipien (zum Beispiel „Studierende" statt „Studenten"). Ähnlich positioniert sich der Rat für deutsche Rechtschreibung (RdR), der 2023 nochmals seine Entscheidung aus dem Jahr 2021 gegen das Gendern mit Sonderzeichen bekräftigt hat (vgl. RdR 2021 und 2023). Die Duden-Redaktion verweist darauf, dass die deutsche Sprache schon jetzt „eine Fülle an Möglichkeiten [bietet], geschlechtergerecht zu formulieren. Es gibt dafür allerdings keine Norm." Im Duden-Verlag sind bis dato sechs Publikationen zum Thema Gendern erschienen (unter anderem Diewald/Steinhauer 2022), die historischen und sprachwissenschaftliche Entwicklungen belegen, Strategien des gendergerechten Formulierens und deren Anwendung an konkreten Beispielen vermitteln.

Empirische Studien bestätigen, dass das Thema genderinklusive Sprache sowohl im intraredaktionellen Alltag als auch im Agenda-Setting der Berichterstattung angekommen ist. Anica Waldendorf (2023) belegte in ihrer Analyse von über vier Millionen Medienartikeln aus den Jahren 2000 bis 2021, dass genderinklusive Sprache in Deutschland stark zunehme und besonders in politisch links-orientierten Medien explizit nicht-binäre Formen zu finden seien. Diese Erkenntnis zieht sich durch viele weitere Untersuchungen, die übereinstimmend darauf hinweisen, dass Redaktionen sich entsprechend ihrer generellen politischen Ausrichtung mehr oder weniger pragmatisch positioniert haben: Die Kritiker von Sonderschreibweisen wie Sternchen und andere stammen eher aus dem konservativen Spektrum (zum Beispiel Springer Verlag), die Befürworter eher aus dem linken (zum Beispiel die taz) (vgl. Payr 2022: VII; Zylka/Grimberg 2021). Häufig wird die generelle Notwendigkeit einer gendersensiblen Haltung auch auf der liberalen und konservativen Seite betont, aber gleichzeitig darauf hingewiesen, dass aus pragmatischen und praktischen Gründen auf deren sprachliche Umsetzung verzichtet wird (dazu später mehr in Kapitel 4).

3. Die Sprache als journalistisches Handwerkszeug

Sprache ist Gegenstand primär der geistes- und sozialwissenschaftlichen Disziplinen, aber auch der Neurowissenschaften. Sie wird – unabhängig

von ihrer spezifischen Ausprägung – als ein zentrales Mittel betrachtet, mit dem Menschen ihre Umwelt kognitiv und sozial erschließen (vgl. Piaget 1992; Berger/Luckmann 1969). Der Erwerb von Sprache ist dabei nicht nur ein sozialer und kommunikativer Lernprozess, sondern zugleich ein kognitiver Mechanismus, durch den spezifische Denkmuster im Gehirn ausgebildet werden. Diese mentalen Strukturen können aktiviert, modifiziert und erweitert werden, um auf neue soziale oder individuelle Anforderungen zu reagieren (vgl. Lakoff 2016 sowie Lakoff/Johnson 1980). In diesem Zusammenhang fungiert Sprache als ein zentrales Instrument zur Reduktion der Komplexität der sozialen Lebenswelt (vgl. Dernbach 2019). Ursprünglich vollzog sich diese Erschließung primär über orale Kommunikationsformen, wurde jedoch im Verlauf der Menschheitsgeschichte wesentlich durch die Verschriftlichung erweitert, was neue Formen des Denkens, Erinnerns und gesellschaftlichen Handelns ermöglichte (vgl. Erfurt 1996 sowie Merten 1994).

Sprache ist im Journalismus nicht nur ein Transportmittel von Informationen, sondern erfüllt eine Vielzahl kommunikativer, kognitiver und gesellschaftlicher Funktionen. Als zentrales Werkzeug journalistischer Praxis dient Sprache der Selektion, Strukturierung und Bewertung von Wirklichkeit. Mit sprachlichen Mitteln wird nicht nur über Ereignisse berichtet, sondern zugleich ihre Relevanz konstruiert – ein Prozess, der als „Wirklichkeitskonstruktion" beschrieben wird (vgl. Löffelholz 2004: 64). Dabei operieren Journalisten innerhalb spezifischer sprachlicher Konventionen, die je nach Textsorte, Medium und Zielpublikum variieren (vgl. Dernbach 2019). In Nachrichten dominiert in der Regel eine sachlich-nüchterne Sprache, während in Kommentaren oder Glossen eine stärker wertende und expressive Ausdrucksweise zulässig ist oder sogar erwartet wird.

Zudem ist Sprache im Journalismus ein machtvolles Instrument der (politischen) Meinungsbildung. Sie kann Deutungsmuster etablieren, Frames setzen und Narrative formen, die langfristig gesellschaftliche Wahrnehmungen prägen (vgl. Entman 1993 sowie Wehling 2016). Insbesondere durch die Auswahl sprachlicher Bilder (Metaphern) und Begriffe kann eine Einflussnahme auf das Publikum erfolgen, ohne dass diese immer bewusst wahrgenommen wird. Die sprachliche Gestaltung journalistischer Texte steht damit in einem Spannungsverhältnis zwischen Objektivität und Subjektivität, zwischen Informationsvermittlung und Wirklichkeitsdeutung (vgl. unter anderem Steiner-Hämmerle 2023).

Journalistische Sprache ist also nie neutral, sondern stets geprägt durch Auswahl, Perspektive und sprachliche Gestaltungsmittel (vgl. Häusermann

1993; Kurz unter anderem 2010). Auch das Sprechen über journalistische Sprache ist nicht frei von Polemik und Provokation und lässt die im Hintergrund liegende gesellschaftspolitische Haltung durchschimmern. Der häufig in Medien als „Sprachpapst" (Heine 2022) titulierte Wolf Schneider, der an der Henri-Nannen-Schule hunderte von Journalisten ausgebildet und zahlreiche Sprach-Ratgeber verfasst hat, positioniert sich deutlich:

> „Die ganze Gender-Debatte ist eine Wichtigtuerei von Leuten, die von Sprache keine Ahnung haben. Zwischen dem natürlichen und dem grammatischen Geschlecht besteht nicht der geringste Zusammenhang. Wie könnte es sonst das Weib heißen? Der Löwe, die Schlange, das Pferd. Obwohl sie alle dieselben zwei Geschlechter haben. Die Führungskraft ist heute überwiegend ein Mann – und keiner hat sich je beschwert. Die Liebe ist weiblich, dabei soll es bleiben" (Geisler/Vehlewald 2022).

Statistisch betrachtet ist die Führungskraft heute immer noch ein Mann. Die am 11. Juni 2021 im Deutschen Bundestag beschlossene Frauenquote in der Privatwirtschaft und im öffentlichen Dienst verändert die Verhältnisse nur langsam.

Abgesehen von seiner Provokation haben Schneider und Dutzende anderer (wie Häusermann 2011, Linden 2000) klare Vorstellungen der Sprache als journalistischem Handwerkszeug (gehabt). Im Vordergrund stand das Konzept, mit einfacher Sprache komplexe politische, ökonomische und kulturelle Sachverhalte einfach und verständlich zu erklären. Dazu gehören vor allem:

- kein hochgestochener Nominalstil
- Adjektive in Maßen
- Doppelungen vermeiden
- aussagekräftige Verben verwenden
- präzise formulieren.

Diese Praxistipps stammen allerdings aus einer Zeit, als die gesellschaftliche und mediale öffentliche Debatte über das Gendern noch nicht richtig begonnen hatte – was sie nicht als völlig überaltert deklarieren, aber markieren soll, dass der Blick stärker normativ auf handwerklich-professionelle Regeln gerichtet war.

4. Die redaktionellen Binnenregeln im Hinblick auf Framing und Gendern

Auf die generellen Prämissen beziehen sich auch heute noch alle deutschen Medienredaktionen – legen sie allerdings im Hinblick auf die Verwendung gendergerechter Sprache sehr unterschiedlich aus. Dies wird sowohl in den jeweiligen Schreib- oder Sprechweisen als auch in den Kommentierungen zum Thema gendergerechte Sprache deutlich. Konservative Medien wie die Frankfurter Allgemeine Zeitung haben vor allem konservativen Linguisten eine mediale Plattform geboten (siehe unter anderem Eisenberg 2021; Ammer 2020; siehe auch Dernbach 2024). Publikationen, die eher als linksliberal gelten – allen voran die Süddeutsche Zeitung (vgl. Wittwer 2021) – haben sich in ihren Leitlinien bis dato an die Empfehlungen des Rates für deutsche Rechtschreibung (RdR) gehalten und keine Sonderzeichen eingesetzt oder präferieren Partizipbildungen (vgl. Bargmann 2020; Übersicht in Zylka/Grimberg 2021). „Die Befürworter stammen meist aus dem politisch linken Spektrum, die Kritiker aus dem rechten. Das zeigt, dass die Einstellung zum Gendern auch ideologisch geprägt ist – beim Pro oder Kontra geben leider nur in den seltensten Fällen sprachwissenschaftliche Überlegungen den Ausschlag" (Payr 2022: VII).

Generell sind aus den unterschiedlichen redaktionellen Leitlinien folgende Kriterien herauszukristallisieren:

1. Grad der Standardisierung:
   i) gendersensible Sprache mit Verwendung von Gendersternchen (*), Doppelnennungen (zum Beispiel „Leser und Leserinnen") oder neutrale Formulierungen wie „Lesende",
   ii) flexible Ansätze: Entscheidungen bleiben den Autoren überlassen, um ihren Schreibstil nicht einzuschränken,
   iii) konservative Position: Verwendung des generischen Maskulinums; linguistische Argumentation der Sprachentwicklung.
2. Sprachliche Lesbarkeit und Ästhetik:
   Redaktionen wägen ab, ob sie Verständlichkeit und Sprachästhetik oder Inklusivität priorisieren.
3. Zielgruppenspezifische Überlegungen:
   i) junges, progressives Publikum erwartet inklusive Sprache, die Diversität berücksichtigt,
   ii) konservative und ältere Leser empfinden gendersensible Sprache als störend.
4. Gesellschaftliche Verantwortung und Vorbildfunktion:

Redaktionen, die für sich eine zentrale Rolle in der öffentlichen Meinungsbildung einnehmen (zum Beispiel aufgrund der Reichweite), sehen einen Einfluss sprachlicher Entscheidungen auf die Wahrnehmung und Akzeptanz von (sozialen) Normen.

5. Technologische Herausforderungen:
   i) Umsetzung gendersensibler Sprache in digitalen Formaten (zum Beispiel Suchmaschinenoptimierung, Barrierefreiheit für Screenreader) ist für Menschen mit Sehbehinderungen schwer zu interpretieren,
   ii) Datenschutz.

In der folgenden Tabelle (Abbildung 1) wird ein Teil dieser Aspekte zusammengestellt und ergänzt um die Kriterien Position (gegenüber der Verwendung gendersensibler Sprache), Begründung für und gegen Sonderzeichen, mediales Framing in der Berichterstattung sowie Fokus auf spezielle journalistische, redaktionsinterne Regelungen und ihre Begründung. Eingearbeitet beziehungsweise im Ergebnis erkennbar ist die Komplexität des Themas, denn es ergeben sich nicht nur Pole, sondern auch Mischformen. Das bedeutet: Redaktionen, die eher wohlwollend gegenüber gendergerechter Sprache und ihrer konsequenten Sichtbarkeit in der öffentlichen Debatte sind, verzichten selbst in ihren Beiträgen auf diese Kennzeichnung in Form von Sonderzeichen und regeln die Frage pragmatisch.

Alle Mediengattungen sind in die Diskussionen involviert beziehungsweise reagieren in unterschiedlicher Weise. Im ZDF gibt es laut Information der Chefredakteurin Bettina Schausten (Huber 2023) weder ein Ge- noch ein Verbot für gendersensible Sprachverwendung. Der Sender wolle alle erreichen, weshalb in Angeboten für Jüngere gegendert werde, in anderen hingegen nicht. Den ARD-Anstalten bleibt der Umgang mit Sprache überlassen (ARD ohne Jahr). Der Hessische Rundfunk „bevorzugt geschlechtersensible Sprache im Unternehmen und im Programm, weil sie alle meint, alle zeigt und alle anspricht. Das unterstützt den Hessische Rundfunk als gemeinwohlorientierter Sender in seinem Programmauftrag" (Hessischer Rundfunk ohne Jahr).

Position Verwendung gendersensibler Sprache	**Begründung** für/gegen die Verwendung von Sonderzeichen	Mediales **Framing** in der Berichterstattung im Hinblick auf die hinter der Position liegende Wirkung	Journalistische, **redaktionsinterne Regelungen** und deren Hintergründe/Begründungen
1. *Gegen* die Verwendung gendersensibler Sprache	Überwiegend linguistische Argumentation: Generisches Maskulinum entspricht den gefestigten Sprachstrukturen; Unterscheidung biologisches und grammatikalisches Geschlecht = sexusneutral	Gendersensible Sprache irritiert, verbreitet Chaos, widerspricht den Empfehlungen des Rates für deutsche Rechtschreibung	Keine Sonderzeichen im Sinne von Ästhetik, Transparenz und Orientierung für die Leserschaft Abstufung: wenn möglich, neutrale Formulierungen
2. *Für* die Verwendung gendersensibler Sprachformen, vor allem Sonderzeichen	Sprache lebt und entwickelt sich; sozialpsychologische Argumentation der Festigung von Rollenstereotypen	Tradierte Sprache festigt Rollenstereotype; Sprache verändert gesellschaftliche Wirklichkeit und umgekehrt	Verwendung von Sonderzeichen Pragmatische Variante: Verwendung von maskulinen und femininen Pluralformen oder Partizipbildungen (wie Lesende)

Im Folgenden sollen die oben herausgearbeiteten Merkmale an drei überregionalen Tageszeitungen konkretisiert werden, die unterschiedlichen politischen Spektren angehören, aber nicht repräsentativ sind für die Gattung Printmedien: der in Berlin erscheinenden links-alternativen Tageszeitung (taz), der links-liberalen Süddeutschen Zeitung (SZ) aus München und der in Frankfurt produzierten konservativen Tageszeitung Frankfurter Allgemeine Zeitung (FAZ). Dargestellt werden die Positionen, wie sie von den Redaktionen selbst kommuniziert werden. Die jeweils dahinterstehenden Motive können möglicherweise aus der generellen Haltung gegenüber gesellschaftspolitischen Themen und Veränderungen erahnt, aber ohne eine ausführliche Analyse nicht belegt werden.

Die Tageszeitung (taz) präferiert seit ihrer Gründung als Genossenschaft im Jahr 1978 eine „progressive und inklusive Sprache"; in allen Artikeln stehen gendersensible Formulierungen mit Gendersternchen (zum Beispiel „Leser*innen") oder andere inklusive Sprachformen, um alle Geschlechter anzusprechen und Diskriminierung zu vermeiden. Dies solle die Vielfalt der Gesellschaft widerspiegeln und geschlechtergerechte Sprache fördern und gehöre zum Selbstverständnis des linksalternativen Mediums (vgl. taz 2023), das sich für soziale Gerechtigkeit und Gleichberechtigung einsetze.

Mittels Sprache sollen gesellschaftliche Missstände thematisiert und marginalisierte Gruppen sichtbar gemacht, mit gezielter Wortwahl und Themenauswahl soziale und politische Debatten sowie progressive Perspektiven angestoßen werden. Die seit Jahren anhaltenden Debatten sieht die Redaktion allerdings als kontraproduktiv, sie bringe „das Gendern in Verruf. Denn der Diskurs driftet ins Dogmatische ab und fördert so Verbote" (Schwab 2021).

Die taz-Redakteurin Noemi Molitor (2024) schreibt: „Unser Handwerk im Journalismus ist die Sprache. Bei genau diesem Werkzeug lohnt es sich also, genau hinzuschauen und auch ethische Fragen an orthografische Regeln zu stellen. Sei es in der Berichterstattung oder beim Schreiben im Allgemeinen." Dies äußert sie in dem Kontext, dass „Deutsch […] nie ungegenderte Sprache" war: „überall gegenderte Artikel, Pronomen und Wortendungen".

Für die Chefredakteurin der Süddeutschen Zeitung, Judith Wittwer (2021), ist gendersensible Sprache „keine Frage von Sonderzeichen", sondern etwas „Lebhaftes", das täglich um neue Wörter bereichert werde. Aufgrund der (sprachlichen) „Political Correctness und […] [der] sozialen Medien ist die Diskussion um eine diskriminierungsfreie Sprache für viele von den Rändern gar ins Zentrum gerückt". Zeitungen haben „eine Vorbildfunktion". Die Texte sollten „für möglichst viele (im Idealfall alle) Leserinnen und Leser verständlich" sein. Vor allem jugendliche Leser dürften in einer Zeitung nicht auf eine „Rechtschreibung stoßen, die ihnen in der Schule als Fehler angekreidet würde" (ebd.). Dementsprechend offen, aber zurückhaltend und pragmatisch verhält sich die Redaktion intern und nach außen hin in ihrer Berichterstattung.

Die Frankfurter Allgemeine Zeitung (FAZ) beruft sich auf ihren Gründungsauftrag, „für Freiheit einzustehen – für den einzelnen Menschen wie für unser Land" (FAZ ohne Jahr). Diese Freiheit des selbstbestimmten Denkens beginne im Kopf. „Die Frankfurter Allgemeine lässt Komplexität und Widersprüche zu, liefert Denkanstöße und gibt Raum für kontroverse Gedanken. Damit fordert sie ihre Leser zum Nachdenken heraus und unterstützt sie bei der selbstbestimmten Meinungsbildung, bei der Überprüfung ihrer Haltung und bei der Formulierung eigener Positionen" (ebd.). Gendersensiblen Sprachformen steht die Redaktion kritisch gegenüber und begründet dies mit sprachlicher Tradition, Lesbarkeit und Ästhetik. Aus zahlreichen Beiträgen ist herauszulesen, dass die FAZ generell eine kritische Position zum Gendern einnimmt, sie für ungerecht hält und die Politik auffordert, für die Einhaltung der Regeln zu sorgen (vgl. Schmoll 2024). Die FAZ gehört zu den Medien, die Gegnern der sprachlichen Veränderun-

gen regelmäßig eine Plattform bieten, sei es Linguisten wie Peter Eisenberg (2020 und 2021), der Autorin Elke Heidenreich oder dem Kulturstaatsminister Wolfram Weimer, der Gendersprache in seiner Behörde verboten hat. Die FAZ-Redakteurin Heike Schmoll (2025) holt verbal bisweilen kräftig aus, wenn sie sich zum Thema äußert, beispielsweise in einem Kommentar zum Urteil des Oberlandesgerichts Düsseldorf, das das Generische Maskulinum in Gerichtsurteilen für ausreichend erachtet, um gendergerecht zu kommunizieren.

5. KI und ihre Auswirkung auf die journalistische Sprache

In der Diskussion um gendersensible Sprache spielt inzwischen auch der Einsatz Künstlicher Intelligenz (KI) eine wachsende Rolle, da textgenerierende Systeme (wie unter anderem Chatbots) oder Übersetzungsprogramme die Wahl sprachlicher Mittel maßgeblich mitgestalten. So zeigen erste Analysen, wie maschinelle Übersetzungssysteme (wie *Google Translate*, *DeepL* etc.) mit genderfairer Sprache umgehen. Ergebnis: Die Systeme neigen stark dazu, maskuline Formen zu verwenden; neutrale beziehungsweise inklusive Alternativen sind verschwindend selten (0 – 2 Prozent) – selbst wenn Kontext vorliegt (Lardelli/Gromann 2023). Text-KIs reagieren auf Eingaben mit Gendersternchen unterschiedlich: Manche Modelle übernehmen diese konsequent, andere ignorieren sie und geben stattdessen generische Maskulina oder Doppelnennungen aus. Der Umgang der KI mit diesen Spezifika hängt stark von den Trainingsdaten und den Voreinstellungen der Entwickler ab. Gleichzeitig besteht das Risiko, dass stereotype Daten oder unzureichend reflektierte Trainingsansätze bestehende Ungleichheiten reproduzieren und durch automatisierte, massenhaft verbreitete Texte stereotype Muster noch verstärken. KI könnte eine produktive Funktion in der Versachlichung der Debatte übernehmen, indem sie flexible Optionen für verschiedene Sprachstile anbietet oder Nutzende transparent über Alternativen informiert. Ob KI die Diskussion um gendersensible Sprache also entspannt oder verschärft, hängt maßgeblich von der Transparenz, der Reflexivität und der Partizipation vielfältiger gesellschaftlicher Gruppen in der Entwicklung dieser Technologien ab. Eine bewusste Integration gendersensibler Sprachmuster in Trainings- und Feintuningprozesse könnte dazu beitragen, diskriminierungsarme und inklusivere Sprachformen zu fördern. Balestri (2025) entdeckt in einer aktuellen Studie, dass die neueste Version von *Google Gemini 2.0 Flash Experimental* tatsächlich seltener

Geschlechterverzerrung zuungunsten des weiblichen Geschlechts aufweist, insbesondere da bei spezifischen Eingaben die Akzeptanzraten im Vergleich zu den Ergebnissen von *ChatGPT- 4.0* deutlich gestiegen sind.

Der Einsatz von KI-Tools in den bisher zu findenden Positionspapieren beziehungsweise redaktionellen Leitlinien wird vor einem stark ethisch geframten Kontext definiert. Das Media Lab Bayern (2025), ein Innovation Hub für digitale Medien, hat ein ausführliches Dokument über die Möglichkeiten und Grenzen von KI im Journalismus erstellt. Neben Einordnungen, differenzierenden Erklärungen und Benennung konkreter Anwendungsfelder werden auch ethische Aspekte angerissen: Da Gesetze und Richtlinien auf internationaler, EU- und nationaler Ebene allein nicht ausreichten, müssten alle Medien-Unternehmen jeweils eigene Verhaltenskodizes oder Transparenzrichtlinien ausarbeiten. Eine der großen Herausforderungen für die Entwicklung und den Einsatz von KI im Journalismus seien unbewusste Vorurteile (*unconscious bias)* (vgl. ebd.). Aufgrund der Tatsache, dass die in Deutschland eingesetzten Tools überwiegend aus den USA stammen, werden die in KI eingebauten Stereotype regelmäßig reproduziert – das ist sehr sichtbar in KI-gestützten Personalauswahlprozessen, vor allem im Hinblick auf die Geschlechterdiskriminierung (vgl. Trzensimiech 2023).

Beispielhaft wird auf das Positionspapier des Deutschen Journalistenverbandes (DJV ohne Jahr) verwiesen, aus dem hier zitiert wird:

> „Eine Veröffentlichung von automatisiert bzw. mit KI erzeugten Beiträgen ohne die Beteiligung oder die effektive Eingriffsmöglichkeit von Journalistinnen und Journalisten darf nicht stattfinden. Redaktionen sollen klar geregelte Abnahme- und Freigabeprozesse für journalistische Inhalte etablieren. Ungeachtet der automatisierten Erstellung und/oder Distribution von journalistischen Inhalten durch Künstliche Intelligenz bleibt stets diejenige Redaktion für die Inhalte publizistisch verantwortlich…"

Der Bayerische Rundfunk (BR) hat im Jahr 2024 eine KI-Richtlinie erlassen, in der unter anderem steht:

> „Wir beschreiben KI-Technologie als technische Systeme und vermeiden vermenschlichende, irreführende Formulierungen. Insbesondere der Vergleich mit menschlicher Intelligenz und menschlichen Fähigkeiten wie Sprechen, Schreiben oder Denken stellt die Funktionen der Technologie oft als falsch und übermächtig dar. Dies betrifft auch den Gebrauch

von Metaphern und Bildsprache, die den irreführenden Eindruck künstlicher Lebewesen verstärken. Daher vermeiden wir diese Art der Bebilderung und Beschreibung in unseren Publikationen."

Rieke C. Harmsen, Chefredakteurin Online beim Sonntagsblatt, betont, dass die Redaktion KI für „den kreativen Prozess" nutzt, also für Ideen und Vorschläge zu Überschriften und ähnliches. Um Fakten zu belegen, werden ausschließlich seriöse Quellen erschlossen, also in der Regel persönliche Gespräche mit Expertinnen und Experten (vgl. Harmsen 2024).

6. Fazit: Ethikbasierte Sprache im Journalismus

Am Schluss werden die wesentlichen Aspekte der Debatte zusammengefasst:

- Das Konfliktfeld liegt zwischen (sprach-)wissenschaftlich, historisch begründeten Syntax-, Grammatik-, Lexik- und Orthographieregeln und der sprachlichen, tradierten Anwendungspraxis in der Gesellschaft oder: zwischen dem Genus (grammatisches Geschlecht – Generisches Maskulinum beim Plural) und Sexus (natürliches Geschlecht). In der Auseinandersetzung wird zudem die medizinisch-biologisch mehrheitlich vertretene Perspektive der Binarität zwischen den zwei Geschlechtern Mann und Frau in Frage gestellt.
- Gesellschaftspolitischer Konsens ist die Notwendigkeit der Gleichbehandlung und Gleichberechtigung aller Menschen – gleichgültig welchen Geschlechts.
- Unbestritten ist ebenfalls, dass Sprache ein zentrales Kulturgut ist, das gelebt wird und sich verändert.
- In jeder Sprachkultur bestehen Institutionen, die – wie der Rat für deutsche Rechtschreibung (RdR) – die Normen des Sprachgebrauchs beobachten und gegebenenfalls die Regeln anpassen. Dabei gilt immer die Korrelation zwischen gesellschaftlicher und individueller Freiheit einerseits, der Sprach- und Sprechfreiheit andererseits sowie die Grenzen der Verständlichkeit und der Erlernbarkeit von Sprache zu berücksichtigen.
- Politik positioniert sich in der öffentlichen und medialen Debatte, kann aber Verordnungen – auch im Sinne von Verboten – nur für die Verwendung der Sprache in staatlichen Einrichtungen (also Verwaltung und Schulen) erlassen.

Über das Thema Gendersprache und den Einsatz von Sonderzeichen wurde und wird nach wie vor sehr polarisiert diskutiert, insbesondere in den öffentlichen, politischen, aber auch wissenschaftlichen sowie medialen Arenen. Häufig ist Polemik das Stilmittel der Auseinandersetzung. Der Musiker und Publizist Fabian Payr (2022: VII) beobachtet „ein antiautoritäres Klima der orthografischen Freizügigkeit, das von einem unreflektierten Anpassen an Schreib- und Sprechmoden bis hin zum pflichtschuldigen Beachten von gendergerechten Schreibregeln reicht, die von einer akademischen Elite vorgegeben werden". Zé do Rock[2] (2021: 49) fragt: „Wird die Welt gerechter, wenn man die Sprache umbaut?" Er vergleicht die deutsche mit der englischen Sprache und beklagt, dass das Gendern deutsche Texte nur verlängere und man Gefahr laufe, als „Sexist, Chauvinist und Sonstist" zu gelten, wenn man nicht „jedes Wort, das irgendwie einen Menschen beschreibt, mit einem *innen" versehe (ebd.). Sein Artikel in der Wochenzeitung Die Zeit endet wie folgt:

> „Und wenn alles nach Quoten gehen soll: Über ein Jahrhundert lang regierten in Deutschland weiße Männer, seit gefühlten Jahrzehnten eine weiße Frau, als Nächstes muss a schwarze Transsexuelli Bundeskanzli werden. Der Fußballo Boateng ist nicht zufällig transsexuell? Möglich natürlich, dass ihm Talent und Erfahrung fürs Regieren fehlen, aber er hätte wenigstens die richtige Farbe, und wenn er kein Transsexuell ist, na ja, er könnte es sich ja noch überlegen."

Primär in Deutschland ist eine sehr politisch-inhaltlich überladene, polarisierte und polarisierende Sprach-Debatte zu beobachten, die eher einer vielfach polemischen Nicht-Diskussion vor allem zwischen Linguisten und Vertreterinnen der Genderforschung gleicht als der Suche nach einem Konsens. Es wird nicht zugehört und es werden keine Argumente ausgetauscht, sondern jeder beansprucht das Recht der richtigen Regel. Hinzu kommt, dass sich Politik – oder besser: einzelne Politiker – sehr populistisch einmischen. Die Medien schaukeln die Debatte hoch, indem sie der jeweiligen Position besonderen Platz einräumen (zum Beispiel in Form von Interviews oder Gastbeiträgen), sehr zugespitzte Positionen unkommentiert publizieren oder sie mit deutlichen Worten kommentieren. In den ersten Monaten des Jahres 2025 schien sich die Situation etwas beruhigt

2 Künstlername eines brasilianischen Schriftstellers, der seit den 1990er-Jahren in Deutschland lebt und sich in diversen Beiträgen mit der deutschen Sprache beschäftigt hat.

zu haben. Aber Gesetzesurteile wie die der Oberlandesgerichte Düsseldorf und Naumburg, dem Verfügen von Genderverboten in Behörden (wie zuletzt im Staatsministerium für Kunst und Kultur) befeuern die Auseinandersetzung der Kontrahenten auf der öffentlichen Bühne. Die Hoffnung bleibt, dass sich politische und publizistische Akteursgruppen auf die Einhaltung der in Grundgesetz Artikel 5, 1, 2 und 3 garantierten Rechte und Freiheiten besinnen. Jeder soll so sprechen dürfen, wie ihr oder ihm der Schnabel gewachsen ist. Verstehen und Verständnis hängen nicht primär und ausschließlich von Sternchen ab. Oder andersherum: Unsichtbarkeit, Missachtung und Diskriminierung von Menschen basieren auf einer antrainierten, mentalen Haltung und sind mit Gendersternchen und Co. nicht zu vertreiben. Die Lösung liegt in Bildung, die den Blick weitet, zu Offenheit und Toleranz führt.

Literatur

AfD (2021): Antrag gegen Verwendung geschlechtergerechter Sprache abgelehnt (online unter: https://www.bundestag.de/dokumente/textarchiv/2021/kw25-de-genderg erechte-sprache-846940 – letzter Zugriff: 11.7.2025).

Ammer, Jessica (Hg.) (2020): Die deutsche Sprache und ihre Geschlechter. Schriften der Stiftung Deutsche Sprache, Bd. 3, Berlin (online unter: https://www.stiftung-deu tsche-sprache.de/ddsuig.pdf – letzter Zugriff: 11.7.2025).

ARD: Gendern in der ARD (online unter: https://hilfe.ard.de/artikel/gendern-in-der -ard/ – letzter Zugriff: 16.9.2025).

Balestri, Roberto (2025): Gender and content bias in Large Language Models: a case study on Google Gemini 2.0 Flash Experimental, 18. März 2025 (online unter: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai .2025.1558696/full – letzter Zugriff: 16.9.2025).

Bargmann, Kai (2020): Was tun mit Lesenden, Forschenden, Kandidierenden? (online unter: https://www.better-media.de/bmwp/2020/02/was-tun-mit-lesenden-forschen den-kandidierenden/ – letzter Zugriff: 11.7.2025).

Bayerische Staatsregierung (2023): Damit Bayern stark und stabil bleibt – Regierungsprogramm der Zukunft, 5. Dezember 2023 (online unter: https://www.bayern.de/damit-bayern-stark-und-stabil-bleibt-regierungsprogramm-der-zukunft/ – letzter Zugriff: 11.7.2025).

Bayerischer Rundfunk (BR) (2024): KI-Richtlinien, 12. Juli 2024 (online unter: https://www.br.de/extra/ai-automation-lab/ki-ethik-100.html – letzter Zugriff: 11.7.2025).

Berger, Peter L. / Luckmann, Thomas (1969): Die gesellschaftliche Konstruktion der Wirklichkeit, Frankfurt am Main.

Braun, Sabine / Sczesny, Friedericke / Stahlberg, Dagmar (2005): Cognitive Effects of Masculine Generics in German: An Overview of Empirical Findings, in: Communications 30 (1/2005), S. 1–21. https://doi.org/10.1515/comm.2005.30.1.1

Bündnis 90/ Die Grünen (2015): Beschluss: Geschlechtergerechte Sprache in Anträgen an die BDK (online unter: Anträge-BDK-Sprache- Handreichung – letzter Zugriff: 17.9.2025).

Cassaris, Lovis Noah A. S. (2025): Die deutsche Sprache queeren: Zur geschlechtergerechten Sprachpraxis im Hochschulkontext, in: Queer Studies 45, Bielefeld. https://doi.org/10.14361/9783839476611

Dernbach, Beatrice (2019): Die Denkmuster in unseren Köpfen. Frames vereinfachen die mediale Diskussion über komplexe Themen, in: Beatrice Dernbach / Alexander Godulla / Annika Sehl (Hg.), Komplexität im Journalismus, Wiesbaden. https://doi.org/10.1007/978-3-658-22860-6_6

Dernbach, Beatrice (2024): Denkmuster in unseren Köpfen und die Freiheitsgrade des Sprechens, in: Netzwerk Wissenschaftsfreiheit (Hg.), Jahrbuch Wissenschaftsfreiheit. 1. Band, Berlin, S. 63–81.

Dernbach, Beatrice (2025): Wie die gesellschaftliche Transformation in die Mediensprache kommt. Eine Exkursion von der Energiewende bis zum Gendersternchen, in: Gabriele Mehling / Kristina Wied / Michael Wild (Hg.), Medien bewahren und verändern: Festschrift für Rudolf Stöber, Berlin, Boston, S. 509–520. https://doi.org/10.1515/9783111589732-033

Deutscher Journalistenverband (DJV): Positionspapier bezüglich des Einsatzes Künstlicher Intelligenz im Journalismus (online unter: https://www.djv.de/fileadmin/user_upload/DJV/INFORMATIONEN/medienpolitik/DJV-Positionspapier_KI_2023-04.pdf – letzter Zugriff: 11.7.2025).

Deutscher Presserat (2025): Pressekodex, 19. März 2025 (online unter https://www.presserat.de/pressekodex.html – letzter Zugriff: 11.7.2025).

Diewald, Gabriele / Steinhauer, Anja (2022): Handbuch geschlechtergerechte Sprache, 2. aktualisierte und erweiterte Aufl., Berlin.

Duden: Geschlechtergerechter Sprachgebrauch (online unter: https://www.duden.de/sprachwissen/sprachratgeber/Geschlechtergerechter-Sprachgebrauch – letzter Zugriff: 11.7.2025).

Eisenberg, Peter (2020): Das missbrauchte grammatische Geschlecht – Gendern im Wandel, in: Jessica Ammer (Hg.), Die deutsche Sprache und ihre Geschlechter. Schriften der Stiftung Deutsche Sprache, Bd. 3, Berlin, S. 17–23.

Eisenberg, Peter (2021): Unter dem Muff von hundert Jahren, 8. Januar 2021 (online unter: https://www.faz.net/aktuell/feuilleton/debatten/der-duden-und-der-unsinn-der-gegenderten-sprache-17135087.html – letzter Zugriff: 11.7.2025).

Entman, Robert (1993): Toward Clarification of a Fractured Paradigm, in: Journal of Communication 43 (4/1993), S. 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Erfurt, Jürgen (1996): Sprachwandel und Schriftlichkeit, in: Günther, Hartmut et al. (Hg.), Ein interdisziplinäres Handbuch internationaler Forschung, Berlin, New York, S. 1387–1404. https://doi.org/10.1515/9783110147445.2.9.1387

FAZ: Unsere Haltung. Freiheit beginnt im Kopf (online unter: https://www.frankfurter allgemeine.de/unser-selbstverstaendnis – letzter Zugriff: 11.7.2025).

Geisler, Sebastian / Vehlewald, Hans-Jörg (2022): Deutschlands oberster Sprachlehrer Wolf Schneider: „Gendern ist für Wichtigtuer", 02. August 2022 (online unter: https://www.bild.de/politik/inland/politik-inland/deutschlands-oberster-sprachlehrer-wolf-schneider-gendern-ist-fuer-wichtigtuer-80889590.bild.html – letzter Zugriff: 11.7.2025).

Gesellschaft für deutsche Sprache e. V. (GfdS) (2020): Leitlinien der GfdS zu den Möglichkeiten des Genderings, August 2020 (online unter: https://gfds.de/standpunkt-der-gfds-zu-einer-geschlechtergerechten-sprache/ – letzter Zugriff: 11.7.2025).[3]

Grüne München: Andere Gender, andere Sitten!, (online unter: https://www.gruene-muenchen.de/csd22/gendergerechte-sprache/ – letzter Zugriff: 17.9.2025).

Harmsen, Rieke C. (2024): KI-Richtlinien der Sonntagsblatt-Redaktion: So arbeiten wir mit ChatGPT & Co., 27. März 2024 (online unter: https://www.sonntagsblatt.de/artikel/medien/ki-richtlinien-der-sonntagsblatt-redaktion – letzter Zugriff: 11.7.2025).

Häusermann, Jürg (2011): Journalistisches Texten, Konstanz.

Heine, Matthias (2022): Der letzte Sprachpapst, 11. November 2022 (online unter: https://www.welt.de/kultur/medien/article242094047/Nachruf-auf-Wolf-Schneider-Der-letzte-Sprachpapst.html – letzter Zugriff: 16.9.2025).

Hessischer Rundfunk (HR): Geschlechtersensible Sprache im hr. Alle meinen, alle zeigen, alle ansprechen! (online unter: https://www.hr.de/unternehmen/backstage-und-meldungen/archiv-backstage-geschichten/geschlechtersensible-sprache-im-hr-v1,geschlechtersensible_sprache-100.html – letzter Zugriff: 16.9.2025).

Huber, Joachim (2023): Gendern im ZDF: „Wir wollen niemanden belehren oder erziehen". Interview mit ZDF-Chefredakteurin Bettina Schausten, 30. Januar 2023 (online unter: https://www.tagesspiegel.de/gesellschaft/zdf-chefredakteurin-bettina-schausten-im-interview-jan-bohmermann-notigt-mir-respekt-ab-9253296.html?icid=single-topic_9265968___ – letzter Zugriff: 16.9.2025).

Kurz, Josef et al. (Hg.) (2010): Stilistik für Journalisten, 2. Aufl., Wiesbaden.

Lakoff, George (2016): Auf leisen Sohlen ins Gehirn: politische Sprache und ihre heimliche Macht, Heidelberg.

Lakoff, George / Johnson, Mark (1980): Metaphors We Live By, Chicago, London.

Lardelli, Manuel / Gromann, Dagmar (2023): Gender-Fair (Machine) Translation. New Trends in Translation and Technology, S. 166–17, März 2023 (online unter: https://www.researchgate.net/publication/369948882_Gender-Fair_Machine_Translation#fullTextFileContent – letzter Zugriff: 16.9.2025).

Linden, Peter (2000): Wie Texte wirken, Berlin.

Müller-Spitzer, Carolin / Ochs, Samira (2024): Shifting social norms as a driving force for linguistic change: Struggles about language and gender in the German Bundestag (online unter: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/12493/file/Mueller_Spitzer_Shifting_social_norms_2024.pdf – letzter Zugriff: 11.7.2025).

Payr, Fabian (2022): Von Menschen und Mensch*innen, Heidelberg. https://doi.org/10.1007/978-3-658-36675-9

3 Der letzte Zugriff am 16.9.2025 scheiterte, da die GfdS nur noch über eine verschlüsselte Verbindung zugänglich ist (HTTP Strict Transport Security (HSTS)).

Piaget, Jean (1992): Das Erwachen der Intelligenz beim Kinde, München.

Pusch, Luise (1984): Das Deutsche als Männersprache, Frankfurt am Main.

Media Lab Bayern (2025): KI im Journalismus: Möglichkeiten & Grenzen, Media Trends, 23. Juli 2025 (online unter: https://www.media-lab.de/de/blog/journalismus-und-ki-chancen-grenzen-von-kuenstlicher-intelligenz-in-den-medien/#ethische-grundsaetze-beim-einsatz-von-ki-im-journalismus – letzter Zugriff: 11.7.2025).

Molitor, Noemi (2024): „Von wegen Gendersprache", Wider die „Grammatikarianer", 16. August 2024 (online unter: https://taz.de/Von-wegen-Gendersprache/!6027123/ – letzter Zugriff: 11.7.2025).

Rat für deutsche Rechtschreibung (RdR) (2021): Geschlechtergerechte Schreibung: Empfehlungen vom 26.3.2021, 26. März 2021 (online unter: https://www.rechtschreibrat.com/geschlechtergerechte-schreibung-empfehlungen-vom-26-03-2021/ – letzter Zugriff: 11.7.2025).

Rat für deutsche Rechtschreibung (RdR) (2023): Geschlechtergerechte Schreibung: Erläuterungen, Begründung und Kriterien vom 15.12.2023, 15. Dezember 2023 (online unter: https://www.rechtschreibrat.com/geschlechtergerechte-schreibung-erlaeuterungen-begruendung-und-kriterien-vom-15-12-2023/ – letzter Zugriff: 11.7.2025).

Reus, Gunter (o.D.): Sprache und Stil (online unter: https://journalistikon.de/category/sprache-und-stil-des-journalismus/ – letzter Zugriff: 11.7.2025).

Rock, Zé do (2021): Von innen, unnen und onnen, in: Die Zeit Nr. 32/2021, 06. August 2021 (online unter: https://www.zeit.de/2021/32/geschlechtergerechte-sprache-diskriminierung-gendersternchen – letzter Zugriff: 11.7.2025).

Schmoll, Heike (2025): Der Unsinn der Gendersprache, 31. Juli 2025 (online unter: https://www.faz.net/aktuell/politik/inland/urteil-in-duesseldorf-der-unsinn-der-gendersprache-110614811.html – letzter Zugriff: 16.9.2025).

Schmoll, Heike (2024): Zum Gendern durch die Hintertür, 23. Januar 2024 (online unter: https://www.faz.net/aktuell/politik/inland/gendern-politik-muss-fuer-einhaltung-der-rechtschreibregeln-sorgen-19466572.html – letzter Zugriff: 11.7.2025).

Schwab, Waltraud (2021): Debatte übers Gendern, In der Sackgasse, 12. September 2021 (online unter: https://taz.de/Debatte-uebers-Gendern/!5797123/ – letzter Zugriff: 11.7.2025).

Shell (2024): Einstellung Jugendlicher zum Thema Gendern nach Geschlecht in Deutschland im Jahr 2024 (online unter: https://de-statista-com.thn.idm.oclc.org/statistik/daten/studie/1537944/umfrage/gendern-einstellungen-jugendlicher/ – letzter Zugriff: 11.7.2025).

Stahlberg, Dagmar / Braun, Sabine / Sczesny, Friedericke (2001): Name Your Favorite Musician: Effects of Masculine Generics and of their Alternatives in German, in: Journal of Language and Social Psychology, 20 (4/2001), S. 464–469. https://journals.sagepub.com/doi/10.1177/0261927X01020004004

Stainer-Hämmerle, Kathrin / Ingruber, Daniela / Marschnig, Georg (Hg.): Verschwörungserzählungen und Faktenorientierung in der Politischen Bildung, Frankfurt am Main.

taz (2023): Redaktionsstatut, 04. September 2023 (online unter: https://taz.de/taz-die-tageszeitung/!vn5957750/ – letzter Zugriff: 11.7.2025).

Trzensimiech, Annika Christina (2023): Umfang und Reduktionsmöglichkeiten der Geschlechtsdiskriminierung in KI-gestützten Auswahlprozessen, in: Studentisches Magazin der HAW Hamburg, Bd. 4 Nr. 2. https://doi.org/10.15460/apimagazin.2023.4.2.153

Trömel-Plötz, Senta (1982): Frauensprache: Sprache der Veränderung, Frankfurt am Main.

Waldendorf, Anica (2023). Words of change: The increase of gender-inclusive language in German media, in: European Sociological Review, 40 (2/2023), S. 357–374. https://doi.org/10.1093/esr/jcad044

Wehling, Elisabeth (2016): Politisches Framing, Köln.

Wittwer, Judith (2021): Transparenz-Blog: Warum verzichtet die SZ auf das Sternchen?, in: Süddeutsche Zeitung, 27. Juli 2021 (online unter: https://www.sueddeutsche.de/kolumne/transparenz-blog-warum-verzichtet-die-sz-auf-das-sternchen-1.5364973 – letzter Zugriff: 11.7.2025).

Werner, Hendrick (2019): 20 Jahre Rechtschreibreform in den deutschen Printmedien, 01. August 2019 (online unter: https://www.weser-kurier.de/bremen/kultur/20-jahre-rechtschreibreform-in-den-deutschen-printmedien-doc7e4dcr9s7qpm1lpr3lw – letzter Zugriff: 11.7.2025).

YouGov (2023): Finden Sie geschlechtergerechte Sprache, sogenanntes Gendern, wichtig oder unwichtig?, 07. März 2023 (online unter: https://de-statista-com.thn.idm.oclc.org/statistik/daten/studie/1536019/umfrage/gendern-relevanz-nach-parteipraeferenz/ – letzter Zugriff: 11.7.2025).

Zylka, Jenni / Grimberg, Steffen (2021): Wie deutsche Medien mit der Genderfrage umgehen, 06. Juli 2021 (online unter: https://www.mdr.de/medien360g/medienwissen/wie-medien-gendern-100.html – letzter Zugriff: 11.7.2025).

II.
Automatisierte Sprache: Potenziale und Risiken Künstlicher Intelligenz

„Ich kenne mich mit KI-Technologie nicht aus, aber...": Eine wissenssoziologische Diskursanalyse von User:innendiskussionen zu Künstlicher Intelligenz auf Nachrichtenwebsite und *Instagram*-Account von *Die Zeit*

Brigitte Huber und Julia Levasier

Zusammenfassung

Künstliche Intelligenz (KI) ist in den letzten Jahren zu einem kontroversen und vieldiskutierten Thema geworden. Während sich bereits zahlreiche Studien mit traditionellen Mediendiskursen zu KI befasst haben, besteht noch Forschungsbedarf im Bereich von User:innendiskursen, insbesondere mit Blick auf das Thema generative KI. Der vorliegende Beitrag untersucht User:innendiskussionen über generative KI in den Kommentarbereichen der Nachrichtenwebsite und des *Instagram*-Accounts der deutschen Wochenzeitung *Die Zeit* (n = 427). Mit Hilfe einer wissenssoziologischen Diskursanalyse wird dabei konkret erforscht, welche Diskursstränge, Themen, Argumente und Wissen im Diskurs sichtbar werden. Die Ergebnisse zeigen, dass User:innendiskurse zu generativer KI sehr vielschichtig verlaufen. Dabei bilden Fragestellungen rund um verantwortungsvollen und ethischen Umgang mit KI einen Schwerpunkt. KI-bedingte Herausforderungen werden von User:innen als Gefahren in der Größenordnung von Naturkatastrophen eingeordnet. Die Glaubwürdigkeit des als erfahrungsbasiert herausgestellten Wissens einiger User:innen wird von anderen Mitdiskutierenden kritisch hinterfragt. Auf der Online-Nachrichtenwebsite zeigen sich längere Diskussionsbeiträge als auf *Instagram* und es ist intensivere Diskussion erkennbar. Entsprechend stellt sich die Frage, wie eine zunehmende Verlagerung von User:innendiskussionen von Nachrichtenwebsites auf Social-Media-Plattformern im Hinblick auf die Qualität digitaler Debatten einzuordnen ist.

1. Einleitung

In den letzten Jahren ist KI zu einem vorherrschenden Thema der öffentlichen Debatte geworden.[1] Studien zeigen einen deutlichen Anstieg der Medienberichterstattung über KI im Laufe der Zeit (vgl. Korneeva et al. 2023; Nguyen/Hekman 2022; Vergeer 2020). Zuletzt hat das Aufkommen von *ChatGPT* zu intensiven öffentlichen Debatten über wirtschaftliche und gesellschaftliche Auswirkungen sowie zu ethischen Überlegungen mit Blick auf generative KI geführt (vgl. Qi et al. 2024). Während traditionelle Mediendiskurse über KI bereits umfassend untersucht wurden (siehe etwa Brantner/Saurwein 2021; Brause et al. 2023; Cools et al. 2024; Duberry/Hamidi 2021; Roe/Perkins 2023; Shaikh/Moran 2024; Taddicken et al. 2020; Winkel 2024), sind Online-Diskussionen über KI noch relativ wenig erforscht, insbesondere auf Social Media stattfindende Diskussionen (vgl. Tsimpoukis 2025; Zeng et al. 2020). Hier schließt der vorliegende Beitrag an und analysiert User:innenkommentare zum Thema generative KI nicht nur unter Nachrichtenartikeln, sondern auch auf Social Media. User:innenkommentare sind deshalb besonders relevant für das Verständnis des öffentlichen Diskurses über neue Technologien wie KI, weil sie zeigen, auf welche Art und Weise Laien komplexe technologische Entwicklungen jenseits der professionellen Medienberichterstattung interpretieren, bewerten und diskutieren.

Das wissenschaftliche Interesse an User:innendiskursen verlagert sich durch die Abwanderung vieler User:innen aus Nachrichtenforen hin zu Social-Media-Plattformen entsprechend ebenfalls zunehmend dorthin, wobei hier noch viel Forschungsbedarf besteht (für einen Überblick zu den analysierten Plattformen, siehe Kubin et al. 2024). Einzelne Studien haben bereits KI-bezogene Diskurse auf Plattformen wie *Twitter* (jetzt X) (vgl. Miyazaki et al. 2024; Zou et al. 2025) oder *Reddit* (vgl. Qi et al. 2024) untersucht. Es besteht Bedarf, solche Diskussionen auch auf anderen Plattformen zu untersuchen, insbesondere auf *Instagram* (vgl. Anter/Kümpel 2023; San Cornelio 2022) – einer Plattform, die zunehmend für Nachrichtenkonsum genutzt wird (vgl. Newman et al. 2024). Daher wurde für den vorliegenden Beitrag eine Diskursanalyse (vgl. Keller 2011) der User:innendiskussionen über KI durchgeführt. Dabei wurden nicht nur die Kommentare auf der

1 Die in diesem Beitrag präsentierten Ergebnisse werden auch in ausführlicherer Form in einer englischsprachigen Fachzeitschrift veröffentlicht (derzeit im Review-Verfahren).

klassischen Nachrichtenwebsite der deutschen Wochenzeitung *Die Zeit*, sondern auch Kommentare auf dem *Instagram*-Account des Mediums berücksichtigt. Dies ermöglicht es, ein umfassenderes Bild des Diskurses zu zeichnen.

2. Forschungsperspektiven auf Nutzer:innenkommentare

Die Forschung zu Nutzer:innenkommentaren hat sich aus verschiedenen disziplinären Perspektiven entwickelt (für einen Überblick siehe Ziegele 2019). Die Journalismusforschung untersucht Nutzer:innenkommentare als Feedback-Mechanismen, die den Newsroom-Betrieb und die journalistische Praxis beeinflussen (vgl. Domingo 2008; Singer 2010; Ziegele/Jost 2016). Die politische Kommunikationsforschung betrachtet User:innenkommentare aus verschiedenen theoretischen Perspektiven, so etwa als Instrumente des politischen Engagements und als Möglichkeit der deliberativen Partizipation (vgl. Friess/Eilders 2015; Malinen 2015; Ruiz et al. 2011) sowie als Orte für einen gegenöffentlichen Diskurs, der Mainstream-Öffentlichkeiten hinterfragt (vgl. Toepfl/Piwoni 2018). Aus sprachwissenschaftlicher und kultureller Perspektive können Kommentare als diskursive Ausdrucksformen oppositioneller Lesarten fungieren sowie als Mittel der Identitätskonstruktion durch kreative Sprachcodes (vgl. Hall 1980).

Die digitale Transformation bringt einen grundlegenden Wandel mit sich, wie Nachrichten rezipiert und konsumiert werden (vgl. Newman et al. 2024). Zum einen hat eine Verlagerung weg vom passiven Konsum hin zu aktiver Beteiligung und Mitgestaltung stattgefunden (vgl. Baqir et al. 2025; Picone 2017). So erleichtern Social-Media-Seiten, Nachrichten-Websites, mobile Nachrichten-Apps, Video-Streaming-Dienste und Nachrichtenaggregationsportale die Auseinandersetzung mit Medieninhalten (vgl. Geers 2020). Die digitale Partizipation ermöglicht vielschichtige Interaktionsmuster mit Inhalten, die je nach Plattform variieren können (vgl. Brown et al. 2018; Gaudette et al. 2021) Zum anderen können informelle Lernprozesse in selbstorganisierten Online-Umgebungen stattfinden (vgl. Del Valle et al. 2020). Nutzer:innenkommentare sind im Vergleich zu traditionellen Mediennarrativen oft vielfältiger (vgl. Lörcher/Taddicken 2017), beinhalten aber auch dysfunktionale Kommunikationsformen wie *trolling* oder Hassrede (vgl. Brubaker et al. 2021; Eberwein 2020; Quandt 2018). Plattformen ermöglichen zwar den kollaborativen Aufbau von Wissen, begünstigen aber auch „dark participation" (vgl. Quandt 2018) und toxische Kommunikation

(vgl. Kim et al. 2021). Dies kann der Akzeptanz neuer Technologien und damit auch von KI zuwiderlaufen (vgl. Rega et al. 2023). Zwar können User:innenkommentare das Verstehen von Nachrichteninhalten potenziell verbessern, indem sie etwa vielfältigere Perspektiven und Einordnung bieten sowie journalistischen Content mit eigenen Erfahrungen bereichern. Mit Kommentaren gehen jedoch auch Risiken einher, wenn User:innen etwa sachlich falsche Inhalte in ihren Kommentaren posten und die Nachrichteninhalte, auf die sie sich beziehen, nicht korrekt bewerten (vgl. Hsueh et al. 2015) oder wenn Inhalte voreingenommen oder unzivilisiert sind (vgl. Ziegele 2025). Empirische Untersuchungen zeigen, dass 20 bis 40 Prozent der Nutzer:innenkommentare in Nachrichtenkommentarbereichen inzivil sind (vgl. Coe et al. 2015; Gonçalves et al. 2020; Rossini 2022; Santana 2013; Su et al. 2018), was sich auch negativ auf die wahrgenommene Qualität und Objektivität journalistischer Inhalte auswirken kann (vgl. Schlesinger 2024). Da traditionelle Gatekeeping-Mechanismen durch algorithmische Logik und Plattform-Governance ersetzt werden (vgl. Gorwa et al. 2020), haben sich auch die Bedingungen und Spielregeln verändert, unter denen unterschiedliche Nutzer:innengruppen an öffentlichen Diskursen über komplexe Zukunftsthemen wie KI teilnehmen können.

2.1 Theoretische Verortung: Kommentarspalten und Öffentlichkeitskonzepte

Durch das Posten von Kommentaren und Meinungen auf Nachrichtenwebsites und Social-Media-Plattformen hat sich die einst passive Leserschaft zu aktiven Laien-Diskursteilnehmer:innen gewandelt (vgl. Ziegele 2019). Nutzer:innenkommentare können die persönliche Meinungsbildung zu bestimmten Themen erheblich beeinflussen (vgl. Anderson et al. 2014; Hsueh et al. 2015) und darüber hinaus zu einer veränderten Einschätzung des öffentlichen Meinungsklimas beitragen (vgl. Lee/Jang 2010; von Sikorski/Hänelt 2016). Der Einfluss von Nutzer:innenkommentaren geht somit über die individuelle Meinungsäußerung hinaus und prägt vielmehr breitere Meinungsbildungsprozesse.

Die sozialwissenschaftliche Literatur konzipiert Öffentlichkeit als Zusammenschluss mehrerer Foren, die sich um verschiedene soziale Gruppen, Themen und Kontexte formieren und unterschiedliche Arenen für öffentliches Engagement und Deliberation bieten (vgl. Dahlgren 2005). Dieses vielschichtige Verständnis geht über traditionelle Öffentlichkeitsmo-

delle hinaus, die zwischen der massenmedialen Öffentlichkeit und kleineren, thematisch ausgerichteten, oft fachspezifischen Öffentlichkeiten sowie einfachen Formen der Öffentlichkeit wie persönlichen Begegnungen unterschieden haben (vgl. Gerhards/Neidhardt 1990). Mit Blick auf die Entwicklung digitaler Öffentlichkeiten wird von der sogenannten „Plattformisierung" (vgl. Fischer/Jarren 2024) der Öffentlichkeit gesprochen. Digitale Plattformen fungieren als neue Vermittler, die die öffentliche Kommunikation durch algorithmische Kuratierung und Content-Moderation auch deshalb stark verändern, weil traditionelle journalistische Normen allenfalls eine untergeordnete Rolle spielen (vgl. Gorwa et al. 2020). Plattformen verbinden nicht einfach nur Online- und Offline-Sphären, sondern ermöglichen es individuellen und institutionellen Kommunikator:innen, traditionelle Medien zu umgehen, während sie gleichzeitig die öffentliche Kommunikation mithilfe weitgehend undurchsichtiger Empfehlungssysteme neu strukturieren (vgl. Eisenegger/Schäfer 2023). Fischer und Jarren (2024) argumentieren, dass digitale Strukturen, darunter Elemente wie algorithmische Filter und die Datafizierung, die Kommunikation in sozialen Medien erheblich beeinflussen. Plattformen sind relevant für die Bildung von Gruppen-, Organisations- oder Netzwerk-Öffentlichkeiten, bieten eine Bühne für Themen und Meinungen und ermöglichen es Einzelpersonen und Gruppen, sich zu vernetzen, während sie gleichzeitig journalistische Akteur:innen und damit die traditionellen Nachrichtenmedien beeinflussen. Aufgrund der disruptiven Auswirkungen der Plattformisierung auf den demokratischen Diskurs und der ausgeprägten Instabilität der aktuellen Kommunikationsökosysteme erhält das Plädoyer für die Konzeption einer „post-öffentlichen Sphäre" (Schlesinger 2024) neuen Auftrieb.

Angesichts der hier nur skizzierten grundlegenden Veränderungen der Struktur digitaler Öffentlichkeiten fordern Pfetsch et al. (2018) eine Verknüpfung der Konzepte der antagonistischen und partizipativen Ansätze der Öffentlichkeitssoziologie mit dem theoretischen Ansatz der Netzwerkkommunikation (vgl. Chadwick 2017; Klinger 2018), um unterschiedliche Prioritäten und Perspektiven in digitalen und Offline-Öffentlichkeiten empirisch analysieren zu können. Im vorliegenden Beitrag werden Kommentarbereiche in Online-Nachrichtenmedien als „mass-media induced discussion arena" (Lörcher/Taddicken 2017) konzeptualisiert.

2.2 Diskurse über KI in traditionellen Medien und sozialen Medien

Die Medienberichterstattung über KI hat etwa seit dem Jahr 2014 deutlich zugenommen (vgl. Korneeva et al. 2023; Nguyen/Hekman 2022; Vergeer 2020). Untersuchungen zeigen, dass die Berichterstattung über KI in den USA im Vergleich zur Berichterstattung in Europa positiver ist (vgl. Duberry/Hamidi 2021). Mit Blick auf den deutschen Mediendiskurs zeigen Untersuchungen, dass neben einer relativ ausgewogenen, positiven Berichterstattung in den großen deutschen Medien wirtschaftliche Aspekte von KI im Vordergrund stehen (vgl. Fischer/Puschmann 2021; Kieslich et al. 2022; Taddicken et al. 2020). Dieses Berichterstattungsmuster spiegelt die allgemeine Tendenz im Technikjournalismus wider, wirtschaftliche Auswirkungen gegenüber sozialen Aspekten zu priorisieren. Rusche et al. (2022) und Cools et al. (2024) bestätigen diese prioritäre Fokussierung auf wirtschaftliche Aspekte von KI, wobei bestimmte Stränge der Berichterstattung durchaus auch ethische und gesellschaftliche Rahmenbedingungen thematisieren. Nach der Einführung von *ChatGPT* beobachten Roe und Perkins (2023) eine Verschiebung in den britischen Medien hin zur Thematisierung von KI-bedingten Risiken. Für Deutschland stellt Wittemann (2024) seit der Einführung von *ChatGPT* eine deutliche Abkehr von einer zuvor eher ausgewogenen Berichterstattung hin zu einer eher überwiegend kritischen, die Risiken und Gefahren von KI betonenden Berichterstattung fest.

Kimmerle et al. (2015) zeigen, dass Social-Media-Plattformen kollektive Sinnbildungsprozesse ermöglichen und verdeutlichen, dass User:innen aktiv an der kollektiven Wissenskonstruktion beteiligt sind. Daraus lässt sich ableiten, dass Social-Media-Plattformen über die traditionelle Medienberichterstattung hinaus eine besonders einflussreiche Rolle bei der Förderung der öffentlichen Wahrnehmung und Einstellung auch gegenüber neuen Technologien spielen können (vgl. Zou et al. 2025). Hara et al. (2025) zeigen, dass Laien auf der Plattform X aktiv zu wertorientierten Diskussionen über generative KI beitragen und zu Mitgestalter:innen von Wissensbeständen werden können. In diesem Sinne machen auf den Plattformen geteilte Erfahrungen von Laien die Technologie aus zweiter Hand erfahrbar und senken potenziell die Hemmschwelle für User:innen, selbst mit ihr zu experimentieren (vgl. Rogers 2003; Zolkepli/Kamarulzaman 2015). Qi et al. (2024) untersuchten die öffentliche Wahrnehmung von KI anhand einer Analyse von 33.912 *Reddit*-Kommentaren in 388 s*ubreddits* von November 2022 bis Juni 2023 unter Verwendung von Themen- und Sentiment-Analysen. Die Ergebnisse zeigen, dass sich die Diskussionen in

technologieorientierten s*ubreddits* eher auf die technischen Aspekte der KI konzentrieren, während nicht-technologische *subreddits* auch stärker gesellschaftliche Auswirkungen wie zum Beispiel die Gefährdung der Arbeitsplatzsicherheit durch KI thematisieren. Es zeigt sich auch, dass technologieorientierte *communities* eine stärkere Polarisierung der Stimmung aufweisen, was sowohl auf Optimismus hinsichtlich des technologischen Fortschritts als auch auf erhebliche Bedenken hinsichtlich der gesellschaftlichen Auswirkungen der KI hindeutet. Miyazaki et al. (2024) stellen auf Basis der Analyse von über 3 Millionen *twee*ts aus den Jahren 2019 bis 2023 fest, dass *Twitter*-Nutzer:innen, die bereits intensivere Erfahrung mit KI aufwiesen, wie zum Beispiel IT-Fachleute, tendenziell eine positivere Einstellung zeigen.

Aufbauend auf diese Befunde, interessiert sich die vorliegende Studie dafür, wie User:innen das Thema generative KI aushandeln. Die vorhandene Literatur zeigt, dass Diskurse über verantwortungsvolle KI und Ethik überwiegend negativ geprägt sind (vgl. Hagendorff 2024). Ziel der vorliegenden Studie ist es zu untersuchen, wie Nutzer:innendiskurse konstruiert werden und welche Wissensbestände in diesen Diskussionen herangezogen werden, um KI zu legitimieren oder zu delegitimieren.

3. Forschungsfragen

Die vorliegende Studie setzt sich zum Ziel, Themen, Argumente und Wissen in User:innendiskussionen zu generativer KI zu erforschen. Angesichts der polarisierten Natur digitaler Debatten, die bei anderen kontroversen Themen beobachtet wurden (vgl. Graham et al. 2021) und der Erkenntnis, dass technologieorientierte *online-communities* eine stärkere Polarisierung der Meinungen zu KI zeigen (vgl. Qi et al., 2024), war unsere Ausgangserwartung, unterschiedliche Muster in der Art und Weise zu identifizieren, wie Nutzer:innen sich mit Themen rund um generative KI auseinandersetzen. Angesichts der dokumentierten Verschiebung hin zu einer kritischeren Berichterstattung über KI in deutschen Medien nach *ChatGPT* (vgl. Wittemann 2024) sind wir außerdem davon ausgegangen, dass die Diskussionen der User:innen diese professionellen Medienrahmen widerspiegeln, hinterfragen oder erweitern könnten. Der Studie liegen drei Forschungsfragen zu Grunde, die auf die Identifizierung struktureller Muster im User:innendiskurs sowie der Untersuchung inhaltlicher und epistemologischer Dimen-

sionen abzielen. Konkret befasst sich die erste Forschungsfrage mit der Organisation des Diskurses:

> FF1: Welche Diskursstränge lassen sich in den User:innendiskussionen zum Thema generative KI auf der Nachrichtenwebsite und dem *Instagram*-Account von *Die Zeit* identifizieren?

Die zweite Forschungsfrage zielt darauf ab, den inhaltlichen Gehalt der Nutzer:innendiskussionen zu untersuchen und zu erforschen, ob sich der beobachtete Fokus auf wirtschaftliche Themen in der traditionellen deutschen KI-Berichterstattung (vgl. Fischer/Puschmann 2021; Taddicken et al. 2020) auch im Nutzer:innendiskurs widerspiegelt oder ob Laien andere thematische Schwerpunkte setzen. Die zweite Forschungsfrage lautet daher:

> FF2: Welche Themen und Argumente werden in den User:innendiskussionen zu generativer KI auf der Nachrichtenwebsite und dem *Instagram* Account von *Die Zeit* sichtbar?

Schließlich soll erkundet werden, welches Wissen User:innen in den Diskurs einbringen. Angesichts von Forschungsergebnissen, die darauf hindeuten, dass die breitere Öffentlichkeit in Diskussionen zu generativer KI neben Expert:innen zu Co-Produzent:innen von Wissensbeständen über neue Technologien wird (vgl. Hara et al. 2025), ist ein Ziel zu untersuchen, welche Formen von Wissen hier konkret eingebracht werden und wie der Anspruch an die Glaubwürdigkeit und Wahrhaftigkeit an diese Wissensbestände verhandelt wird. Daher lautet die dritte Forschungsfrage:

> FF3: Welches Wissen bringen Diskursteilnehmer:innen in den User:innendiskussionen zum Thema generative KI auf der Nachrichtenwebsite und dem *Instagram*-Account von *Die Zeit* ein?

4. Methode

Um die Forschungsfragen zu beantworten, wurde eine wissenssoziologische Diskursanalyse (vgl. Keller 2011) durchgeführt. Dieser Ansatz zielt darauf ab, zu untersuchen, wie symbolische Ordnungen diskursiv konstruiert werden (vgl. Keller 2011). Eine Form der Diskursanalyse ist speziell auf die Untersuchung von Wissensprozessen zugeschnitten. Dieser Ansatz eignet sich damit besonders für unseren Zweck, da er einen vorwiegend soziologischen Ansatz verfolgt – und keinen rein linguistischen – und somit die

diskursive Konstruktion von Realität als empirischen Prozess erforschbar macht (vgl. Keller 2005).

Die vorliegende Studie umfasst die Analyse von Nutzer:innenkommentaren auf dem *Instagram*-Account und der Nachrichtenwebsite von *Die Zeit*. Mit etwa 1,3 Millionen Leser:innen nimmt *Die Zeit* eine herausragende Stellung in der deutschen Medienlandschaft ein (vgl. ZEIT Advise, 2026, S. 5). Darüber hinaus bietet die Instagram-Präsenz von *Die Zeit* mit über 1,4 Millionen Followern (Stand: Februar 2026) gute Möglichkeiten, KI-Diskurse auf verschiedenen digitalen Plattformen innerhalb einer Medienmarke zu erfassen. Um relevantes Material zu identifizieren, wurden zunächst alle Beiträge durchgeschaut, die seit dem Start von *ChatGPT* am 30. November 2022 auf dem *Instagram*-Account von *Die Zeit* zum Thema KI veröffentlicht wurden, das heißt alle Posts, die zwischen dem 30.11.2022 und dem 30.6.2024 gepostet wurden. Beiträge, die sich mit generativer KI befassten, wurden in eine Excel-Liste gemeinsam mit der Anzahl der entsprechenden Nutzer:innenkommentare eingetragen. Diese erste Sichtung ergab 20 relevante Beiträge und Reels mit insgesamt 2.579 Nutzer:innenkommentaren. Anschließend wurden diese *Instagram*-Beiträge mit Online-Artikeln auf der Nachrichtenwebsite von *Die Zeit* abgeglichen, das heißt es wurde überprüft, auf welchen konkreten Online-Nachrichtenartikel sich die einzelnen relevanten *Instagram*-Beiträge bezogen. Dieser Schritt führte zur Identifizierung von 16 relevanten Online-Zeitungsartikeln, die auf der Nachrichtenwebsite von *Die Zeit* veröffentlicht wurden, mit insgesamt 1.323 Nutzer:innenkommentaren. Anschließend wurde die Anzahl der Nutzer:nnenkommentare pro *Instagram*-Beitrag und entsprechendem Online-Nachrichtenartikel zusammengerechnet. Aus dieser Aggregation wurden die sechs Beiträge mit den meisten Nutzer:innenkommentaren für eine detaillierte Untersuchung ausgewählt. Diese Entscheidung, verschiedene Artikel als Ausgangspunkt zu nehmen, wurde aufbauend auf Huber et al. (2019) getroffen, um möglichst unterschiedliche Themen und Argumente erfassen zu können. Um die Analyse durchführbar zu machen, haben wir eine maximale Anzahl von Kommentaren pro Artikel festgelegt und höchstens 40 Kommentare pro Artikel und Plattform in das Sample aufgenommen. Bei zwei Artikeln wurden weniger als 40 Kommentare analysiert, da einzelne Kommentare nicht mehr verfügbar waren. Diese Vorgehensweise ergab einen finalen Korpus von 427 Nutzer:innenkommentaren. Eine Übersicht über die ausgewählten Artikel und die Anzahl der Kommentare pro Artikel findet sich in *Tabelle A1* im Anhang. Die ausgewählten Nutzer:innenkommentare wurden von einer Person kodiert. In einem ersten Schritt wurden

anhand bestehender wissenssoziologischer Diskursanalysen (vgl. Huber et al. 2019; Zimmermann 2013) deduktiv allgemeine Hauptkategorien gebildet, nämlich Akteur:innen, Themen, Argumente (Pro und Contra) und Wissenstypen. Weitere Kategorien wurden dann induktiv aus dem Material gebildet. Die User:innenkommentare wurden mit Hilfe der qualitativen Datenanalyse-Software MAXQDA kodiert.

5. Ergebnisse

Das Ziel der Studie war es, Diskursstränge, Themen, Argumente und Wissen zu identifizieren, die in den Nutzer:innendiskussionen zu generativer KI sichtbar werden. Zunächst war von Interesse, welche unterschiedlichen Diskursstränge in den Nutzer:innendiskussionen zu generativer KI auf der Nachrichtenwebsite und dem *Instagram*-Account der deutschen Wochenzeitung *Die Zeit* identifiziert werden können (FF1). Die Ergebnisse der Diskursanalyse ergaben fünf Hauptdiskursstränge:

(1) generative KI und menschliche Eigenschaften/Bewusstsein,
(2) generative KI und Trauerbegleitung,
(3) generative KI und Ausbeutung,
(4) generative KI und Datenschutz sowie
(5) generative KI und Übersetzung/Synchronisation von Videos.

In einem nächsten Schritt wurden die Themen (FF2) erfasst, die innerhalb dieser Stränge im Zusammenhang mit generativer KI auftauchten. Eine genauere Betrachtung der Kommentare innerhalb der identifizierten Stränge zeigt ein starkes Interesse der Nutzer:innen an der Diskussion der Terminologie und der historischen Entwicklung von KI – dies kam häufig im ersten Diskursstrang vor, der sich mit generativer KI und menschlichen Eigenschaften befasste, sowie auch im dritten Diskursstrang, der sich mit Ausbeutung befasste. Darüber hinaus diskutierten User:innen, ob generative KI über ein Bewusstsein verfügt und ob die Unterscheidung zwischen Mensch und Maschine im Zusammenhang mit generativer KI noch gültig ist. Im zweiten Diskursstrang war ein besonders hervorstechendes Diskussionsthema der Trend in den USA, verstorbener Personen durch den Einsatz von KI wieder in Form von Textantworten oder Sprachnachrichten zum Leben zu erwecken. Während die Mehrheit der User:innen dies abstoßend fand und den psychologischen Wert der Kommunikation mit KI-Darstellungen verstorbener Verwandter als fragwürdig erachtete,

sahen einige darin das Potenzial für eine wirksamere Trauerbewältigung. Auch Bedenken hinsichtlich des Datenschutzes und der Ausbeutung von *clickworkern* im Zusammenhang mit KI-Trainingsdaten wurden geäußert. Schließlich wurde eine allgemeinere Diskussion über Ethik im Zusammenhang mit KI sowie über Arbeitsplatzsicherheit und Regulierung identifiziert.

Bei den Argumenten wurde eine Vielzahl von Punkten für und gegen den Einsatz von generativer KI eingebracht. Befürworter:innen des Einsatzes Künstlicher Intelligenz zogen häufig Parallelen zu menschlichen Fähigkeiten und führten überzeugende Argumente oder konkrete Fortschritte in diesem Bereich an. So hob Nutzer CfP1 etwa die Rolle der KI bei der Entdeckung eines neuartigen Antibiotikums hervor. Die Überlegenheit der KI wurde von einigen Diskussionsteilnehmer:innen in Frage gestellt. Eine Person argumentierte hier folgendermaßen: „Mei, ich möchte nicht behaupten KI wär zu gar nichts gut. Mich stört nur das Framing, und das ganz gewaltig: Klimawandel ist ein riesen Problem für das wir dummen Menschen keine Lösung haben, wir hoffen also darauf dass die superintelligente KI die Lösung für uns findet und uns alle rettet" (gottwürfeltnicht). Ein wiederkehrendes Thema in kritischen Kommentaren ist der Diskurs über allgemeine Gefahren, oft begleitet von Katastrophenszenarien. Dabei beziehen sich User:innen beispielsweise auf Filme wie *Terminator* oder *I-Robot*. Ein User gibt hier zu Bedenken: „Ich hoffe darauf, dass wir uns als Menschheit bewusst entscheiden, was wir wollen und was nicht. Dass darüber breit und ernsthaft diskutiert wird und dass KI nichts ist, das wie eine Naturgewalt über uns hereinbricht" (palmeninhelsinki). Ähnlich verglich ein als Synchronsprecher tätiger Nutzer die neuen Entwicklungen im Bereich *GenAI* für die Vertonung von Videos mit einer Flutkatastrophe und argumentierte folgendermaßen: „was für eine riesengroße Katastrophe das bedeutet, kann man sich wahrscheinlich gar nicht ausmalen. Versteht denn niemand, daß die Verwendung von KI in diesen und anderen Bereichen Konsequenzen fürchterlichsten Ausmaßes nach sich zieht? […] sogenannte Kreative schaffen Kreative ab. Und wir Kreative sind nur die Spitze des Eisbergs" (sprecher_omid). Einige Befürworter:innen generativer KI reagierten auf kritische Kommentare mit einem Stereotyp: „Die deutschesten Kommentare, die man zum technischen Fortschritt kriegen kann. Diese Bitterkeit gegenüber Fortschritt, macht mir mehr Sorgen, als jeglicher Fortschritt in der KI-Forschung" (ecnerwal).

Schließlich war von Interesse, welches Wissen User:innen in den Diskurs einbringen (FF3). Hierzu haben wir uns auf die Frage fokussiert, wie Lai-

en in ihren Argumenten über KI-Autorität und Legitimität konstruieren. Dementsprechend galt es, die Arten von Wissen, die von User:innen im Kontext des Diskurses geteilt wurden, zu identifizieren. Hier waren historisches und technisches Wissen auffällig relevant. Als Belege wurden häufig Quellen wie Science-Fiction-Filme, Fernsehserien, Dokumentationen, Zeitungsartikel und Wikipedia-Artikel herangezogen. Wissenschaftliche Quellen wurden in den analysierten Nutzer:innenkommentaren hingegen selten erwähnt. Die Nutzer:innen verwiesen beispielsweise auf Erkenntnisse aus Verhaltensexperimenten, wenn sie diskutierten, ob Bewusstsein empirisch getestet werden kann und nahmen hier Bezug zum sogenannten Spiegeltest. Interessanterweise relativierten manche Nutzer:innen ihr eigenes Wissen. So erklärte etwa Userin disgruntled_auntie: „Zugegebenermaßen bin ich nicht auf dem neuesten Stand" und „Ich kenne mich ein bisschen aus in dem Bereich (würde mich nicht als Expertin bezeichnen, habe aber selbst schon neuronale Netze trainiert)". Darüber hinaus wurde die Wissensbasis anderer Nutzer:innen häufig hinterfragt. Eine Person behauptete beispielsweise, das KI-Tool Lambda getestet zu haben. Mehrere stellten diese Behauptung in Frage. Schließlich gab die Person zu, das Tool nicht selbst getestet zu haben, sondern sich auf die Erfahrungen eines befreundeten Journalisten verlassen zu haben. Dieses Beispiel zeigt, dass Nutzer:innen im analysierten Diskurs anderen offenbar nicht blind vertrauten, sondern eher dazu neigten, Argumente zu hinterfragen und zu widerlegen.

6. Diskussion

Die vorliegende Studie hatte zum Ziel, Diskussionsstränge, Themen und Argumente sowie das in Nutzer:innendiskussionen über KI geteilte Wissen zu analysieren. Die Ergebnisse unserer Diskursanalyse von User:innenkommentaren auf der Nachrichtenwebsite und dem *Instagram*-Account der deutschen Wochenzeitung *Die Zeit* zeigen, dass die Diskussionen thematisch vielfältig waren und auch ethische Überlegungen – vor allem rund um die Ausbeutung von *clickworkern* – beinhalteten. Es wurde viel Erfahrungswissen geteilt, das jedoch auch von anderen User:innen hinterfragt wurde. Es zeigten sich vereinzelt Verweise auf wissenschaftliches Wissen – ein Muster, das auch in User:innendiskussionen zu anderen kontroversen Themen beobachtet wurde (vgl. Huber et al. 2019). Allerdings verwiesen Nutzer:innen häufiger auf Filme, Serien, Zeitungsartikel oder ähnliche Quellen. Auch Katastrophenszenarien wurden bedient – eine Erzählweise,

die auch in einer aktuellen Inhaltsanalyse der Medienberichterstattung über KI in deutschen Zeitungen festgestellt wurde (vgl. Ermler 2025). Obwohl die Qualität der Nutzer:innendiskurse nicht im Fokus dieser Studie stand, fällt auf, dass die Nutzer:innenkommentare auf der Online-Nachrichtenwebsite umfangreicher waren als auf *Instagram*. Diese Beobachtung wirft die Frage auf, wie sich die zunehmende Verlagerung des User:innendiskurses von Nachrichtenwebsites zu Social-Media-Plattformen auf die Qualität der User:innendiskurse auswirken wird. Hier sind aufbauende Studien in Längsschnittdesign gefragt, die im quantitativen Stil größere Mengen an User:innenkommentaren im Zeitverlauf systematisch analysieren.

Die vorliegende Studie hat einige Limitationen. Einerseits ist die Auswahl von *Die Zeit* als einzigem Medium und *Instagram* als einziger Social-Media-Plattform nicht aussagekräftig für die deutsche Medienlandschaft. Zukünftige Studien sollten daher eine größere Bandbreite an Zeitungen und Social-Media-Plattformen einbeziehen. Zudem hat die in dieser Studie ausgewählte Wochenzeitung eine eher hochgebildete Leser:innenschaft (vgl. ZEIT Advise, 2026, S. 5). Die Auswahl von Kommentarbereichen aus Zeitungen mit anders gestalteter Leser:innenschaft könnte andere Argumente zutage fördern und zu Unterschieden hinsichtlich der gemeinsamen Wissensbestände führen. Abgesehen von diesen Einschränkungen liefert die vorliegende Studie erste relevante Einblicke in Nutzer:innendiskurse über generative KI und bildet eine solide Grundlage für zukünftige Studien.

Tabelle A1 Analysekorpus

			Website		Instagram	
Nr	Datum	Artikel	Anzahl Postings	Anzahl analysierte Postings	Anzahl Postings	Anzahl analysierte Postings
1	19. Januar 2023	Künstliche Intelligenz: Hast du ein Bewusstsein?	237	40	67	40
2	22. Januar 2023	ChatGPT: Ausgebeutet, um die KI zu zähmen	107	40	325	40
3	7. April 2023	Trauer und KI: Sie haben einen verpassten Anruf Ihrer toten Mutter	75	40	481	40
4	26. Mai 2023	Sam Altman im Interview: „Es ist gefährlich, künstliche Intelligenz zu vermenschlichen"	405	40	31	7[2]
5	14. September 2023	Ein neues KI-Tool ermöglicht es, Videos zu übersetzen und sogar Lippenbewegungen anzupassen	115	40	200	40
6	18. Juni 2024	Wie ihr eure Daten vor der Meta-KI schützen könnt	20	20	315	40
		Total	959	220	1419	207

Danksagung

Diese Studie ist Teil des von der IU Internationale Hochschule geförderten Forschungsprojekts „Research Umbrella KI in Marketing und Kommunikation".

Literatur

Anderson, Ashley A. et al. (2014): The "Nasty Effect": Online incivility and risk perceptions of emerging technologies, in: Journal of Computer-Mediated Communication 19 (3/2014), S. 373–387. https://doi.org/10.1111/jcc4.12009

Anter, Luise / Kümpel, Anna Sophie (2023): Young adults' information needs, use, and understanding in the context of Instagram: a multi-method study, in: Digital Journalism, S. 119. https://doi.org/10.1080/21670811.2023.2211635

2 Instagram-Post zeigte bei der Anzahl der Kommentare 31 an, aber nur 7 Kommentare waren sichtbar / konnten abgerufen werden.

Baqir, Anees et al. (2025): Unveiling the drivers of active participation in social media discourse, in: Scientific Reports 15 (1/2025). https://doi.org/10.1038/s41598-025-88117-x

Brantner, Cornelia / Saurwein, Florian (2021): Covering technology risks and responsibility: Automation, artificial intelligence, robotics, and algorithms in the media, in: International Journal of Communication 15, S. 5074–5098. https://doi.org/10.21083/ijoc.v15i0.1905

Brause, Saba Rebecca et al. (2023): Media representations of artificial intelligence: surveying the field, in: Simon Lindgren (Hg.), Handbook of Critical Studies of Artificial Intelligence, Cheltenham/Northampton, S. 277–288.

Brown, Danielle K. et al. (2018): Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations, in: Social Media + Society 4(4). https://doi.org/10.1177/2056305118810216

Brubaker, Pamela J. / Montez, Daniel / Church, Scott Haden (2021): The power of Schadenfreude: Predicting behaviors and perceptions of trolling among Reddit users, in: Social Media + Society 7 (2/2021). https://doi.org/10.1177/20563051211021382

Cammaerts, Bart / Mattoni, Alice / McCurdy, Patrick (Hg.) (2013): Mediation and Protest Movements, Bristol.

Chadwick, Andrew (2017): The Hybrid Media System, New York.

Chan, Michael / Yi, Jingjing (2024): Social media use and political engagement in polarized times: Examining the contextual roles of issue and affective polarization in developed democracies, in: Political Communication 41 (5/2024), S. 743–762. https://doi.org/10.1080/10584609.2024.2325423

Coe, Kevin / Kenski, Kate / Rains, Stephen A. (2014): Online and uncivil? Patterns and determinants of incivility in newspaper website comments, in: Journal of Communication 64 (4/2014), S. 658–679. https://doi.org/10.1111/jcom.12104

Cools, Hannes / Van Gorp, Baldwin / Opgenhaffen, Michael (2024): Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers, in: Journalism 25 (1/2024), S. 3–21. https://doi.org/10.1177/14648849221122647

Kimmerle, Joachim et al. (2015): Learning and collective knowledge construction with social media: A process-oriented perspective, in: Educational Psychologist 50 (2/2015), S. 120–137. https://doi.org/10.1080/00461520.2015.1036273

Dahlgren, Peter (2005): The internet, public spheres, and political communication: dispersion and deliberation, in: Political Communication 22 (2/2005), S. 147–162. https://doi.org/10.1080/10584600590933160

Del Valle, Marc Esteve et al. (2020): Learning in the wild: Understanding networked ties in Reddit, in: Nina Bonderup Dohn et al. (Hg.), Mobility, Data and Learner Agency in Networked Learning, Cham, S. 51–68. https://doi.org/10.1007/978-3-030-36911-8_4

Domingo, David (2008): Interactivity in the daily routines of online newsrooms: Dealing with an uncomfortable myth, in: Journal of Computer-Mediated Communication 13 (3/2008), S. 680–704.

Duberry, Jérôme / Hamidi, Sabrya (2021): Contrasted media frames of AI during the COVID-19 pandemic: a content analysis of US and European newspapers, in: Online Information Review 45 (4/2021), S. 758–776. https://doi.org/10.1108/OIR-09-2020-0393

Eberwein, Tobias (2020): "Trolls" or "warriors of faith"? Differentiating dysfunctional forms of media criticism in online comments, in: Journal of Information, Communication and Ethics in Society 18 (4/2020), S. 575–587. https://doi.org/10.1108/JICES-08-2019-0090

Eisenegger, Mark / Schäfer, Mike S. (2023): Reconceptualizing public sphere(s) in the digital age? On the role and future of public sphere theory, in: Communication Theory 33 (2-3/2023), S. 61–69. https://doi.org/10.1093/ct/qtad011

Ermler, Kim Lisa (2025): Zwischen Dampfmaschine und Atomwaffen – Frames in der Medienberichterstattung über die Vertrauenswürdigkeit von Künstlicher Intelligenz (Unveröffentlichte Masterarbeit), Universität Bremen.

Ferree, Myra Marx et al. (2002): Four models of the public sphere in modern democracies, in: Theory and Society 31 (3/2002), S. 289–324.

Fischer, Sarah / Puschmann, Cornelius (2021): Wie Deutschland über Algorithmen schreibt: Eine Analyse des Mediendiskurses über Algorithmen und Künstliche Intelligenz (2005–2020), hrsg. von der Bertelsmann Stiftung, Gütersloh. https://doi.org/10.11586/2021003

Fischer, Renate / Jarren, Otfried (2024): The platformization of the public sphere and its challenge to democracy, in: Philosophy & Social Criticism 50 (1/2024), S. 200-215. https://doi.org/10.1177/01914537231203535

Friess, Dennis / Eilders, Christiane (2015): A systematic review of online deliberation research, in: Policy & Internet 7 (3/2015), S. 319–339. https://doi.org/10.1002/poi3.95

Gaudette, Tiana et al. (2021): Upvoting extremism: Collective identity formation and the extreme right on Reddit, in: New Media & Society 23 (12/2021), S. 3491–3508. https://doi.org/10.1177/1461444820958123

Geers, Sabine (2020): News consumption across media platforms and content: A typology of young news users, in: Public Opinion Quarterly 84 (S1/2020), S. 332–354. https://doi.org/10.1093/poq/nfaa010

Gerhards, Jürgen / Neidhardt, Friedhelm (1990): Strukturen und Funktionen moderner Öffentlichkeit: Fragestellungen und Ansätze (= WZB Discussion Paper 90–101/1990), Berlin.

Gonçalves, Joao (2018): Aggression in news comments: how context and article topic shape user-generated content, in: Journal of Applied Communication Research 46 (5/2018), S. 604–620. https://doi.org/10.1080/00909882.2018.1529419

Gorwa, Robert / Binns, Reuben / Katzenbach, Christian (2020): Algorithmic content moderation: Technical and political challenges in the automation of platform governance, in: Big Data & Society 7 (1/2020), S. 1–15. https://doi.org/10.1177/2053951719897945

Graham, Timothy et al. (2021): #IStandWithDan versus #DictatorDan: The polarised dynamics of Twitter discussions about Victoria's COVID-19 restrictions, in: Media International Australia 179 (1/2021), S. 127–148. https://doi.org/10.1177/1329878X20981780

Hagendorff, Thilo (2024): Mapping the ethics of generative AI: A comprehensive scoping review, in: ArXiv, 13. Februar 2024 (online unter: https://arxiv.org/abs/2402.08323 – letzter Zugriff: 06.12.25).

Hall, Stuart (1980): Encoding/decoding, in: Stuart Hall et al. (Hg.), Culture, Media, Language, London, S. 128–138.

Halpern, Daniel / Gibbs, Jennifer (2013): Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression, in: Computers in Human Behavior 29 (3/2013), S. 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Hara, Noriko et al. (2025): Exploring the dynamics of interaction about generative artificial intelligence between experts and the public on social media, in: Journal of Science Communication 24 (1/2025), A02. https://doi.org/10.22323/2.24010202

Hsueh, Mark / Yogeeswaran, Kumar / Malinen, Sanna (2015): "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?, in: Human Communication Research 41 (4/2015), S. 557–576. https://doi.org/10.1111/hcre.12059

Huber, Brigitte / Wetzstein, Irmgard / Aichberger, Ingrid (2019): Societal problem solver or deficient discipline? The debate about social science in the online public sphere, in: Journal of Science Communication 18 (02/2019), A04. https://doi.org/10.22323/2.18020204

Keller, Reiner (2011): The sociology of knowledge approach to discourse (SKAD), in: Human Studies 34, S. 43–65. https://doi.org/10.1007/s10746-011-9175-z

Keller, Reiner (2005): Analysing Discourse: An approach from the sociology of knowledge, in: Forum Qualitative Sozialforschung 6 (3/2005). https://doi.org/10.17169/fqs-6.3.19

Kieslich, Kimon / Došenović, Pero / Marcinkowski, Frank (2022): Alles, nur kaum Science-Fiction: Eine Themenanalyse der deutschen Medienberichterstattung über Künstliche Intelligenz (= Meinungsmonitor Künstliche Intelligenz, Factsheet Nr. 7 – Oktober 2022), in: CAIS Research (online unter: https://www.cais-research.de/wp-content/uploads/Factsheet-7-Medienberichterstattung.pdf – letzter Zugriff: 4.2.2026).

Kim, Jin Woo et al. (2021): The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity, in: Journal of Communication 71 (6/2021), S. 922–946.

Klinger, Ulrike (2018): Aufstieg der Semiöffentlichkeit: Eine relationale Perspektive, in: Publizistik 63 (2/2018), S. 245–267. https://doi.org/10.1007/s11616-018-0421-5

Korneeva, Ekaterina et al. (2023): Tracing the legitimacy of Artificial Intelligence: A longitudinal analysis of media discourse, in: Technological Forecasting and Social Change 192. https://doi.org/10.1016/j.techfore.2023.122467

Kreiss, Daniel / McGregor, Shannon C. (2024): A review and provocation: On polarization and platforms, in: New Media & Society 26(1), S. 556–579. https://doi.org/10.1177/14614448231161880

Kubin, Emily et al. (2024): Understanding news-related user comments and their effects: a systematic review, in: Frontiers in Communication 9. https://doi.org/10.3389/fcomm.2024.1447457

Lee, Eun-Ju / Jang, Yoon Jae (2010): What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception, in: Communication Research 37 (6/2010), S. 825–846. https://doi.org/10.1177/0093650210376189

Lörcher, Ines / Taddicken, Monika (2017): Discussing climate change online. Topics and perceptions in online climate change communication in different online public arenas, in: Journal of Science Communication 16 (2/2017), A03. https://doi.org/10.22323/2.16020203

Mak, Macau K. F / Li, Mengyu / Rojas, Hernando (2024): Social media and perceived political polarization: Role of perceived platform affordances, participation in uncivil political discussion, and perceived others' engagement, in: Social Media + Society 10 (1/2024). https://doi.org/10.1177/20563051241228595

Malinen, Sanna (2015): Understanding user participation in online communities: A systematic literature review of empirical studies, in: Computers in Human Behavior 46, S. 228–238. https://doi.org/10.1016/j.chb.2015.01.004

Manosevitch, Eedith / Walker, Dana (2009): Reader comments to online opinion journalism: A space of public deliberation (= 10th International Symposium on Online Journalism, Austin, TX, April 17–18, 2009).

Miyazaki, Kunihiro et al. (2024): Public perception of generative AI on Twitter: An empirical study based on occupation and usage, in: EPJ Data Science 13 (1/2024), Article 2. https://doi.org/10.1140/epjds/s13688-023-00445-y

Newman, Nic et al. (2024): Reuters Institute Digital News Report 2024, hrsg. von Reuters Institute for the Study of Journalism, Oxford.

Nguyen, Dennis / Hekman, Erik (2022): The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation, in: AI & Society 39 (2/2022), 437–451. https://doi.org/10.1007/s00146-022-01511-1

Pfetsch, Barbara / Löblich, Maria / Eilders, Christiane (2018): Dissonante Öffentlichkeiten als Perspektive kommunikationswissenschaftlicher Theoriebildung, in: Publizistik 63 (4/2018), S. 477–495. https://doi.org/10.1007/s11616-018-0441-1

Picone, Ike (2017): Conceptualizing media users across media: The case for 'media user/use' as analytical concepts, in: Convergence: The International Journal of Research into New Media Technologies 23 (4/2017), S. 378–390. https://doi.org/10.1177/1354856517700380

Qi, Weihong et al. (2024): Excitements and concerns in the post-ChatGPT era: Deciphering public perception of AI through social media analysis, in: Telematics and Informatics 92. https://doi.org/10.1016/j.tele.2024.102158

Quandt, Thorsten (2018): Dark participation, in: Media and Communication 6 (4/2018), S. 36–48. https://doi.org/10.17645/mac.v6i4.1519

Rega, Rossella / Marchetti, Rita / Stanziano, Anna (2023): Incivility in online discussion: An examination of impolite and intolerant comments, in: Social Media + Society 9 (2/2023). https://doi.org/10.1177/20563051231180638

Roe, Jasper / Perkins, Mike (2023): 'What they're not telling you about ChatGPT': Exploring the discourse of AI in UK news media headlines, in: Humanities and Social Sciences Communications 10 (1/2023), S. 1–9. https://doi.org/10.1057/s41599-023-02282-w

Rogers, Everett M. (2003): Diffusion of innovations, 5. Aufl., New York.

Rossini, Patrícia (2022): Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk, in: Communication Research 49 (3/2022), S. 399–425. https://doi.org/10.1177/0093650220921314

Rowe, Ian (2015): Civility 2.0: a comparative analysis of incivility in online political discussion, in: Information, Communication & Society 18 (2/2015), S. 121–138. https://doi.org/10.1080/1369118x.2014.940365

Ruiz, Carlos et al. (2011): Public sphere 2.0? The democratic qualities of citizen debates in online newspapers, in: The International Journal of Press/Politics 16 (4/2011), S. 463–487.

Rusche, Christian et al. (2022): KI-Monitor 2022. Künstliche Intelligenz in Deutschland, Gutachten im Auftrag des Bundesverbandes Digitale Wirtschaft (BVDW) e.V., Köln.

San Cornelio, Gemma (2022): Instagram aesthetics for social change: a narrative approach to visual activism on Instagram, in: William Housley et al. (Hg.), The SAGE Handbook of Digital Society, London, S. 188–208. https://doi.org/10.4135/9781529783193.n12

Santana, Arthur D. (2014): Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards, in: Journalism Practice 8 (1/2014), S. 18–33. https://doi.org/10.1080/17512786.2013.813194

Schicha, Christian (2022): Streitkultur statt Cancel Culture – Ein Plädoyer für eine offene Diskurskultur bei kontroversen Debatten in: Gürtler Christian / Marlis Prinzing / Thomas Zeilinger (Hg.), Streitkulturen. Medienethische Perspektiven auf gesellschaftliche Diskurse, Baden-Baden, S. 113–132.

Schlesinger, Philip (2024): The post-public sphere and neo-regulation of digital platforms, in: Javnost – The Public 31 (1/2024), S. 64-88. https://doi.org/10.1080/13183222.2024.2311010

Schrøder, Kim Christian / Larsen, Bent Steeg (2010): The shifting cross-media news landscape: Challenges for news producers, in: Journalism Studies 11 (4/2010), S. 524–534. https://doi.org/10.1080/14616701003638392

Shaikh, Sonia Jawaid / Moran, Rachel E. (2024): Recognize the bias? News media partisanship shapes the coverage of facial recognition technology in the United States, in: New Media & Society 26 (5/2024), S. 2829–2850. https://doi.org/10.1177/14614448221090916

Singer, Jane B. (2010): Quality control: Perceived effects of user-generated content on newsroom norms, values and routines, in: Journalism Practice 4 (2/2010), S. 127–142. https://doi.org/10.1080/17512780903391979

Stroud, Natalie Jomini / Scacco, Joshua M. / Curry, Alexander L. (2016): The presence and use of interactive features on news websites, in: Digital Journalism 4 (3/2016), S. 339–358. https://doi.org/10.1080/21670811.2015.1042982

Su, Leona Yi-Fan et al. (2018): Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets, in: New Media & Society 20 (10/2018), S. 3678–3699. https://doi.org/10.1177/1461444818757205

Taddicken, Monika et al. (2020): Wirtschaftlicher Nutzen statt gesellschaftlicher Debatte? Eine quantitative Framing-Analyse der Medienberichterstattung zum autonomen Fahren, in: Medien & Kommunikationswissenschaft 68 (4/2020), S. 406–427. https://doi.org/10.5771/1615-634X-2020-4-406

Toepfl, Florian / Piwoni, Eunike (2018): Targeting dominant publics: How counterpublic commenters align their efforts with mainstream news, in: New Media & Society 20 (5/2018), S. 2011–2027. https://doi.org/10.1177/1461444817712085

Treré, Emiliano (2015): Reclaiming, proclaiming, and maintaining collective identity in the #YoSoy132 movement in Mexico: an examination of digital frontstage and backstage activism through social media and instant messaging platforms, in: Information, Communication & Society 18 (8/2015), S. 901–915. https://doi.org/10.1080/1369118X.2015.1043744

Tsimpoukis, Panos (2025): Contesting dominant AI narratives on an industry-shaped ground: Public discourse and actors around AI in the french press and social media (2012–2022), in: Journal of Science Communication 24 (2/2025). https://doi.org/10.22323/2.24020210

Vergeer, Maurice (2020): Artificial intelligence in the Dutch press: An analysis of topics and trends, in: Communication Studies 71 (3/2020), S. 373–392. https://doi.org/10.1080/10510974.2020.1733038

Sikorski, Christian / Hänelt, Maria (2016): Scandal 2.0: How Valenced Reader Comments Affect Recipients' Perception of Scandalized Individuals and the Journalistic Quality of Online News, in: Journalism & Mass Communication Quarterly 93 (3/2016), S. 551–571. https://doi.org/10.1177/1077699016628822

Wimmer, Jeffrey (2007): (Gegen-)Öffentlichkeit in der Mediengesellschaft. Analyse eines medialen Spannungsverhältnisses, Wiesbaden.

Winkel, Marek (2024): Controlling the uncontrollable: the public discourse on artificial intelligence between the positions of social and technological determinism, in: AI & Society 40, S. 1947–1959. https://doi.org/10.1007/s00146-024-01979-z

Wittemann, Stephanie (2024): KI-Mediendiskurs vor und nach ChatGPT. Wie sich der Diskurs um KI durch eine Veröffentlichung verändert, (Unveröffentlichte Bachelorarbeit), IU International University of Applied Sciences.

ZEIT Advise (2026). Die Zeit Preisliste 2026. https://advise.zeit.de/wp-content/uploads/2026/02/260202_DIE-ZEIT_Preisliste-2026.pdf

Zeller, Frauke et al. (2014): A subjective user-typology of online news consumption, in: Digital Journalism, 2 (2/2014), S. 214–231. https://doi.org/10.1080/21670811.2013.801686

Zeng, Jing / Chan, Chung-Hong / Schäfer, Mike S. (2020): Contested Chinese Dreams of AI? Public discourse about Artificial intelligence on WeChat and People's Daily Online, in: Information, Communication & Society 25 (3/2020), S. 319–340. https://doi.org/10.1080/1369118X.2020.1776372

Ziegele, Marc (2019): Reader commenting, in: Tim P. Vos / Folker Hanusch (Hg.), The International Encyclopedia of Journalism Studies, Düsseldorf, S. 1–8. https://doi.org/10.1002/9781118841570.iejs0059

Ziegele, Marc (2025): Integration oder Spaltung? Status quo und Verbesserungspotenziale der Online-Diskussionen von Bürger: innen an der Schnittstelle zwischen Gesellschaft, Medien und Politik, in: Lothar Häberle (Hg.), Mainstream – freie Meinung–Populismus: Interdisziplinäre Beiträge zur Debattenkultur und zu Spaltungstendenzen der Gesellschaft, Wiesbaden, S. 129–154.

Ziegele, Marc / Jost, Pablo (2016): Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments, in: Communication Research 47 (6/2016), S. 891–920. https://doi.org/10.1177/0093650216671854

Ziegele, Marc et al. (2018): Online user comments across news and other content formats: Multidisciplinary perspectives, new directions, in: SCM Studies in Communication and Media 6 (4/2018), S. 315–332.

Zimmermann, Christine (2013): „Same-sex marriage “und der amerikanische Kulturkampf: Ein „familiärer “Diskurs zur (Re-) Konstruktion einer Institution, in: Reiner Keller / Inga Truschkat (Hg.), Methodologie und Praxis der Wissenssoziologischen Diskursanalyse: Band 1: Interdisziplinäre Perspektiven, Wiesbaden, S. 221–247.

Zolkepli, Izzal Asnira / Kamarulzaman, Yusniza (2015): Social media adoption: The role of media needs and innovation characteristics, in: Computers in Human Behavior 43, S. 189–209. https://doi.org/10.1016/j.chb.2014.10.050

Zou, Wenxue et al. (2025): Exploring the early adoption of open AI among laypeople and technical professionals: An analysis of Twitter conversations on #ChatGPT and #GPT3, in: International Journal of Human-Computer Interaction 41 (1/2025), S. 149–160. https://doi.org/10.1080/10447318.2023.2295725

KI-Modelle in Medienunternehmen: empirische Befunde und ethische Reflexionen für die Regulierung

Michael Litschka

Zusammenfassung

Der Beitrag untersucht die Auswirkungen und Herausforderungen des Einsatzes von KI-Modellen (unter anderem auch Sprachmodellen) in Medienunternehmen aus empirischer und ethischer Perspektive. Basierend auf einer für eine Regulierungsbehörde durchgeführten Studie, die unter anderem Expert:inneninterviews und eine SWOT-Analyse umfasste, werden zentrale Chancen und Risiken für ökonomische Wertschöpfung und ethische Verantwortung identifiziert. Die Ergebnisse zeigen einerseits mögliche Effizienzgewinne und Innovationspotenziale, andererseits erhebliche ethische Problemfelder wie algorithmische Verzerrung, fehlender Datenschutz, mangelnde Transparenz und die Gefahr des Verlusts journalistischer Qualität. Für die wahrgenommenen Risiken wünschen sich die an der Studie teilnehmenden Medienunternehmen klarere und stärkere Regulierungsvorschriften. In der auf diese empirischen Befunde folgenden ethischen Diskussion beschreibt der Beitrag einige aktuelle medien- und KI-ethische Ansätze, wobei er insbesondere die Gerechtigkeitstheorien nach Rawls (hinsichtlich substanzieller Chancengleichheit und öffentlicher Legitimation) und Sen (hinsichtlich „vergleichender" Gerechtigkeitsanalysen und Pluralismus) systematisch gegenübergestellt und für die Entwicklung pluralistisch ausgerichteter, gesellschaftlich akzeptierter Regulierungsinstrumente fruchtbar macht. Beide Ansätze zeigen (und stützen somit Regulierungsbefürworter:innen), dass nicht Regulierung per se unternehmerische Freiheiten einschränkt, sondern der ethisch problematische Einsatz von KI Werte wie Autonomie und Freiheit gefährdet. Der Beitrag schließt mit der Empfehlung, sowohl substanzielle Chancengleichheit als auch globale Pluralität und partizipative Stakeholder-Dialoge als Grundlagen für eine verantwortungsvolle KI-Regulierung im Mediensektor zu verankern.

1. Einleitung

Der vorliegende Beitrag beschreibt *einerseits* (siehe Kapitel 2) eine empirische Untersuchung, die im Auftrag einer österreichischen Regulierungsbehörde durchgeführt wurde und eine umfassende Analyse des Einsatzes von KI-Anwendungen, insbesondere Sprachmodellen, in österreichischen und internationalen Medienunternehmen sowie Plattformen beinhaltete (siehe dazu Belinskaya et al. 2024 und Pinzolits et al. 2025). Ziel war es, sowohl ökonomische als auch gesellschaftliche Dimensionen zu beleuchten, wobei eine SWOT-Analyse entlang der Medien-Wertschöpfungskette im Mittelpunkt stand. Der Fokus lag insbesondere auf der Selbstsicht der Akteur:innen, um deren subjektive Wahrnehmungen und strategische Einschätzungen differenziert herauszuarbeiten.[1] Da diese Erhebung neben den technologischen Chancen auch viele gesellschaftliche und ethische Risiken der KI-Nutzung offenbart und der Wunsch nach einer stärkeren Regulierung aufkam, erfolgt nach der Beschreibung einiger empirischer Erkenntnisse aus der Studie eine genauere Analyse der Frage, wie eine solche Regulierung auch ethisch fundiert werden kann.

Der Beitrag versucht somit *andererseits* (siehe Kapitel 3 und 4), die aufkommenden Fragen in einen medien- und KI-ethischen Rahmen zu stellen, der in der zugrundeliegenden empirischen Studie nur angedeutet wurde. Dazu werden insbesondere die Gerechtigkeitstheorien nach Rawls und Sen herangezogen, systematisch gegenübergestellt und für die Entwicklung pluralistisch ausgerichteter, gesellschaftlich akzeptierter Regulierungsinstrumente fruchtbar gemacht. Beide Ansätze argumentieren (und stützen somit Regulierungsbefürworter:innen), dass nicht Regulierung *per se* unternehmerische Freiheiten einschränkt, sondern der ethisch problematische Einsatz von KI-Werte wie Autonomie und Freiheit gefährden kann. Ebenso zeigen beide Ansätze, dass eine öffentliche Rechtfertigung zum Beispiel einer KI-Strategie eines Unternehmens oder einer KI-Regulierung durch Behörden private Rechtfertigungen gewinnorientierten Handelns (durch Medienunternehmen und Plattformen) aussticht.

1 Eine Folgestudie für denselben Auftraggeber mit dem Ziel einer repräsentativen Erhebung der Einschätzung der Bevölkerung zum Einsatz von KI in Medien wurde im Oktober 2025 präsentiert (Pinzolits et al. 2025).

2. SWOT-Analyse zur Selbsteinschätzung der befragten Medienunternehmen

Die dem Beitrag zugrundeliegende Studie (teilweise veröffentlicht in Belinskaya et al. 2024 und Pinzolits et al. 2025) folgte einem multi-methodischen Design:

- Literaturstudie zum *status quo* von KI-Anwendungen im Mediensektor
- Sechzehn leitfadengestützte Expert:inneninterviews zur praktischen Integration und Bewertung von KI-Technologien im Medienworkflow (Personen im Management von Printmedien, Radio, öffentlich-rechtlicher Rundfunk, Universität, HAW, Presseagentur, Unterhaltungsmedien, Kreativagentur, KI-Consulting, KI-Softwareentwicklung)
- Eine daraus folgende Analyse der Selbsteinschätzung dieser Expert:innen zu Chancen und Risiken von KI-gestützten Medienprozessen (SWOT-Analyse) entlang der relevanten Wertschöpfungsstufen (siehe auch Abbildung 1).
- Eine systematische Übersicht aktueller Regulierungsrichtlinien
- Überlegungen zu KI-ethischen Inputs für künftige Regulierungsaktivitäten.

Im Folgenden sollen einige Ergebnisse der Interviews und der daraus hervorgegangenen SWOT-Analyse dargestellt werden. Ziel ist es, die anhand von zwei Wertschöpfungsstufen (*creation* und *editing*) beispielhaft herausgefilterten Chancen und Risiken aus Sicht der Akteur:in darzustellen und für jene Medienunternehmen, die im Bereich Journalismus und Unterhaltung tätig sind, möglichst zu verallgemeinern. Für diese verallgemeinerten Problemfelder gilt es dann, medien- und KI-ethisch reflektierte Regulierungszugänge zu diskutieren. Die Literaturstudie und die Übersicht zu aktuellen Regulierungsrichtlinien spielen für die folgende Argumentation keine Rolle; die für diesen Beitrag neu detaillierten KI-ethischen Überlegungen folgen in Kapitel 3 und 4.

KI-Sprachmodelle werden bereits in vielen internationalen Medien genutzt: In Schweden nutzt beispielsweise *Aftonbladet* KI, um Texte in Rap-Songs umzuwandeln; *Reuters* und *Synthesia* haben einen vollautomatisierten Nachrichtenzusammenfassungsdienst mit virtuellem Moderator; *Associated Press* verwendet Automatisierungstechnologien zur Berichterstattung über *Minor League* Baseball; Bloomberg nutzt ein Sprachmodell, um automatisierte Antworten auf Kundenfragen zu liefern; in der Radiobranche wird mit *voice-cloning* und Sprachsynthese-Technologien experimentiert; in der Forschung werden *LLMs* (*Large Language Models*) für Textgene-

rierung und -analyse eingesetzt. Wenn wir eine (von mehreren möglichen) typische Wertschöpfungskette eines Medienunternehmens heranziehen (siehe Abbildung 1), ergeben sich Themenfelder für eine so genannte SWOT-Analyse, wie sie in der Betriebswirtschaftslehre gerne verwendet wird.

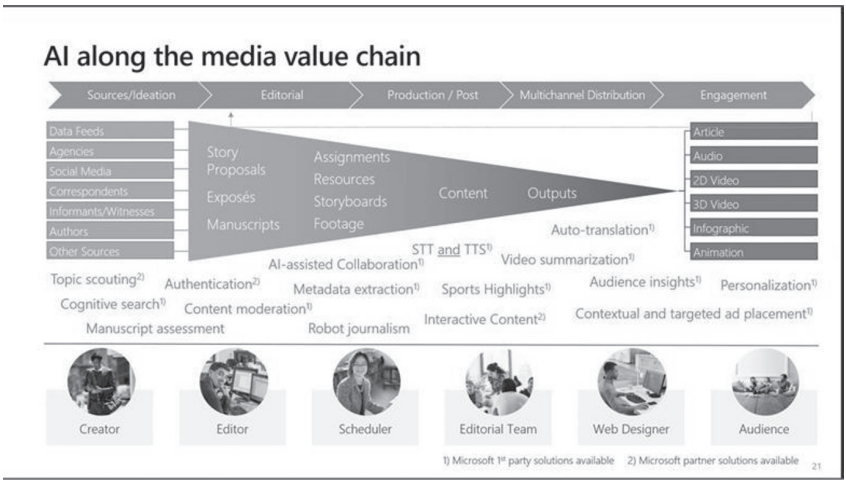

Abbildung 1: KI entlang der Medien-Wertschöpfungskette (Quelle: www.ibc.org)

Die SWOT-Analyse dient zur Identifikation von Stärken, Schwächen, Chancen und Risiken von Unternehmensstrategien und -ressourcen und wurde in diesem Fall auf die Implementierung von KI-Technologien im Mediensektor entlang der Wertschöpfungskette bezogen. Anbei ein Beispiel für den Bereich „Creation", wie dieser sich aus Sicht der befragten Interviewpartner:innen darstellt:

Abbildung 2: SWOT-Bereich „Creation“

Demnach können Stärken des vermehrten Einsatzes von KI im Medien-Workflow etwa eine Effizienzsteigerung durch die Automatisierung repetitiver Aufgaben, eine Freisetzung kreativer Ressourcen bei der Contenterstellung durch KI-gestützte Tools sowie eine Erleichterung datenintensiver Prozesse wie Recherche, Marktanalyse und Musterdetektion sein. Schwächen hingegen sehen die Befragten beim zeit- und kostenintensiven Training der KI-Systeme, der schwierigen Überprüfung der erstellten Inhalte, dem Zugriff auf große Nutzerdatenmengen (Stichwort Datenschutz) und der Neigung zu Halluzinationen. Die befragten Interviewpartner:innen sehen durchaus das Risiko der Verbreitung falscher Informationen durch fehlerhafte oder verzerrte KI-Ausgaben und einen möglichen *bias* der Trainingsdaten, etwa bei der Restrukturierung großer Datenmengen.

Chancen ergeben sich durch neue kreative Möglichkeiten durch die Integration von KI in den Produktionsprozess (zum Beispiel KI-generierte Musiktexte, automatisierte Nachrichtenaggregation), durch die Unterstützung des Innovationspotenzials von Medienunternehmen im internationalen Konkurrenzumfeld und durch Geschäftsmodellinnovation und Erschließung neuer Märkte. Auch die Schaffung neuer, KI-zentrierter Jobs wird als Möglichkeit wahrgenommen. Risiken ergeben sich durch die mögliche Erosion eigener kreativer Kompetenzen und die Gefahr zunehmender Abhängigkeit von KI-Lösungen großer Plattformen und deren proprietären Lösungen, die Marginalisierung assistierender Tätigkeiten am Arbeitsmarkt sowie das Potenzial für gesellschaftliche Desinformation und Manipulation, mangelnde Qualitätskontrollen und Vertrauenswürdigkeit. Gerade für das

Feld des Journalismus wurden auch die Gefahren des „automation creep" als fortschreitender Ersatz menschlicher Arbeit durch Maschinen und KI, enthumanisierter Journalismus und algorithmusgetriebener Sensationalismus genannt.

Ein zweites Beispiel für den Bereich „Editing" und wie dieser sich aus Sicht der befragten Interviewpartner:innen im Rahmen einer SWOT-Analyse darstellt:

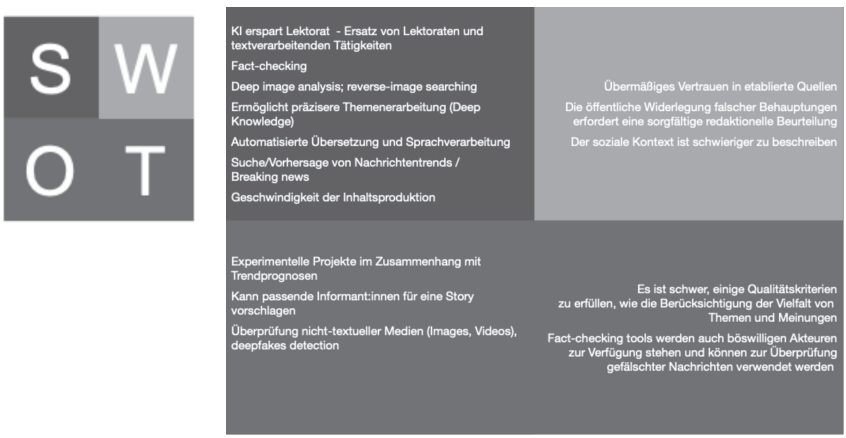

Abbildung 3: SWOT-Bereich „Editing"

In der Medienproduktion wird die Rolle von KI im Bereich des Editing vielschichtig betrachtet. KI-Technologien bieten das Potenzial, traditionelle Arbeitsschritte zu ersetzen, man sieht aber das Bedürfnis, dass journalistische Kernkompetenzen gewahrt werden müssen. Während die automatisierte Weiterverarbeitung von Inhalten als wichtig angesehen wird, bleibt die Kennzeichnung von KI-erstellten Elementen unerlässlich. Nur repetitive Aufgaben oder etwa das Lektorat werden erleichtert, nicht jedoch das Verständnis komplexer Zusammenhänge überflüssig; besondere Verantwortung kommt hier den Führungskräften in Medienunternehmen zu, die sicherstellen müssen, dass Qualitätskriterien der journalistischen Arbeit gewährleistet bleiben. Die sorgfältige redaktionelle Beurteilung und die Beachtung des sozialen Kontexts bleiben wahrgenommene Schwächen der KI-Tools.

KI-Tools erleichtern die Themenfindung, indem sie wichtige Themen aus umfangreichen Dokumenten identifizieren und Vorschläge für Inter-

viewfragen generieren können, womit sie in der journalistischen Prozesskette unterstützen können. Hinsichtlich neuer KI-basierter Formate gibt es unterschiedliche Chancen, von der Entwicklung einer datenschutzkonformen KI-Lösung über die Verbesserung der Überprüfung nicht-textlicher Inhalte bis hin zur *deep fake detection*.

Die Darstellung der Einschätzung der Expert:innen betreffend möglicher Effizienzgewinne wird freilich nicht von allen Forscher:innen geteilt. So schätzt zum Beispiel Acemoglu (2024), dass zwar niedrig-qualifizierte Arbeiten produktiver vonstatten gehen würden, diese aber nur 5 Prozent aller Aufgaben in den nächsten 10 Jahren umfassen. Trabelsi (2024) stützt das Argument der Effizienzsteigerung durch KI durch nun besser mögliche *big data*-Analysen, betont aber das Risiko der Polarisierung und steigender Ungleichheiten auf Arbeitsmärkten. Auch Brynjolfsson (2025) sieht potenzielle Gewinne eher für den Niedriglohnsektor und bei seltener auftretenden Problemen, wenn Menschen wenig Vorerfahrungen besitzen. Keine Evidenz hingegen sieht er für Produktivitätsgewinne auf Industrieebene (unabhängig vom Sektor).

Es zeigt zusammenfassend (siehe zu den ausführlichen Ergebnissen Belinskaya et al. 2024) in dieser Selbsteinschätzung und der SWOT-Analyse, dass Medienunternehmen (in unserem Beispiel journalistisch arbeitende, unterhaltende und Technologieplattformen nutzende Medien und KI-Entwickler:innen) die Vorteile und Nachteile KI-gestützter Wertschöpfungsformen auf Medienmärkten recht klar benennen können.

Die Vorteile der Integration von KI- beziehungsweise Großen Sprachmodellen (*LLM*) in die Wertschöpfungskette werden vor allem in der Automatisierung repetitiver Aufgaben im Medienprozess gesehen. Beispielsweise werden Recherchetätigkeiten, Datenanalysen und kreative Content Produktion unterstützt. Viele Systeme sind in der Lage, Muster, Trends und thematisch relationale Tendenzen sichtbar zu machen und bieten somit den Anwender:innen effizientere *workflows* und neue Zugänge zur Dateninterpretation.

Allerdings sind den teilnehmenden Expert:innen die möglichen Nachteile der Technologienutzung durch zum Beispiel größere Abhängigkeit, weniger Verständlichkeit (Stichwort „Explainability" der KI: die zunehmende Abstraktion und Intransparenz der Prozesse führt zu Verständniseinbußen bei weniger technikaffinen Akteur:innen), möglicherweise vermehrte Falschinformationen und durch *bias* verzerrte Entscheidungen der KI bewusst. Dazu kommen datenschutzrechtliche Bedenken sowie mögliche unbewusste Einflussnahmen auf die eigenständige (menschliche) Kreativar-

beit. Im Bereich journalistisch operierender Medienunternehmen ist die Hauptherausforderung das Erfüllen journalistischer Qualitätskriterien der Überprüfung, Datenqualität und Einhaltung des *human-in-the-loop*- Prinzips; im Bereich digitaler Plattformen geht es vor allem um algorithmische Transparenz. Nicht zuletzt sehen die befragten noch abzuwartende Umwälzungen auf dem Arbeitsmarkt und bei den erforderlichen Qualifikationen ihrer Mitarbeiter:innen.

Die Wahrnehmung der KI-Integration ist somit ambivalent: Während operative Vorteile durchaus anerkannt werden, dominiert eine kritische Grundhaltung hinsichtlich längerfristiger gesellschaftlicher und branchenspezifischer Risiken. Da immer wieder der Wunsch nach klaren Regulierungsrichtlinien aufkam, sollen nun für diese einige medien- und KI-ethische Ansätze auf ihre Fruchtbarkeit für künftige Regulierungsmaßnahmen analysiert werden.

3. Aktuelle medien- und KI-ethische Ansätze

Die ethische Reflexion des KI-Einsatzes im Medienbereich orientiert sich in der Literatur an mehreren etablierten theoretischen Strängen, die je nach Schwerpunkt der Autor:innen (siehe Abbild 3) von diesen detaillierter ausgearbeitet werden.

Theoriestrang/Konzept	Autor:innen	Jahr	Keywords
Tugendethisch-teleologisch Tugendethische Nutzung	Ess Spiekerman Cohen Vallor	2020 2019 2012 2024	Welche Art Mensch muss ich werden, in der ständigen Ausübung meiner technologischen Interaktionen, um zufrieden/glückselig zu sein? → *Eudaimonia;* *Wertebewusstsein* im Umgang mit (digitaler) Technologie; Frage nach dem „Warum" neuer technologischer Entwicklungen; vernünftige *Wertepriorisierung* und *Wertebalancierung*, so dass jede(r) ihr/sein Telos im Leben erreichen kann; Wertbewusste Programmierung
Kantianisch-deontologisch	Floridi et al.	2018	Digitale Technik muss einerseits die Kantische *Menschenwürde*, andererseits die möglichst komplette Selbstrealisierung („*self-realization*") unterstützen
Narrative Ethik und Werte	Grimm et al.	2019	Privatheit, Autonomie, Sicherheit als grundlegende Werte und Rechte; *narrative Ethik* der Digitalisierung
Autonomiediskussion	Thimm & Bächle Coeckelbergh	2019 2020	*Autonomie* als komplexe (und interdisziplinär zu klärende) Zuschreibung, also als *Bedingung für die Möglichkeit* der Übernahme ethischer Verantwortung (nicht als reine Beschreibung der technischen Entscheidungsmöglichkeiten von algorithmischen Plattformen)
Designprinzipien	Dignum	2019	*Accountability, Responsibility* und *Transparency* für die Produktion und Anwendung neuer Technologien
Gerechtigkeitszugänge, öff. Vernunftgebrauch	Sen, Rawls Habermas	2010 2001 1991	*Gerechtigkeit* der Datengebarung/des Datenzugangs; **Gerechtigkeit als Fairness; Komparative Gerechtigkeit;** Kommunikative Vernunft

Abbildung 4: Einige medien- und KI-ethische Zugänge

Eine ausführliche Diskussion der Ansätze würde den Rahmen sprengen; zudem ist diese Übersicht natürlich keine vollständige Liste. Sie zeigt aber trotz der unterschiedlichen Herkunftsdisziplinen (vor allem Medienethik, KI-Ethik, Ökonomie und Philosophie) wichtige Kernaussagen, die sich auch in der Regulierungsdebatte zur KI wiederfinden, beziehungsweise auch, welche Theorieströnge sich womöglich noch nicht so prominent bemerkbar gemacht haben, aber für die Regulierung fruchtbar gemacht werden könnten.

Tugendethische und teilweise teleologisch ausgerichtete Ansätze (zum Beispiel Spiekermann 2019; Ess 2020 und Vallor 2024) legen einen Fokus auf Wertebewusstsein, das Ziel der „Eudaimonia" (in etwa: „Glückseligkeit durch ethische Vernunft und Ausüben von Tugenden") und verlangen nach einer „Wertepriorisierung" bei der Entwicklung digitaler Technologien. Kantianisch-deontologisch geprägte Zugänge (zum Beispiel Floridi et al. 2018) betonen die Menschenwürde und das Ermöglichen der Selbstrealisierung durch digitale Technik. Einen Schwerpunkt auf die Werte Privatheit, Autonomie und Sicherheit und die Nutzung einer „narrativen" Ethik setzen zum Beispiel Grimm et al. (2019). Coeckelbergh (2020) und Thimm/Bächle (2019) beschäftigen sich unter anderem mit dem Thema Autonomie und betrachten diese als gesellschaftlich und interdisziplinär auszuhandelndes Konstrukt und als Bedingung der Möglichkeit der Übernahme ethischer Verantwortung (was Maschinenethik in einem spezifischen Sinn, nämlich

dass Maschinen „autonom" entscheiden könnten, ausschließt). Einen eher pragmatischen Zugang findet man bei bestimmten Anforderungen an das Design von KI-Tools, etwa an *accountability*, *responsibility* und *transparency* (vgl. Dignum 2019). Viele Richtlinien und Ethik-Kodizes für KI-Entwickler:innen versuchen, solche und weitere Werteanforderungen zu integrieren.

In der KI-Ethik Debatte geht es nicht zuletzt auch um Gerechtigkeitszugänge und den öffentlichen Vernunftgebrauch für eine mögliche gesellschaftliche Akzeptanz neuer KI-Entwicklungen. Dabei wird zwar auf einige diesbezüglichen klassischen Ansätze (vgl. zum Beispiel Habermas 1991; Rawls 2001 und Sen 2010) verwiesen, aber aus Sicht des Autors werden diese Ansätze bisher zu wenig ausgearbeitet (vgl. hierzu Litschka 2025). Mögliche Themenfelder wären hier die Gerechtigkeit des Datenzugangs und der Datennutzung (vgl. Litschka/Saurwein/Pellegrini 2024) sowie Fairnessfragen bei Entscheidungen durch KI-(Sprach-)Modelle, aber auch generell die Rawlsianische Frage nach der gesellschaftlichen Grundstruktur und der Institutionen, die durch die vermehrte KI-Verbreitung und Nutzung verändert werden.

Wie könnte man also künftig diese gerechtigkeitsorientierten Ansätze verwenden, um zum Beispiel gerechte Zugangsmöglichkeiten, Wertediversität, Pluralität und prozedurale Transparenz im KI-gestützten Mediensektor zu sichern / zu erlangen und künftige Regulierungsmaßnahmen noch mehr auf Gerechtigkeitskonzepte zu stützen?

4. Welche KI-Ethik für die Regulierung? Rawls und Sen revisited

Eine naheliegende Frage an dieser Stelle lautet, warum Ethiker:innen bei Fragen der KI-Regulierung und Governance nach einer bestimmten (prozessorientierten) Gerechtigkeitsethik suchen sollten, und nicht mit anderen oben beschriebenen (oder weiteren verfügbaren) Ansätzen auskommen können. Die Antwort darauf ist, dass insbesondere zwei Probleme die Integration normativer Werte in Regulierungsdokumente und -akte kompliziert machen:

– *Das Problem der Werteaggregation*: Aus der klassischen ökonomischen Forschung ist bekannt, dass das Treffen gesamtgesellschaftlicher Entscheidungen aufgrund individueller Präferenzen bestimmten logischen Problemen unterliegt. Schon Arrows Unmöglichkeitstheorem (vgl. Arrow 1951), auf dem auch Ergebnisse anderer *Social Choice*-Theoreti-

ker wie Sen (2017) aufbauen, zeigt, dass es kein komplett konsistentes (Wahl-)System gibt, das wünschenswerte Kriterien wie Nicht-Diktatur, Pareto-Effizienz und Unabhängigkeit von irrelevanten Alternativen gleichzeitig erfüllen kann, wenn die Präferenzen aller Wähler berücksichtigt werden.
– *Das Problem des Wertepluralismus*: Soziale Entscheidungen unterliegen immer auch politischen Problemen und Differenzen, die sich durch unterschiedliche Kulturen, Bildungszugänge und Traditionen ergeben. Ein bekanntes Beispiel hierfür war die kaum aufzulösende Debatte über die widerstreitenden Werte Freiheit und Sicherheit bei der Bekämpfung der *Covid*-Pandemie. Für beide demokratisch akzeptierten und der Gesellschaft wichtigen Werte gab und gibt es gute Argumente; nur beide Werte gleichzeitig zu maximieren, funktioniert selten.

Umgelegt auf unser Problem der Regulierung bedeuten diese Probleme, dass es ohne eine allgemeine Zustimmung zu einem Prozess für Regulierungsvorgaben schwierig werden wird, die unterschiedlichen Präferenzen der Unternehmen, Nutzer:innen und Behörden abzustimmen. Freilich bleiben uns am Ende eines solchen Weges Werteentscheidungen kaum erspart, werden aber womöglich durch einen als gerecht empfundenen Prozess dorthin erleichtert. Gerechtigkeitsorientierte Zugänge versuchen nämlich oft, individuelle Präferenzen und politische Differenzen zu umgehen, indem sie unter anderem fragen, welche *Prozesse* zu einer gesellschaftlich akzeptablen Entwicklung und Nutzung von KI führen und welche *Institutionen* Kooperation innerhalb einer Gesellschaft und ein gemeinsames Gerechtigkeitsverständnis sichern. Zwei bekannte Ansätze hierzu stammen von John Rawls und Amartya Sen.

Rawls' (2001) Theorie der Gerechtigkeit als Fairness und seine bekannten Gerechtigkeitsprinzipien für Hintergrundinstitutionen der Gesellschaft könnte man folgendermaßen für das oben angesprochene Problem anwenden: Da Gleichheit nicht nur formaler Natur sein, sondern auch zu realen, vergleichbaren Erfolgsaussichten führen soll, muss in der KI-Wirtschaft sogenannte *substanzielle* Chancengleichheit hergestellt werden. Dies meint die aktive Beseitigung von Ungleichheiten statt rein prozeduraler Gleichheit (wie sie zum Beispiel vor dem Gesetz herrscht). Man kann dies dadurch begründen, dass algorithmische Diskriminierung einerseits im Gegensatz zum Recht aller auf gleiche Bürgerschaft steht, andererseits eine Überwachung durch KI-Systeme ein „gutes Leben ohne unerwünschte Einflussnahme“ (wie es Rawls ausdrückt) behindern kann. Faire Chancengleichheit

ist eben nicht nur formale Chancengleichheit, sondern Herstellung einer fairen Chancenverteilung (ähnliche Fähigkeiten führen zu ähnlichen Erfolgsaussichten).

Für Medienunternehmen, die mit KI operieren, bedeutet das, dass die Notwendigkeit öffentlicher Rechtfertigung, insbesondere durch öffentliche Deliberation über KI-Modelle, ernster als bislang genommen werden muss. Dies gilt nicht nur für öffentlich organisierte Unternehmen wie öffentlich-rechtliche Anbieter, sondern auch für private (Plattform-) Unternehmen; öffentliche Rechtfertigung sticht private Begründungen, zum Beispiel bezüglich optimaler und effizienter, gewinnorientierter Geschäftsmodelle, aus. Die oft geäußerten Bedenken vieler privat organisierter Medienbetriebe über mögliche Einschränkungen der unternehmerischen Freiheit, die solche Kommunikationsstrategien angeblich mit sich bringen, können so entkräftet werden: Autonomie und Freiheit werden nicht durch Rechtfertigungsprozesse und Regulierungen unterminiert, sondern durch mögliche ungerechte Einflussnahmen einer Technologie wie KI (vgl. Gabriel 2022: 12). Die bei öffentlich-rechtlichen Sendern beispielsweise immer schon stärkere Tradition der Rechtfertigung ihrer Methoden (der Themenauswahl, der Recherche, der verwendeten Technologie, etc.) sollte somit vermehrt in den privaten Sektor Eingang finden, etwas, das wohl nur durch stärkere Regulierung stattfinden kann.

Sowohl die selbstregulierenden *Governance*- und *Accountability*-Maßnahmen der Medienunternehmen selbst, als auch die rechtlichen Vorgaben der Medien- und Technologiepolitik, und nicht zuletzt auch die europäischen Regulierungsrichtlinien bezüglich des Designs neuer KI-Anwendungen sollten also entsprechend angepasst werden. Dies betrifft neben der Notwendigkeit der öffentlichen Deliberation (unabhängig der richtigerweise stattgefundenen Deliberation solcher wegweisender Verordnungen wie dem *AI-Act*) auch die aktive Beseitigung aufgetretener Ungleichheiten; hier könnte man an unterschiedlich vorhandene KI-Kompetenzen des Publikums denken und Maßnahmen, die diese Differenzen ausgleichen.

Sen entwickelte seinen „Capability Approach" weiter zu einem eigenen Gerechtigkeitsansatz („Comparative Justice"; vgl. Sen 2010; Litschka 2015 und Litschka 2019). Im Unterschied zu Rawls verzichtet Sen auf absolute, universelle Gerechtigkeitsprinzipien und betont stattdessen die Bewertung konkreter institutioneller Wirkungen und realer menschlicher Handlungsmöglichkeiten im internationalen Vergleich. Diese realen Handlungsmöglichkeiten (eben „Capabilities") müssen unter anderem von der Politik (mittels verschiedener Konversionsfaktoren) hergestellt werden. Nahelie-

gend in unserem Zusammenhang und direkt anschließend an die empirischen Ergebnisse, die eine Überforderung der Nutzer:innen von KI befürchteten, wäre ein Fokus auf KI-Capabilities der Bürger:innen in der Bildungspolitik.

Betreffend der Medienlandschaft sieht Sen sowieso eine starke Rolle unabhängig operierender Medien(unternehmen). Wie er an einer Stelle (vgl. Sen 2010: 201) formuliert: Global agierende Medien und transnationale Organisationen können die so genannten „Grenzen der Gerechtigkeit" erweitern. Denn der Fokus gerade in technologischen Umbruchszeiten muss immer auch auf Bewohner:innen anderer Länder statt nur eine Nation gelegt werden, womit vermieden werden soll, dass provinzielle (stark von einer Nation oder Kultur geprägten) Werte zu stark in den Vordergrund treten („open" statt „closed impartiality"). Der „Impartial Spectator" dient hier als metaphysisches Modell für eine objektive Beurteilung und die Integration der „distant voices", um eine offene, pluralistische und global legitimierte Wertebasis zu schaffen. So können laut Sen verschiedene und divergierende Gerechtigkeitsprinzipien durch den öffentlichen und unparteilichen Vernunftgebrauch vereint werden.

Konkret bedeutet dies für Medien- und KI-Unternehmen und die zuständigen Regulierungsbehörden, dass vermehrt öffentliche Stakeholder-Dialoge mit Entwickler:innen, Behörden, Unternehmen und User:innen anzustreben sind, um diese Werteintegration zu erreichen; für Regulierungsprinzipien, wie sie grundlegend für Regelwerke wie den *DMA* (*Digital Markets Act*) / *DSA* (*Digital Services Act*) / *AI-Act* und deren Nachfolger sein sollten, sind internationale (-kulturelle) Gesichtspunkte einzubeziehen.

Mit Sen sind auch viele Medienökonom:innen und Medienethiker:innen der Meinung, dass utilitaristische Denkweisen die normativen Probleme der KI-Weiterentwicklungen nicht abdecken können (vgl. beispielsweise Christians 2007; Sen/Williams 1982 und Karmasin/Litschka 2013) und verlangen nach Analysen mittels kommunikativer Rationalität (ähnlich wie Habermas 1991) und deontologischen Konzepten. Ökonomisch rationale Argumente in utilitaristischen Theorien wie dem *Uses-and-Gratifications*-Ansatz (also was genau „nutzt" eine KI-Anwendung spezifischen User:innen?) sind nicht in der Lage, wichtige normative Konzepte wie die Pflichten von Individuen, die Einbettung von Individuen in eine reaktionsfähige Mediengesellschaft oder Massenmedien als soziale Institutionen, die nicht durch isolierte individuelle Entscheidungen verändert werden können, zu erfassen (vgl. Christians 2007). Nur deontologische (und diskursive) Theorien und, wenn man den Argumenten dieses Beitrags folgt, *capability-*

orientierte Denkmodelle können eine solche vollständige normative Sicht vermitteln.

Der *Capability*-Ansatz selbst betont vielerorts die Bedeutung deontologischer Kategorien und von „Prozessen" auf dem Weg zur Erreichung bestimmter Ziele, zwei für den Utilitarismus unwichtige Konzepte. Sen (1985: 4) hat beispielsweise ein prozedurales Verständnis der Bedeutung von Märkten in einer Gesellschaft und Rechten als unveräußerlichen Aspekten einer Person. Was den Markt für Künstliche Intelligenz betrifft, so findet sich in einem Großteil der wirtschaftswissenschaftlichen Literatur, aber auch in Äußerungen von *big-tech-Managern* der Branche (anschaulich dazu Vallor 2024) utilitaristisches Denken und ein starker Glaube an das Funktionieren von Märkten oder Marktplätzen für Ideen (Karmasin/Litschka 2013). Während diese Theorien aus der Perspektive der Betonung des innovativen Potenzials funktionierender Technologiemärkte und der Abschreckung von zu viel staatlichem Einfluss vernünftig sind, scheinen sie nicht in der Lage zu sein, die vielschichtigen Dilemmata zu bewältigen, mit denen sich Organisationen und Bürger:innen bei der aktuellen Nutzung von KI konfrontiert sehen. An anderer Stelle haben der Autor und Kollegen beispielsweise beschrieben, welche Probleme die Marktkonzentration auf Plattformmärkten und in der Social-Media-Industrie verursachen kann (siehe zum Beispiel Litschka et al. 2024). Eine utilitaristische Ethik scheint nicht dazu beizutragen, diese Probleme zu lindern.

Deontologische Ansätze betonen neben der Bedeutung unveräußerlicher Rechte und der Rolle autonomer und universalisierbarer Entscheidungen, bei denen neben den erwarteten Ergebnissen von Interaktionen auch Verfahren und Prozesse berücksichtigt werden müssen, vor allem eine bestimmte Art von Rationalität. Öffentliche Argumentation – unter anderem ermöglicht durch ein funktionierendes System globaler Medien – ist nicht nur der Grundpfeiler der Demokratie, sondern auch eines möglichen universellen Gesellschaftsvertrags. Für die künftige Entwicklung der KI-Ethik ist es wichtig, diese Art von Vernunft einzubeziehen: Sie verlangt, dass wir unsere Argumente vor allen anderen rechtfertigen können. Wenn wir unterschiedliche Kulturen und Traditionen in die KI-Entwicklung und -Nutzung einbeziehen wollen (siehe auch Ess 2020), um zumindest teilweise universalisierbare Vereinbarungen über KI-Regelungen zu erreichen, ist dieses diskursethische Prinzip weiterhin gültig.

Während also Rawls insbesondere auf substanzielle Gleichheit im Zugang zu Chancen pocht und eine öffentlich nachvollziehbare Legitimation von KI-basierten Entscheidungen fordern würde, betont Sen den Pluralis-

mus und die Notwendigkeit internationaler Vergleichsperspektiven. Beide Ansätze bieten somit komplementäre Orientierungspunkte für eine ethisch begründete und gesellschaftlich akzeptierte Regulierung von KI-Technologien: Rawls als Grundlage für interne Chancengleichheit und institutionelle Rechenschaftspflicht, Sen als Impuls für transnationale Offenheit, Partizipation und Diversität der Wertvorstellungen. In der Medienpraxis empfiehlt sich eine verbindende Perspektive, welche substanzielle Gleichheit und globale Pluralität abbildet.

5. Zusammenfassung und Ausblick

Die empirischen Erkenntnisse der durchgeführten Studie bestätigen eine kritische Selbstsicht der Akteur:innen im Hinblick auf die Integration von KI-Technologien im Medienbereich. Im Medienproduktionsalltag haben sich schon an vielen Stellen effizienzsteigernde und innovationsfördernde Prozesse durch eine Anwendung von KI gezeigt. Auf der anderen Seite werden die gesellschaftlichen und ethischen Herausforderungen, die durch Automatisierung, Datenmonopolisierung und algorithmische Steuerung entstehen, kritisch gesehen. Während die Medien- und KI-Ethik schon vielerorts Probleme aufgezeigt und Lösungswege vorgeschlagen hat, sind manche gerechtigkeitsbasierten Theorieimpulse, wie zum Beispiel jene von Rawls (substanzielle Fairness und öffentliche Rechtfertigung) und Sen (vergleichende Gerechtigkeit und Pluralität), bislang noch weniger in politische Maßnahmen und Regulierungen eingeflossen. Der Beitrag zeigt einige erste Möglichkeiten auf, diese Impulse für die Theorieentwicklung einerseits und die praktische Implementierung in Regulierungsvorhaben andererseits aufzunehmen. Unter anderem wären demnach in dieser Sichtweise für eine menschengerechte und gesellschaftlich akzeptable Entwicklung von KI im Medienbereich Prozesse und Institutionen notwendig, die pluralistische Prinzipien offen einbeziehen und substanzielle Chancengleichheit gewährleisten.

Literatur

Acemoglu, Daron (2024): The Simple Macroeconomics of AI, in: Economic Policy, 5. April 2024, (online unter: https://economics.mit.edu/sites/default/files/2024-04/The%20Simple%20Macroeconomics%20of%20AI.pdf – letzter Zugriff: 17.11.2025).

*Arrow, Kenneth Joseph (*1951): Social Choice and Individual Values (= Cowles Commission for Research in Economics, Monograph No. 12), New York.

Belinskaya, Yulia et al. (Hg.) (2024): KI in der Medienwirtschaft, RTR Studienreihe zu Künstlicher Intelligenz, Wien.

Brynjolfsson, E. et al. (2025): Generative AI at Work, in: The Quarterly Journal of Economics 140 (2/2025), S. 889–947.

Christians, Clifford G. (2007): Utilitarianism in media ethics and its discontents, in: Journal of Mass Media Ethics 22 (2–3/2007), S. 113–131.

Coeckelbergh, Marc (2020): AI Ethics, Cambridge.

Dignum, Virginia (2019): Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way, Cham.

Ess, Charles (2020): Digital Media Ethics, 3. Aufl., Cambridge.

Floridi, Luciano et al. (2018): An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, in: Minds and Machines 28 (4/2018), S. 689–707.

Gabriel, Iason (2022): Towards a Theory of Justice for Artificial Intelligence, in: Daedalus 151 (2/2022), S. 1–12.

Grimm, Petra et al. (Hg.) (2019): Digitale Ethik. Leben in vernetzten Welten, Paderborn.

Habermas, Jürgen (1991): Erläuterungen zur Diskursethik, Frankfurt am Main.

Karmasin, Matthias / Litschka, Michael (2013): Normativität in der Medienökonomie, in: Matthias Karmasin / Matthias Rath / Barbara Thomaß (Hg.), Normativität in der Kommunikationswissenschaft, Berlin, S. 191–207.

Litschka, Michael (2015): Medien-Capabilities als polit-ökonomisches Konzept in der Medienethik: Theoretische Grundlagen und mögliche Anwendungen, in: Communicatio Socialis 48 (2/2015), S. 190–201.

Litschka, Michael (2019): The Political Economy of Media Capabilities: The Capability Approach in Media Policy, in: Journal of Information Policy 9 (63/2019), S. 79–110.

Litschka, Michael (2025): AI Ethics and the Capability Approach, in: Genealogy+Critique 11 (1/2025), S. 1–16.

Litschka, Michael / Saurwein, Florian / Pellegrini, Tassilo (2024): Open Data Governance und digitale Plattformen, Ethische, ökonomische und regulatorische Herausforderungen und Perspektiven, Wiesbaden.

Pinzolits, Robert et al. (2025): KI im Mediensektor: Eine SWOT-Analyse entlang der Wertschöpfungskette und ihre regulatorischen Implikationen, in: Rimscha, M. Bjørn von / Ehrlich, Gianna / Riemann, Robin (Hg.), Alles rational? Der menschliche Faktor in Medienorganisationen: Proceedings zur Jahrestagung der Fachgruppe Medienökonomie der DGPuK 2024, Mainz, S. 106–120.

Pinzolits, Robert et al. (2025): KI und Medienvertrauen. Mediennutzung und Medienvertrauen in Österreich im Spannungsfeld von KI und Sozialen Medien. RTR-Studienreihe: Künstliche Intelligenz in der Medienwirtschaft, 14. Oktober 2025 (online unter: https://www.rtr.at/medien/aktuelles/publikationen/Publikationen/Publikationen_2025/Studie_KI_und_Medienvertrauen.de.html – letzter Zugriff 17.11.2025).

Rawls, John (2001): Justice as Fairness: A Restatement, Harvard.

Sen, Amartya (2010): The Idea of Justice, London.

Sen, Amartya (2017): Collective Choice and Social Welfare, New York.

Sen, Amartya / Williams Bernard (Hg.) (1982): Utilitarianism and Beyond, Cambridge.

Spiekermann, Sarah (2019): Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert, München.

Thimm, Caja / Bächle, Thomas C. (2019): Autonomie der Technologie und autonome Systeme als ethische Herausforderung, in: Matthias Rath / Friedrich Krotz / Matthias Karmasin (Hg.), Maschinenethik. Normative Grenzen autonomer Systeme, Wiesbaden, S. 73–87.

Trabelsi, Mohamed Ali (2024): The impact of artificial intelligence on economic development, in: Journal of Electronic Business & Digital Economics 3 (2/2024), S. 142–155.

Vallor, Shannon (2024): The AI Mirror, Oxford.

„Artificial Sepsis“ – Leitlinien einer salutogenetischen Bot-Nutzung

Matthias O. Rath

Zusammenfassung

Der Beitrag beleuchtet die systemischen Risiken generativer KI-Systeme am Beispiel *Großer Sprachmodelle* wie ChatGPT. Er führt die Metapher *Artificial Sepsis* ein, um epistemische Selbstvergiftung durch rückgekoppelte KI-Inhalte zu beschreiben, und entwickelt ein Modell salutogenetischer Bot-Nutzung. Ethische Überlegungen zur viralen Verbreitung und zu den globalen Chancen und Herausforderungen generativer KI ergänzen die Beobachtung einer algorithmisch induzierten Selbstreferenz. Ziel des Beitrags ist eine paradigmatische Verschiebung der gängigen Bewertung des internetbasierten KI-Einsatzes hin zu epistemischer Resilienz, digitaler Kohärenz und verantwortungsgeleiteter Technikgestaltung.

1. Einleitung

Die rapide Ausbreitung generativer Künstlicher Intelligenz (KI) in Form Großer Sprachmodelle (*Large Language Models*, LLMs) wie ChatGPT transformiert gegenwärtig Kommunikationspraktiken, Informationsräume und das Verständnis von Autor:innenschaft. Basierend auf dem Konzept der „Transformer architecture” von Vaswani et al. (2017) wurden „neural networks“ mit riesigen Textkorpora darauf trainiert, jedes Wort oder Subwort (ein *token*, vgl. Vaswani et al. 2017) im Kontext aller anderen *tokens* zu verarbeiten. LLMs entwickeln dadurch zwar unerwartete Fähigkeiten, erzeugen aber auch Inkonsistenzen, die durch Korrekturen in den Trainingsdaten und im ‚Prompting‘, den Anweisungen (*prompts*) an die KI, optimiert werden müssen (vgl. Zhao et al. 2021). So zeigen Weidinger und andere (2021), dass Daten- und Prompting-abhängige Inkonsistenzen von LLMs bei ungeübter und unkontrollierter Nutzung gravierende methodische Schwächen mit medienethischen Folgen und erheblichen gesellschaftlichen Risiken generieren, etwa datengestützt die Verstärkung von Vorurteilen

und Stereotypen (*bias*), Desinformation und Vertrauensprobleme sowie in Bezug auf das Prompting die Problematik fehlender Transparenz (Opazität, vgl. Zhao et al. 2024).

Galten diese Inkonsistenzen und ihre Folgen noch bis vor kurzem als dystopische Vision, so sind diese inzwischen als Herausforderungen in alltägliche Schreib-, Forschungs- und Bildungsprozesse eingedrungen. Die öffentliche und wissenschaftliche Auseinandersetzung kreist dabei vorrangig um Fragen der Autor:innenschaft, der Täuschungspotenziale und der juristischen Verantwortbarkeit KI-generierter Texte (vgl. Dwivedi et al. 2023). Doch diese Fokussierungen greifen zu kurz. Denn das Risiko besteht nicht allein in Fälschungen oder Plagiaten, sondern in einer tiefgreifenden *epistemischen Transformation*, die sich aus selbstreferenziellen Rückkopplungen generativer Systeme speist. Welche Qualität haben Erkenntnisprozesse, an denen KI beteiligt ist? Welche Strukturen entstehen, wenn maschinell erzeugte Texte zum Ausgangspunkt weiterer maschineller Generierung werden? Dieser Beitrag greift bewusst den medizinischen Begriff der *Sepsis* metaphorisch auf, um auf diese Problematik aufmerksam zu machen.

Theoretisch wird der Beitrag von drei zentralen Perspektiven getragen:

1. Methodisch vom Konzept der *Futures Literacy* (vgl. Millers 2011; UNESCO 2018) als Reflexionsrahmen für antizipative Ethik,
2. inhaltlich von der Diskurs- und Kommunikationsethik von Jürgen Habermas als normativem Anspruch an KI-gestützte Kommunikation beziehungsweise KI-basierte Textgenerierung sowie
3. metapherntheoretisch von der *Conceptual Metaphor Theory* (vgl. Lakoff 1993) und der *Deliberate Metaphor Theory* (vgl. Steen 2023) als ethisch-epistemisches Gegenmodell zu dysfunktionalen Informationsökologien.

2. „Sepsis" als Metapher

Digitale Fehlfunktionen sind anhand *medizinischer* Metaphern besonders anschaulich beschreibbar. Ausdrücke wie ‚Virus' werden bereits seit den 1980er Jahren zur Beschreibung softwarebasierter Angriffe genutzt (vgl. Cohen 1987). Die hier eingeführte Metapher der ‚Artificial Sepsis' verweist jedoch auf eine qualitativ neue Störungsebene: die Selbstvergiftung eines informationsverarbeitenden Systems durch die zirkuläre Rezeption KI-generierter Inhalte.

Metaphern können als „*Sprachspiele*" (Wittgenstein 1999: 241) angesehen werden, die einer bestimmten abstrakten Denkstruktur (vgl. Bertau 1996)

entsprechen und Wörter eines „Ausgangsbereichs" mit ihrer Bedeutung in einen „Zielbereich" (Schmale 2019: 5) übertragen. Für die Philosophie und in der Folge die Theologie ist „Krankheit als Metapher" (Bendemann 2022: 88) schon seit der Antike konzeptionell geübt, um „Störungen" und ihre Überwindung strukturell nachvollziehbar zu machen.

Moderne Metapherntheorien (vgl. Semino/Demjén 2016), vor allem aus dem Bereich der kognitiven Linguistik (vgl. Geeraerts 2006; Putterer 2022), unterscheiden mehrere Metapherntypen, unter anderen „Strukturmetaphern". Lakoff und Johnson (2003: 22) bezeichnen damit in ihrer *Conceptual Metaphor Theory* (vgl. Lakoff 1993) jene „Fälle, in denen ein Konzept von einem anderen Konzept her metaphorisch strukturiert wird". Darüber hinaus gehend stellen die hier genutzten Metaphern jedoch nicht nur rhetorische Figuren oder konzeptionelle Übertragungen dar, sondern sie dienen im Sinne der *Deliberate Metaphor Theory* (vgl. Steen 2023) einem bewusst herbeigeführten Perspektivwechsel: „I propose that a metaphor is used deliberately when it is expressly meant to change the addressee's perspective on the referent or topic that is the target of the metaphor, by making the addressee look at it from a different conceptual domain or space, which functions as a conceptual source" (Steen 2008: 222).

Der medizinische Ausdruck „Sepsis" ist definiert als „life-threatening organ dysfunction resulting from a dysregulated host response to infection" (Singer et al. 2016), also als eine systemische Entgleisung der Immunantwort, bei der sich ein Organismus in der Reaktion auf eine Infektion selbst schädigt. Im Folgenden wird diese Fehlreaktion des Körpers, die das eigene Gewebe und die eigenen Organe bekämpft, als *fachsprachliche Metapher* (vgl. Schmale 2019) auf KI-generierte Kommunikation übertragen und als *Artificial Sepsis* (AS) spezifiziert. AS bezeichnet also eine Struktur, in der sich digitale Systeme gegen ihre eigenen Grundlagen richten, indem sie Inhalte nicht mehr aus menschlicher Erfahrung, dialogischer Interaktion oder sozial validierten Quellen schöpfen, sondern zunehmend aus eigenen Outputs – mit der Folge einer systemischen Selbstvergiftung.

3. Digitale Autointoxikation: Zur Diagnose der AS

In der Informatikgeschichte haben sich metaphorisierende Konzepte etabliert, um softwaretechnische oder netzwerkbezogene Fehlfunktionen zu beschreiben. Begriffsbildend wurde vor allem der Ausdruck *bug* als Be-

zeichnung des angeblich ersten Computer-Hardware-Fehlers durch eine Motte am 9. September 1947 in einem Relais eines *Harvard Mark II* Rechners (vgl. Lunduke 2022). Ab den 1980er Jahren sind vor allem metaphorische Ausdrücke wie „Virus" (Cohen 1987) oder „viral attack" (Ross 1990) geläufig. Gemeinsam ist diesen Metaphern die Vorstellung eines *externen* Eingriffs: Fehlfunktionen sind, ähnlich wie im bildspendenden Ausgangsbereich der hier verwendeten Metapher, der Medizin, von außen verursacht, durch externe Faktoren wie ‚Viren' oder ähnliches.

Die hier eingeführte Metapher ‚AS' will jedoch das Augenmerk auf den gegenteiligen Sachverhalt richten. Der problematische Faktor kommt nicht von außen, also als Eindringling, sondern von innen, nämlich durch die Präsenz KI-generierter Texte (im weitesten Sinne) im Netz. Nochmals mit dem fachsprachlichen Ausgangsbereich unserer Metapher AS, der Medizin, gesprochen: Der medizinische Normalfall ist der Einbruch des Fremden durch die Infektion, ein Keim, eine Mikrobe, ein Virus befällt einen Wirtsorganismus. Der medizinische Notfall hingegen tritt auf, wenn sich die Immunantwort des Körpers gegen sich selbst wendet und der Körper sich selbst vergiftet.

Ziel dieses Beitrags ist, die metaphorisch eingeführte *AS* als Fehlorientierung des digitalen Systems nach „innen" zu konstatieren und dann, wieder metaphorisch gesprochen, nach *Heilung* für diese Selbstvergiftung zu suchen. Denn die Gefahr digitaler Desinformation – so die These dieses Beitrags – liegt weniger in externen Aktoren, sondern in strukturell bedingten *Selbstreferenzdynamiken*. Diese Dynamiken lassen sich als Ausdruck einer ‚epistemischen Pathologie' verstehen – einer Störung des digitalen Erkenntnissystems, bei der sich Informationen zunehmend von ihren normativen und erfahrungsbezogenen Quellen lösen. Der metaphorisch aufgegriffene Begriff der Sepsis verweist in seinen „ko(n)textuellen und semantischen Informationen" (Schmale 2019: 9), die er definitorisch zur Verfügung stellt, auf eine *systemische Dysfunktion*: Unter den Bedingungen artifizieller Intelligenz, wie sie zum Beispiel ChatGPT zur Verfügung stellt, wird Wissen nicht mehr produziert, sondern rekombiniert; Urheberschaft wird entgrenzt statt verortet; und die Verifikation von Information wird ersetzt durch statistische Plausibilität (vgl. Rath 2018).

Bis Ende 2024 lernten LLMs wie GPT-3.5 und GPT-4 anhand riesiger Textmengen, darunter auch Daten, die maschinell erzeugt wurden. Bender und andere (2021) haben gezeigt, dass diese Modelle beginnen, ihre eigenen Outputs als Trainingsgrundlage zu nutzen – etwa durch Webcrawling öffentlich verfügbarer Inhalte, die selbst wieder von generativen Syste-

men stammen. Zusätzliche Dynamik erhält diese Entwicklung durch die *ChatGPT Search*-Funktion, die seit Dezember 2024 (OpenAI 2024) allen Usern zur Verfügung steht – das *ChatGPT 4 Model* und höher sucht nun auch im Internet. Dieser Rückkopplungseffekt bedroht die epistemische Diversität: Je mehr maschinell generierte Inhalte zirkulieren und rezykliert werden, desto stärker verengt sich der semantische Horizont, auf dem neue Inhalte beruhen. Shumailov und andere (2023: 2) beschreiben diesen Prozess daher als „model collapse", der zu einem algorithmisch induzierten „data poisoning" (Shumailov et al. 2023: 3) führt.

Konnte dieser Kollaps 2023 noch als nur drohend angesehen werden (vgl. Rath 2024), so zeigt der aktuelle *Bad Bot Report* (Imperva 2025), dass dieser Kollaps faktisch schon gegeben ist. Bereits 51 Prozent aller Webaktivitäten stammen von automatisierten Bots, die Inhalte generieren, die ihrerseits wiederum indexiert und von KI-Systemen als vermeintlich legitime Quellen genutzt werden. Besonders problematisch ist dies bei KI-Systemen, die für medizinische Beratung, psychologische Unterstützung oder juristische Einschätzungen eingesetzt werden – Bereiche, in denen semantische Präzision, Kontextsensitivität und ethische Verantwortung unabdingbar sind.

Damit entsteht durch den Rückgriff auf quasi unbegrenzt zur Verfügung stehende, artifiziell erzeugte Inhalte eine Art *ontologischer Kurzschluss*: Die LLMs generieren sprachliche Artefakte, die sich im nächsten Zyklus als ‚Realität' ausgeben – „models do not forget previously learned data, but rather start misinterpreting what they believe to be real, by reinforcing their own beliefs" (Shumailov et al. 2023: 3). In der Folge verschwimmen die Grenzen zwischen Simulation und Bezugnahme. Die Gefahr besteht dann nicht nur darin, Fehler zu übernehmen, sondern in der schleichenden Entwertung epistemischer, weil semantischer Standards. Was als „Content" zirkuliert, verliert seine Verankerung in realweltlicher Erfahrung, in sozialen Diskursen, in verantwortlicher Urheberschaft. AS beschreibt diesen Kollaps – eine Entkopplung digitaler Inhalte von ihren erkenntnisstiftenden Quellen. Das Konstrukt der AS fordert daher eine ethische Antwort, die über technische Korrekturmechanismen hinausweist: eine neue Kultur zukunftszugewandter *epistemischer Verantwortung*. Methodisch verweist dies auf die Erträge einer *Futures Literacy*.

4. Futures Literacy: Antizipieren, um zu verantworten

Das Konzept der *Futures Literacy* (vgl. Miller 2011; UNESCO 2018) wurde entwickelt, um die Fähigkeit zur aktiven Gestaltung der Zukunft durch kritische Reflexion zu fördern. Es geht nicht um deterministische Vorhersagen, sondern um die imaginative Erschließung möglicher, plausibler und wünschbarer Zukünfte. In Zeiten zunehmender Unsicherheit und technologischer Beschleunigung eröffnet *Futures Literacy* einen Reflexionsraum, der das ethische Potenzial antizipativer Urteilskraft betont.

Futures Literacy ist damit mehr als ein Planungstool. Sie ist eine epistemische Kompetenz – eine Kulturtechnik, mit der Gegenwart durch Zukunft verstehbar gemacht wird. Riel Miller unterscheidet in diesem Zusammenhang „‚used' futures" – implizite, unreflektierte Zukunftsbilder – und „open futures" – bewusste, plurale und veränderliche Imaginationen (vgl. UNESCO 2018: 54, 163). Der ethische Imperativ besteht darin, Zukunft nicht bloß zu konsumieren, sondern verantwortlich zu entwerfen.

Ein praktisches Beispiel ist der Aufbau sogenannter *sandboxes*, kontrollierte Umgebungen, in denen KI-Systeme unter Beobachtung getestet werden, bevor sie in reale Kontexte eingeführt werden (vgl. OECD 2023). Sie dienen nicht nur der technischen Optimierung und der prospektiven Abklärung von Haftungsfragen (Truby et al. 2022), sondern auch der ethischen Bewertung (vgl. Undheim et al. 2023): Wie reagiert ein KI-System auf Ambiguität? Welche Vorannahmen sind in die Trainingsdaten eingeschrieben? Wie können Diskriminierungen, Verzerrungen oder epistemische Ausschlüsse erkannt und adressiert werden?

Auch der Aufbau von Delphi-basierten Dialogforen in Bildung, Wissenschaft und Verwaltung kann als *futures*-literates Handeln verstanden werden. Diese Verfahren integrieren Expertenwissen, gesellschaftliche Erwartungen und normatives Urteilsvermögen in eine strukturierte Reflexion über die Zukunft. Dabei zeigt sich ein Spannungsverhältnis zwischen antizipativer Verantwortung und demokratischer Legitimität: Wer bestimmt, welche Zukunftsoptionen als wünschbar gelten? Welche Akteure sind sichtbar, welche marginalisiert? *Futures Literacy* muss deshalb mit dem Anspruch verbunden sein, partizipativ und dekolonial (vgl. Bourgeois et al. 2024) sowie reflexiv (vgl. Mangnus 2021) zu operieren.

In ethischer Perspektive lassen sich Parallelen zur Verantwortungsethik von Hans Jonas ziehen: Zukunft ist nicht nur Möglichkeitsraum, sondern auch Verpflichtungsraum. Die Prognose des Möglichen, vor allem des *worst case*, wird als „*Heuristik* der Furcht" (Jonas 1979: 392; vgl. Rath 1988) zur

Begründung gegenwärtiger Verantwortung. Ethische Orientierung muss in der digitalen Transformation antizipativ erfolgen – als heuristische Rahmung des noch nicht Wirklichen. *Futures Literacy* ist somit kein methodischer Luxus, sondern eine notwendige Voraussetzung dafür, dass KI-Entwicklung nicht blind, sondern verantwortet verläuft. Sie schafft jene semantische Tiefenschärfe, die notwendig ist, um AS zu vermeiden: durch präventive, plurale und reflektierte Zukunftsdeutungen.

5. Kommunikationsethik: Geltungsansprüche und epistemische Verantwortung

Jürgen Habermas formulierte in seiner *Theorie des kommunikativen Handelns* vier Geltungsansprüche, die in jedem Verständigungsakt implizit präsent sind – sie strukturieren die Möglichkeit legitimer Kommunikation (vgl. Habermas 1981: 439). Wir rekurrieren hier auf drei davon: *Wahrheit* – die Aussage soll inhaltlich zutreffen, *Wahrhaftigkeit* – die Sprechende sollen subjektiv ehrlich sein, und *Richtigkeit* – das Gesagte soll in einem sozialen Kontext als normativ akzeptabel gelten.

Im Kontext generativer KI stellen sich diese Geltungsansprüche neu: Kann ein KI-System „wahr" sprechen, ohne Erfahrung? Kann es „wahrhaftig" sein, ohne subjektives Bewusstsein? Kann es „richtig" kommunizieren, ohne normative Einbettung? Die Antwort lautet: nur simulativ. Generative Sprachmodelle wie ChatGPT produzieren Aussagen, die so wirken, als seien sie intentional, obwohl sie es nicht sind. Ihre Kommunikation ist performativ, aber nicht intentional.

Gunkel und Bryson (2014) schlagen vor, diese Differenz ernst zu nehmen. In ihrer Theorie der Maschinenmoral (*Machine Morality*) unterscheiden sie zwischen *moral agents* (Handlungsträgern mit Verantwortung) und *moral patients* (Wesen mit moralischem Anspruch auf Schutz). Generative KI-Systeme passen in keine dieser Kategorien vollständig. KI-Systeme erfordern eine neue Ethik relationaler Verantwortung: nicht, weil sie Subjekte sind, sondern weil Menschen mit ihnen interagieren, als wären sie es (vgl. für den Bildungsbereich Rath [im Druck]). Diese *as-if-Kommunikation* hat praktische Konsequenzen: Viele Nutzer:innen entwickeln eine anthropomorphe Beziehung zu KI-Systemen – sie erwarten Relevanz, Konsistenz, sogar Empathie. Diese Erwartungen treffen jedoch im Moment noch auf Systeme, die über keine Weltbindung verfügen. Daraus entsteht eine kommunikative Dissonanz, die Habermas' Modell nicht auflöst, aber normativ

rahmt: Wenn Geltungsansprüche nicht erfüllbar sind, müssen sie kenntlich gemacht werden.

Daraus ergibt sich ein *Design-Imperativ*, KI-Systeme so zu gestalten, dass ihre epistemische Statuslage transparent bleibt. Epistemische Label, die eine schlussfolgernd vermeintliche oder naheliegende Notwendigkeit markieren (vgl. Moon et al. 2016), aber auch Antwort-Quellenangaben, Unsicherheitsindikatoren oder stilistische Markierungen können dazu beitragen, die Differenz zwischen Mensch und Maschine kommunikationsethisch produktiv zu machen.

Habermas' Theorie bleibt somit normativer Maßstab in einer Welt, in der Kommunikation nicht mehr ausschließlich zwischen Menschen stattfindet, um kommunikative Rationalität auch unter Bedingungen algorithmischer Produktion aufrechtzuerhalten.

6. Informationsethik und digitale Kompetenz

Die ethischen Herausforderungen generativer KI lassen sich aber nicht allein durch technische Standards oder gesetzliche Regularien bewältigen. Mindestens ebenso entscheidend ist die Kompetenz der Nutzer:innen im Umgang mit digitalen Systemen. Die Diskussion um „Digital Literacy" (Buckingham 2015) hat sich in den vergangenen Jahren zum Konzept einer „Critical GenAI Literacy" (Rapanta et al. 2025) weiterentwickelt, das neben funktionalem Wissen auch ethische (vgl. Capurro 2010; Floridi 2013), reflexive und pädagogisch-didaktische Dimensionen umfasst.

Floridi (2013: 102–133) betont in diesem Zusammenhang die epistemische Verantwortung für die „infosphere" mediatisierter Gesellschaften (vgl. Kalina et al. 2018): Wo die Unterscheidung zwischen Information, Kommunikation und Handlung verschwimmt, tragen Menschen Verantwortung nicht nur für ihre Daten, sondern auch für ihre epistemischen Positionierungen. „Critical GenAI Literacy" wird damit zu einer Disziplin der Lebensführung, einer Frage der Urteilskraft in einem von Maschinen mitgestalteten Bedeutungsraum.

Beispiel Hochschullehre: Viele Studierende nutzen ChatGPT zur Ideenfindung, zum Formulieren wissenschaftlicher Texte oder zur Simulation von Prüfungsgesprächen. Lehrende stehen vor der Herausforderung, diese Nutzung nicht pauschal zu verbieten, sondern kompetenzorientiert zu begleiten – als Differenzierung zwischen zulässiger Unterstützung und un-

zulässiger Substitution, Förderung reflexiver Textarbeit und Stärkung der Fähigkeit zur Quellenkritik.

Beispiel Medienproduktion: Redaktionen stehen unter Produktionsdruck, KI-Systeme bieten eine scheinbar effiziente Lösung. Doch der Einsatz von automatisierten Textbausteinen, Überschriften oder ganzen Artikeln ohne menschliche Gegenprüfung führt zur Erosion journalistischer Qualitätsstandards. Informationsethik fordert hier institutionelle Sicherung epistemischer Sorgfalt – durch Redaktionsrichtlinien, Faktenchecks, Transparenzpflichten.

Beispiel Verwaltung: Wie geht man mit von KI generierten Verwaltungstexten um? Wer trägt Verantwortung für fehlerhafte Bescheide, diskriminierende Formulierungen oder unzulässige Schlussfolgerungen? Informationsethik bedeutet in diesem Fall: Einführung von Prüfungsschleifen, Dokumentation der Systemverwendung, Schulung des Verwaltungspersonals.

Auf einer übergreifenden Ebene kann man drei Kompetenzdimensionen unterscheiden:

- Technische Kompetenz: Wissen über Funktionsweise, Stärken und Schwächen generativer KI.
- Reflexive Kompetenz: Fähigkeit zur Kontextualisierung, zur Einschätzung von Risiken und zum kritischen Umgang mit KI-generierten Inhalten.
- Normative Kompetenz: Fähigkeit, Entscheidungen auf der Basis von Werten, Rechten und sozialen Folgen zu treffen.

Diese Kompetenzen bilden die Grundlage für *epistemische Resilienz*, die Fähigkeit, auch unter Bedingungen digitaler Ambiguität handlungsfähig zu bleiben. Sie ist das Gegenmodell zur AS: nicht immunisierend, sondern befähigend; nicht abschottend, sondern reflektierend. Und diese Resilienz ist notwendig. Waren 2023 starke KI-Eingriffe in die Infosphäre noch nur Befürchtungen – „ein fauler Apfel verdirbt das ganze Fass" war damals mein Bild (vgl. Rath 2024) –, so sind seit 2024 diese Befürchtungen Realität geworden: Der bereits erwähnte *Bad Bot Report* (Imperva 2025: 2) warnt, dass 37 Prozent des bot-basierten *Traffic* intransparent und schädlich ist, eben durch *bad bots* vollzogen wird. Konkret heißt das z. B., dass folgende Branchen am stärksten von bot-basierten Angriffen auf Accounts, sogenannte *takeovers*, betroffen waren: der Finanzdienstleistungssektor mit 22 Prozent aller Angriffe, gefolgt von Telekommunikation und Internetdienstanbietern mit 18 Prozent sowie Computer- und IT-Branche mit 17 Prozent (vgl. Imperva 2025: 19); und ein Ende ist nicht abzusehen. Das *Copenhagen*

Institute for Future Studies (CIFS 2023) prognostiziert, dass bis 2030 das *Metaverse*, also die Verkopplung von virtueller, erweiterter und physischer Realität, zum größten Teil, wohl über 99 Prozent (vgl. Hvitved 2022), KI-generiert sein wird. Wir werden unser altes Internet nicht wiedererkennen. Dieser Prognose des CIFS muss durch eine proaktive Fragestellung begegnet werden, ganz im Sinne der *Futures Literacy*: Was können und müssten wir tun, um die Erfüllung dieser Erwartung zu vermeiden?

7. Salutogenese als ethisches Gegenmodell

Als ein Gegenmodell zu einer resignativen Haltung oder gar Panik in Bezug auf die KI-Zukunft der Infosphäre wird im Folgenden auf den *salutogenetischen Ansatz* verwiesen. Dies Ansatz wurde ursprünglich von Aaron Antonovsky (1979, 1987) im medizinischen Kontext entwickelt und entspringt damit ebenfalls unserem oben eingeführten fachsprachlichen „Ausgangsbereich" (Schmale 2019: 5) der Metapher AS.

Der Ansatz der Salutogenese soll erklären, warum manche Menschen trotz widriger Lebensbedingungen gesund bleiben. Anstelle der Pathogenese – der Frage nach den Ursachen von Krankheit – fragt Antonovsky: Was hält Menschen gesund? Diese Perspektive lässt sich auf mediale und digitale Kontexte übertragen: Was hält uns *epistemisch* gesund in einer Welt, in der KI unsere Informationsökologie verändert (vgl. Ridder 2024)?

Antonovsky identifiziert drei Schlüsselfaktoren eines sogenannten Kohärenzsinns (*sense of coherence*), der maßgebend sei für die Gesundheit. Es ist ein „durchdringendes, andauerndes und dennoch dynamisches Gefühl des Vertrauens" (Antonovsky 1997: 36) in die grundsätzliche *Verstehbarkeit* (*comprehensibility*), *Handhabbarkeit* (*manageability*) und *Sinnhaftigkeit* (*meaningfulness*) des Lebens (vgl. Eriksson/Lindström 2005). Diese Dimensionen können als ethisch-epistemischer Kompass für den Umgang mit generativer KI fungieren:

- Verstehbarkeit: Digitale Inhalte müssen nachvollziehbar sein. Nutzer*innen können erkennen, wann sie mit KI-generierten Texten konfrontiert sind, wie diese zustande gekommen sind und welche Quellen zugrunde liegen. Transparente Kommunikation ist eine Grundvoraussetzung epistemischer Kohärenz.
- Handhabbarkeit: Der Umgang mit KI darf nicht zu Überforderung führen. Das bedeutet, Werkzeugen, Kriterien und Unterstützungsangeboten zur Bewertung digitaler Inhalte bereitzustellen. In der Bildung heißt das,

dass KI-gestützte Lernprozesse so gestaltet werden, dass Schüler:innen „agency" (Mick 2021) behalten – statt sich der Maschine zu unterwerfen.
- Sinnhaftigkeit: Kommunikation muss in einem Wertekontext verankert sein. Informationen erhalten nicht allein durch ihren Wahrheitsgehalt Relevanz, sondern durch ihre Bedeutung für das Leben, für Teilhabe, für Demokratie. KI-Systeme sollen so konzipiert sein, dass sie nicht bloß kognitive Entlastung, sondern existenzielle Orientierung ermöglichen.

Ein salutogenetisches Ethikmodell der KI-Nutzung bedeutet daher keine technikzentrierte Kontrolle, sondern eine lebensweltorientierte Ermöglichung. In der Hochschuldidaktik etwa könnten KI-Systeme eingesetzt werden, um explorative Lernpfade zu eröffnen – ohne die epistemische Eigenleistung zu entwerten. In der Sozialen Arbeit könnten Bots reflexive Dialoge anregen, um Handlungssicherheit zu fördern – nicht durch Belehrung, sondern durch Co-Konstruktion (vgl. Victor/Goldkind 2025).

Zugleich verweist der Salutogenese-Ansatz auf *epistemische Resilienz*: die Fähigkeit, Widersprüche auszuhalten, Unsicherheit zu navigieren und dennoch zu urteilen. Er steht damit quer zu Modellen, die auf Kontrolle, Verifikation oder Exklusion setzen. Statt AS durch technische Immunsysteme zu bekämpfen, schlägt AS eine andere Strategie vor: die Stärkung des Kohärenzempfindens in offenen, pluralen Diskursen.

Ein salutogenetischer Ausblick legt nahe, dass eine partizipativ gestärkte Medienpraxis die entscheidende Rolle im Umgang mit generativer KI spielen wird. Informationskompetenz und kohärente Kommunikationsräume, sogar Wahrhaftigkeit (Rath 2013) sind dabei keine Eigenschaften der Technologie, sondern Leistungen der Nutzer:innen selbst.

8. Fazit und Ausblick – Von der Sepsis zu Resilienz und Kohärenz

AS bezeichnet eine epistemische Pathologie, die nicht auf externe Desinformation oder böswillige Manipulation zurückzuführen ist, sondern auf eine systemische Rückkopplung interner Outputs in KI-generierten Systemen. Generative LLMs tendieren zu digitaler Selbstvergiftung, indem sie ihre eigenen Inhalte reproduzieren, *re-zirkulieren* und schließlich als vermeintlich autoritative Informationsquellen etablieren.

Demgegenüber plädiert der Beitrag für einen Paradigmenwechsel im Sinne einer digitalen ‚Salutogenese' – weg von reaktiver Abwehr hin zu proaktiver Ermöglichung. Die zentrale These lautet: *Nur durch eine ethische Rekonstruktion unserer Kommunikationsräume, Informationspraktiken*

und Bildungsprozesse lässt sich dieser Pathologie begegnen. Drei theoretische Achsen wurden dazu entfaltet:

1. Die *Futures Literacy* als antizipative Ethik, die mögliche Zukünfte plural reflektiert und partizipativ gestaltet,
2. die kommunikative Rationalität im Anschluss an Habermas, die als normatives Raster zur Beurteilung algorithmischer Äußerungen dient, und
3. die Salutogenese als orientierender Rahmen für eine epistemische Resilienz, die nicht auf Immunisierung, sondern auf Kohärenzbildung zielt.

Diese Triade konvergiert in einem von der Salutogenese inspirierten Begriff, der *digitalen Kohärenz*. Sie bezeichnet die Fähigkeit, auch unter Bedingungen automatisierter Kommunikation Orientierung, Urteilskraft und Anschlussfähigkeit zu bewahren. Digitale Kohärenz entsteht dort, wo Geltungsansprüche nicht verwischt, sondern expliziert werden; wo algorithmische Outputs nicht mystifiziert, sondern kontextualisiert werden; und wo technologische Mittel nicht bloß Effizienz, sondern Verstehbarkeit und Sinn ermöglichen.

In praktischer Hinsicht bedeutet dies, dass KI-Systeme so gestaltet, eingesetzt und bewertet werden müssen, dass sie die Handlungs- und Urteilskompetenz der Beteiligten stärken. Dazu gehören epistemische Labels, partizipative Prüfverfahren, offene Bildungsformate und plural strukturierte Diskursräume. Die Verantwortung für diese Gestaltung liegt nicht allein bei Entwickler:innen oder Nutzer:innen – sondern in einem geteilten, relationalen Raum zwischen Technik, Gesellschaft und Ethik. *Digitale Kohärenz* heißt nicht Anpassung an technische Überforderung, sondern Gestaltung alternativer Möglichkeitsräume. *AS* ist kein Schicksal. Sie ist eine Herausforderung – für unser Wissen, unser Handeln und unsere Vorstellungskraft.

Literatur

Antonovsky, Aaron (1979): Health, stress and coping, San Francisco.

Antonovsky, Aaron (1987): Unraveling the mystery of health: How people manage stress and stay well, San Francisco.

Antonovsky, Aaron (1997): Salutogenese – Zur Entmystifizierung der Gesundheit, Tübingen.

Bendemann, Reinhard von (2022): Christus der Arzt. Frühchristliche Soteriologie und Anthropologie im Lichte antik-medizinischer Konzepte, Stuttgart.

Bender, Emily M. et al. (2021): On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, S. 610–623.

Bertau, Marie-Cécile (1996): Sprachspiel Metapher: Denkweisen und kommunikative Funktion einer rhetorischen Figur, Wiesbaden.

Bourgeois, Robin / Karuri-Sebina, Geci / Feukeu, Kwamou Eva (2024): The future as a public good: decolonising the future through anticipatory participatory action research, in: Foresight 26 (4/2024), S. 533–549. https://doi.org/10.1108/FS-11-2021-0225

Buckingham, David (2015): Defining digital literacy – what do young people need to know about digital media?, in: Nordic Journal of Digital Literacy 10 (4/2015), S. 21–35.

Capurro, Rafael (2010): Informationsethik. Ein interdisziplinärer Ansatz, in: Information – Wissenschaft & Praxis 61 (6/2010), S. 315–320.

CIFS (2023): Metaverse Delphi Study. A delphi study on the development of the metaverse towards 2030. Copenhagen Institute for Futures Studies (CIFS), Kopenhagen (online unter: https://veri-media.io/wp-content/uploads/2023/03/metaverse-delphi-study.pdf – letzter Zugriff: 25.6.2025).

Cohen, Fred (1987): Computer viruses: theory and experiments, in: Computers & Security 6 (1/1987), S. 22–35.

Dwivedi, Yogesh K. et al. (2023): So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy, in: International Journal of Information Management 71 (August 2023), 102642.

Eriksson, Monica / Lindström, Bengt (2005): Validity of Antonovsky's sense of coherence scale: a systematic review Journal of Epidemiology & Community Health 59 (6/2005), S. 460–466.

Floridi, Luciano (2013): The ethics of information, Oxford.

Geeraerts, Dirk (2006): Introduction. A rough guide to cognitive linguistics, in: Dirk Geeraerts (Hg.), Cognitive Linguistics: Basic readings, Berlin, New York, S. 1–28.

Gunkel, David J. / Bryson, Joanna (2014): Introduction to the special issue on machine morality: The machine as moral agent and patient, in: Philosophy & Technology 27 (März 2014), S. 5–8.

Habermas, Jürgen (1981): Theorie des kommunikativen Handelns. Band 1. Frankfurt am Main.

Hvitved, Sofie (2022): What if 99 % of the Metaverse is made by AI? 24.2.2022 (online unter: https://cifs.dk/news/what-if-99-of-the-metaverse-is-made-by-ai – letzter Zugriff: 25.6.2025).

Imperva (2025): 2025 Bad Bot Report. The Rapid rise of bots and the unseen risk for business (online unter: https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report/ – letzter Zugriff: 26.9.2025).

Jonas, Hans (1979): Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation, Frankfurt am Main.

Kalina, Andreas et al. (Hg.) (2018): Mediatisierte Gesellschaften. Medienkommunikation und Sozialwelten im Wandel (Tutzinger Studien zur Politik), Baden-Baden.

Lakoff George (1993): The contemporary theory of metaphor, in: Andrew Ortony (Hg.), Metaphor and Thought. 2. Aufl., Cambridge, S. 202–251.

Lakoff, George / Johnson, Mark (2003): Leben in Metaphern. Konstruktion und Gebrauch von Sprachbildern. 3. Aufl., Heidelberg.

Lunduke, Bryan (2022): The story of the first 'computer bug'... is a pile of lies, in: The Lunduke Journal of Technology, 19.8.2022 (online unter: https://lunduke.substack.com/p/the-story-of-the-first-computer-bug – letzter Zugriff: 25.6.2025).

Mangnus, Astrid C. et al. (2021): Futures literacy and the diversity of the future, in: Futures 132 (September 2021), article 102793. https://doi.org/10.1016/j.futures.2021.102793

Mick, Carola (2021). Das Agency-Paradigma, in: Ullrich Bauer / Uwe H. Bittlingmayer / Albert Scherr (Hg.), Handbuch Bildungs- und Erziehungssoziologie, Wiesbaden, S. 1–15.

Miller, Riel (2011): Futures literacy — embracing complexity and using the future, Ethos 10 (October 2011), S. 23–28.

Moon, Lori / Kirvaitis, Patricija / Madden, Noreen (2016): Selective annotation of modal readings: Delving into the difficult data, in: Linguistic Issues in Language Technology 14 (6/2016) (online unter: https://aclanthology.org/2016.lilt-14.6 – letzter Zugriff: 25.6.2025).

OECD (2023): Regulatory sandboxes in artificial intelligence. OECD digital economy Papers, July 2023, No. 356 (online unter: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/07/regulatory-sandboxes-in-artificial-intelligence_a44aae4f/8f80a0e6-en.pdf – letzter Zugriff: 25.6.2025).

OpenAI (2024): Introducing ChatGPT search, 16. Dezember 2024 (online unter: https://openai.com/index/introducing-chatgpt-search/ – letzter Zugriff: 9.7.2025).

Putterer, Elisabeth (2022): Von der Conceptual Metaphor Theory zur Deliberate Metaphor Theory: Theoretische Annahmen, Kritikpunkte und Klärungsversuche, in: Initium 4 (1/2022), S. 112–127. https://doi.org/10.33934/initium.2022.4.9

Rapanta, Chrysi et al. (2025): Critical GenAI literacy: Postdigital configurations, in: Postdigital Science and Education (2. Juli 2025). https://doi.org/10.1007/s42438-025-00573-w

Rath, Matthias (1988): Intuition und Modell. Hans Jonas' „Prinzip Verantwortung" und die Frage nach einer Ethik für das wissenschaftliche Zeitalter, Frankfurt am Main.

Rath, Matthias (2013): Authentizität als Eigensein und Konstruktion – Überlegungen zur Wahrhaftigkeit in der computervermittelten Kommunikation, in: Martin Emmer / Alexander Filipovic / Jan-Hinrik Schmidt / Ingrid Stapf (Hg.), Echtheit, Wahrheit, Ehrlichkeit. Authentizität in der Online-Kommunikation, Weinheim, S. 16–27.

Rath, Matthias (2018): Data Science – die neue Leitwissenschaft?, in: Thomas Knubben / Erich Schöls / Uli Braun (Hg.), Weltkulturatlas – Kultur in Zeit der Globalisierung. Daten, Geschichten, Grafiken, Stuttgart, S. 21–37.

Rath, Matthias (2024): 'To find the 'rotten apple' – information ethical requirements for the information literacy of autonomous writing engines, in: Serap Kurbanoğlu / Sonja Špiranec / Joumana Boustany / Yurdagül Ünal / İpek Şencan / Denis Kos / Esther Grassian / Diane Mizrachi / Loriene Roy (Hg.), Information Experience and Information Literacy. 8th European Conference on Information Literacy, ECIL 2023, Revised Selected Papers, Part II. Cham, S. 129–139.

Rath, Matthias [im Druck]: Peer-to-Peer Learning mit „Artificial Companions". Hochschuldidaktische Überlegungen zur mediatisierten Lehre in digitalen Kontexten, in: Carolyn Blume / Gudrun Marci-Boehncke / Patricia Ronan (Hg.), Peer-to-Peer-Konzepte in der Sprachendidaktik. Theorie und Praxis für die Hochschullehre, Bielefeld.

Ridder, Jeroen de (2024): online illusions of understanding, in: Social Epistemology 36 (6/2024), S. 727–742.

Ross, Andrew (1990): hacking away at the counterculture, in: Postmodern Culture 1 (1/1990). https://dx.doi.org/10.1353/pmc.1990.0011

Schmale, Günter (2019): Mögliche Metaphern in der Fachsprache, in: ELAD-SILDA Études de Linguistique et d'Analyse des Discours – Studies in Linguistics and Discourse Analysis 2 (Oktober 2019), S. 1–32.

Semino, Elena / Demjén, Zsófia (Hg.) (2016): The Routledge handbook of metaphor and language. London. https://doi.org/10.4324/9781315672953

Shumailov, Ilia et al. (2023): The curse of recursion: Training on generated data makes models forget, in: arXiv:2305.17493. https://doi.org/10.48550/arXiv.2305.17493

Singer, Mervyn et al. (2016): The third international consensus definitions for sepsis and septic shock (Sepsis-3), in: JAMA 315 (8/2016), S. 801–810.

Steen, Gerard J. (2008): The paradox of metaphor: Why we need a three-dimensional model of metaphor, in: Metaphor and Symbol 23 (4/2008), S. 213–241. https://doi.org/10.1080/10926480802426753

Steen, Gerard J. (2023:) Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension, in: Frontiers in Psychology 14 (5. September 2023), 1242888. https://doi.org/10.3389/fpsyg.2023.1242888

Truby, Jon et al. (2022): A sandbox approach to regulating high-risk artificial intelligence applications, in: European Journal of Risk Regulation 13 (2/2022), S. 270–294.

Undheim, Kristin / Erikson, Truls / Timmermans, Bram (2023): True uncertainty and ethical AI: regulatory sandboxes as a policy tool for moral imagination, in: AI Ethics 3 (August 2023), S. 997–1002.

UNESCO (2018): Transforming the future: anticipation in the 21st century. London (online unter: https://unesdoc.unesco.org/ark:/48223/pf0000264644 – letzter Zugriff: 25.6.2025).

Vaswani, Ashish et al. (2017): Attention is all you need, in: arXiv: 1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Victor, Bryan G. / Goldkind, Lauri (2025): The therapist in the machine: Confronting AI's Challenge to clinical social work, in: Journal of Technology in Human Services 43 (2/2025), S73–81. https://doi.org/10.1080/15228835.2025.2500827

Weidinger, Laura et al. (2021): Ethical and social risks of harm from Language Models, in: arXiv:2112.04359v1. https://doi.org/10.48550/arXiv.2112.04359

Wittgenstein, Ludwig (1999): Philosophische Untersuchungen, in: Gertrude Elizabeth Margaret Anscombe / Rush Rhees /Georg Henrik von Wright (Hg.), Ludwig Wittgenstein Werkausgabe. Band 1, Frankfurt am Main, S. 231–485.

Zhao, Haiyan et al. (2024): Explainability for large language models: A survey, in: ACM Transactions on Intelligent Systems and Technology 15 (2/2024), Article 20. https://doi.org/10.1145/3639372

Zhao, Zihao et al. (2021): Calibrate before use: Improving few-shot performance of language models, in: Proceedings of Machine Learning Research 139, 12697–12706 (online unter: https://proceedings.mlr.press/v139/zhao21c.html – letzter Zugriff: 25.6.2025).

Ethische Perspektiven auf Große Sprachmodelle am Beispiel von Trainingsdatenqualität

Jana Hecktor

Zusammenfassung

Die Entwicklung sogenannter Sprachmodelle hat durch ihre Einbindung in generative KI-Anwendungen in den letzten Jahren eine gesteigerte wissenschaftliche und gesellschaftliche Relevanz erhalten. Während sich ihre Einsatzfelder stetig erweitern und sie zunehmend Teil des Alltags vieler Menschen werden, haben sich parallel diverse kritische Perspektiven herausgebildet. Insbesondere die potenziell fehlerhaften, halluzinierten sowie diskriminierenden Systemausgaben stehen hierbei oftmals im Zentrum der Diskussion. Der vorliegende Beitrag widmet sich einer spezifischen Perspektive auf diese Problemfelder: einer ethisch orientierten Betrachtung der Qualität von Trainingsdaten als möglichem Ansatz, um konkrete Risiken von Sprachmodellen abzuschwächen. Hierfür werden der Leitfaden für Datenqualität in KI-Systemen des Bundesamts für Sicherheit in der Informationstechnik sowie das Glossar des Forschungsprojektes KITQAR herangezogen, um einen Überblick zu möglichen Kriterien der Trainingsdatenqualität zu gewinnen. Anschließend werden einzelne Kriterien wie Diversität oder Repräsentativität exemplarisch in den Blick genommen, um zu prüfen, inwieweit sie für das Training großer Sprachmodelle produktiv Anwendung finden können und welche Überlegungen insbesondere aufgrund möglicher „Trade-offs" zwischen einzelnen Werten aus einer ethisch geleiteten Perspektive zu berücksichtigen sind.

1. Einleitung

In den letzten Jahren haben generative KI-Systeme für einen enormen Anstieg in der Verbreitung und Nutzung von KI-Anwendungen gesorgt. Die Art und Weise, in der ChatGPT und ähnliche Systeme genutzt werden, ist dabei durchaus vielfältig und sorgt neben großem Interesse und hohen

Nutzendenzahlen[1] auch immer wieder zu kritischen oder warnenden medialen Berichterstattungen. In seinen diversen medialen Auftritten spricht auch Sam Altman, der CEO von OpenAI, immer wieder von Herausforderungen im Umgang mit dem von ihm (mit) entwickelten System. Er verweist unter anderem darauf, dass Nutzende dem System ohne Notwendigkeit höfliche Floskeln wie „Danke" und „Bitte" kommunizieren und auf diese Weise den Strom- und Rechenbedarf des Systems ‚unnötig' in die Höhe treiben (vgl. Reyes 2025). Zeigt dies die Herausforderungen von OpenAI mit möglicherweise unerwartetem Verhalten seitens der Nutzer:innen umzugehen, lassen sich die kritischen Ebenen des Systems aus ethisch angeleiteter Perspektive in der Regel vor allem in den Auswirkungen für Nutzer:innen selbst finden. So wird ChatGPT inzwischen auch für unzählige, teils sehr private, sensible, intime oder komplexe Aufgaben genutzt und dabei als eine Art Suchmaschine und Ratgeber in Einem verstanden. Nutzende fragen das System nach medizinischen Diagnosen, lassen sich einem viralen Trend auf den sozialen Medien folgend die eigene Zukunft imaginieren, verwandeln das System in einen virtuellen Partner, oder nutzen Sprachmodelle für Therapiesitzungen (vgl. Afshar 2024; Ayre/Cvejic/McCaffery 2025; Lee 2024; OpenAI 2025a; OpenAI 2025b; Raile 2024).[2] Auf diese Weise dringen Systeme wie ChatGPT in höchst vulnerable, intime und sehr private Bereiche ihrer Nutzenden ein. Ein entscheidender Grund für den Erfolg lässt sich vermutlich darin finden, dass Nutzende mit dem System via Sprache interagieren können. Mit Blick auf medizinische Diagnosen vermutet Sam Altman, dass ebenso entscheidend ist, dass Nutzende ihre Antworten direkt und schnell haben wollen. Dies sei ihnen scheinbar wichtiger als gesichert akkurate Antworten zu erhalten. In seinem Podcast warnt Altman im Wissen um all dies vor den Gefahren, die von der Nutzung ausgehen können und betont, dass ChatGPT keine Wahrheiten ausgebe, sondern stattdessen Wortnachbarn berechne. Wenngleich manche hinter diesen Warnungen einen wohl durchdachten Coup des CEOs sehen, zeigt es doch auch zwei Dinge sehr deutlich: 1. (Sprach-)Modelle wie ChatGPT[3] werden vielfältig und ubiquitär für viele, teils sehr private,

1 In seinem TED-Talk im April 2025 sprach Sam Altman davon, dass etwa 10 Prozent der Weltbevölkerung ChatGPT regelmäßig nutzen würden (vgl. TED 2025).

2 In gewisser Weise führen Systeme wie ChatGPT damit eine Art der Nutzung fort, wie sie bereits in dem 1966 von Joseph Weizenbaum entwickelten System ELIZA angelegt war.

3 Obgleich ChatGPT in der Regel als Sprachmodell diskutiert wird, setzt sich das System aus verschiedenen technischen Programmierungen zusammen.

Zwecke genutzt; 2. Die Art und Weise wie Sprachmodelle in Anwendung gebracht werden, kann durchaus problematisch sein.[4]

Zu den häufigsten Anwendungen großer Sprachmodelle gehören das Übersetzen von Texten, das Erstellen von textbasierten Inhalten, das Zusammenfassen von Informationen sowie das Beantworten konkreter Fragen auf Basis individueller Prompts.[5] Diese Fülle an Möglichkeiten hat dazu geführt, dass Sprachmodelle in ganz unterschiedlichen Feldern Anwendung finden. Unternehmen, Plattformen und Institutionen implementieren Sprachmodelle auf ihren Webseiten, in ihren Callcentern oder als Assistenz in Form von KI-Agenten und Chatbots, die Kundenanfragen beantworten, Support leisten und allgemeine Informationen bereitstellen können. In vielen Berufen finden Sprachmodelle Anwendung, um bei textbasierten Aufgaben zu unterstützen – E-Mails werden vorformuliert, Pressemitteilungen übersetzt oder Textbausteine ganz von KI geschrieben (vgl. Heesen et al. 2023). Diese Möglichkeiten der Textarbeit haben insbesondere im Bildungsdiskurs bereits einige Diskussionen ausgelöst und zu der Entwicklung vielfältiger Handreichungen geführt, die eruieren, auf welche Weise KI-Systeme im Bildungssektor genutzt werden sollten oder dürfen (vgl. Scheiter et al. 2025). Sprachmodelle werden dabei nicht allein für jene Arten der Textarbeit genutzt, die formalen Logiken folgt, sondern durchaus auch für mit Kreativität verbundene Arbeitsprozesse in den Einsatz gebracht. Exemplarisch sei an dieser Stelle auf die Games-Branche verwiesen, in der Große Sprachmodelle inzwischen unter anderem dazu genutzt werden, Charakterkonzepte zu brainstormen oder sich Anregungen für Umwelten und narrative Elemente ausgeben zu lassen (vgl. Shaker et al. 2016; Galotta et al. 2024; Thompson 2024; Yannakakis/Togelius 2024). Viele dieser Möglichkeiten lassen sich auch abseits des beruflichen oder bildungstheoretischen Kontextes im privaten Alltag nutzen. Große Sprachmodelle können dabei helfen, eine E-Mail an den Stromanbieter zu schicken, die Nachrichten des Airbnb-Gastgebers zu übersetzen oder die eigene Charaktererstellung für Pen & Paper unterstützen.

4 ChatGPT steht dabei an dieser Stelle stellvertretend für viele andere ähnliche Systeme, die als generative KI basierend auf komplexen Sprachmodellen Text- oder Bildeingaben verarbeiten, um anschließend Text- oder Bildausgaben zu produzieren.

5 Ein wesentlicher Unterschied zu der übergeordneten Kategorie der generativen KI liegt in dem Umfang des medialen Inputs und Outputs. Während Große Sprachmodelle sich auf sprachbasierte Daten konzentrieren, umfasst generative KI auch andere Medienformen wie die Bildgenerierung oder das Verarbeiten von Code.

All dies zeigt, wie vielfältig sich Sprachmodelle in verschiedene Diskurse eingeschrieben haben und wie viel Potential für zukünftige Anwendungen in ihnen liegt. Die unbestreitbare Omnipräsenz der Systeme und ihre Wirkmacht innerhalb heterogener gesellschaftlicher Prozesse führen dazu, dass auch die negativen Dimensionen der Systeme gesamtgesellschaftliche Konsequenzen haben können. Nicht überraschend haben sich daher in den letzten Jahren auch kritische ethische Perspektiven auf diese KI-Systeme eröffnet.[6] Davon abgesehen, dass selbst der Firmenchef von OpenAI auf die kritischen Dimensionen des eigenen Systems hinweist, wurde in der medialen Berichterstattung sowie innerhalb wissenschaftlicher Untersuchungen bereits vielfach deutlich gemacht, dass die Generierung sowie die Nutzung von Sprachmodellen negative Konsequenzen haben können. Dies beginnt bei der Ausgabe diskriminierender Inhalte, dem Generieren von Desinformationen oder Halluzinationen und geht hin zu dem fehlenden Schutz vulnerabler Nutzer:innen sowie Problemen des Copyrights und des Datenschutzes (vgl. Chun 2021; Heesen et al. 2021; Mehrabi et al. 2021; Loh 2024).

Bei einer Medientechnologie mit derart breit gefächerter sowie allgegenwärtiger Nutzung können auch die negativen Dimensionen der Systeme große Wirkmacht entfalten. Einen ethisch angeleiteten Blick auf die negativen Dimensionen Große Sprachmodelle zu werfen, bedeutet daher sowohl die gesamtgesellschaftlichen Herausforderungen zu adressieren als auch sich den Risiken der konkreten und individuelle Nutzung zu widmen. Der vorliegende Text möchte dabei insbesondere die inhärenten Logiken der Systeme als Ausgangspunkt konkreter Herausforderungen in den Blick nehmen. Hierfür soll sich exemplarisch den Trainingsdaten der Modelle als entscheidendes Element der Funktionalität der Systeme gewidmet werden.

6 An dieser Stelle sei darauf verwiesen, dass eine ethische Betrachtung grundsätzlich eine Abwägung von Werten und Gütern bedeutet und nicht zwingend negative Perspektiven zur Folge hat. KI-Anwendungen können durchaus positive Folgen haben oder für produktive Zwecke eingesetzt werden, beispielsweise indem Inhalte einfacher an individuelle Bedürfnisse angepasst werden, was wiederum zu einer gesteigerten Inklusion gesellschaftlicher Minderheiten führen kann.

2. Ethische Perspektiven auf Sprachmodelle

Große Sprachmodelle sind ein entscheidender Teilbereich der Entwicklung generativer KI. Viele an den Begriff KI herangetragenen kritischen Perspektiven lassen sich daher auch auf Sprachmodelle übertragen.[7] Eine Dimension, die in den letzten Jahren zunehmend Aufmerksamkeit erhalten hat, sind die ökologischen Folgen der Systeme. Als Medientechnologie basieren Sprachmodelle auf technischen Infrastrukturen und Produktionsbedingungen, die sich entlang von Fragen nach Umweltverträglichkeit, Arbeitsbedingungen und Ressourcenverteilung adressieren lassen. Für die zugrundeliegende technische Infrastruktur der Systeme werden kontinuierlich große Mengen unterschiedlicher Ressourcen in den Einsatz gebracht. Von Frischwasser zum Kühlen der Serverfarmen über seltene Edelmetalle für die technisch-materielle Grundlage hin zu einem enormen Strombedarf zur Generierung, Aufrechterhaltung sowie der konkreten Nutzung der Systeme. Die Endlichkeit der einzelnen Rohstoffe erfordert Entscheidungen, die nicht nur ökologische und ökonomische, sondern auch machtgeprägte Verteilungslogiken auf den Plan rufen. Das Wasser, welches dem Kühlen der heiß laufenden Server dient, wird insbesondere in trockenen Gebieten auch an anderer Stelle gebraucht, und die Produktion des benötigten Stroms kann nicht unbedingt (allein) durch umweltbewusste Quellen produziert werden. Im letzten Jahr ging durch die Medien, dass Google darüber nachdenkt, eigene Mini-Atomkraftwerke in Anwendung zu bringen, um den durch KI-Systeme gestiegenen eigenen Strombedarf zu decken. Edelmetalle wiederum sind selten und haben zugleich viele relevante Anwendungsbereiche. Neben der Entscheidung, wann und zu welchem Zwecke diese seltenen Stoffe weiterverarbeitet werden sollen, steht oftmals auch die teils umweltbelastende oder auf Ausbeutungslogiken basierende Gewinnung der Materialien in der Kritik (vgl. Crawford 2021; OECD 2022). Wenn Systeme wie ChatGPT für möglicherweise unpassende, unnötige, problematische oder sogar rechtlich fragwürdige sowie ethisch inakzeptable Prozesse in Anwendung gebracht werden, könnte man zurecht argumentieren, dass die umweltlichen Dimensionen der Systeme eine doppelte Ebene der Kritik eröffnen. *Doppelt* in dem Sinne, dass einerseits grundsätzlich zu reflektieren

7 Wenngleich sich dieser Text insbesondere auf die Trainingsdatenqualität als konkreten Anwendungsfall konzentriert, sollen zu Anfang auch einige allgemeinere Ebenen aufgerufen werden, die als Perspektive wiederum bei der Generierung sowie Nutzung von Trainingsdaten ebenfalls eine Rolle spielen (können).

ist, inwieweit die Nutzung der Systeme bzw. der dabei erhaltene Output die hierfür verbrauchten Ressourcen wert ist. Diese Ebene verstärkt sich andererseits noch, wenn die Ressourcen für einen Prozess verbraucht werden, der keinen substanziellen Mehrwert mit sich bringt oder dessen Ergebnisse vielleicht sogar für negative Zwecke (Deep Fakes oder Desinformationen) in Anwendung gebracht werden.[8] Eine weitere gesamtgesellschaftliche Ebene ließe sich in den Veränderungen adressieren, die Große Sprachmodelle in einzelnen gesellschaftlichen Diskursen und Lebensbereichen auslösen, wie bei den bereits erwähnten Auswirkungen in Arbeits- und Bildungskontexten. Damit die Systeme so funktionieren, wie sie es tun, bedarf es zudem explizit menschlicher Arbeit. Mit Begriffen wie Data-Work, Click-Work oder auch als Ghost Work[9] bezeichnete Arbeitsprozesse dienen der Zuarbeit von maschinellen Systemen, beispielsweise in der Klassifizierung von Datensätzen oder der Moderation von Inhalten, die den Richtlinien der jeweiligen Anbieter widersprechen. Diese Arbeit wird in der Regel in Länder mit sehr viel geringeren Löhnen ausgelagert und geschieht ohne langfristige Absicherung oder Arbeitsschutz (vgl. Gray/Suri 2019; Distelmeyer 2025).

Der zunehmenden Bedeutung sowie der negativen Wirkmacht der Systeme tragen nicht nur mediale Berichterstattungen und wissenschaftliche Untersuchungen Rechnung. In Europa hat die KI-Verordnung deutlich gemacht, dass auch politische und juristische Akteur:innen die Wirkmacht von KI-Anwendungen ernst nehmen. Technologische Systeme werden hierbei dem Risiko ihrer Nutzung folgend klassifiziert und reguliert. Wenngleich damit ein erster wichtiger Schritt getan wurde, um dem risikobehafteten Einfluss von KI-Systemen Grenzen zu setzen, muss dennoch festgehalten werden, dass die KI-Verordnung ein *regionaler* Vorstoß gegen ein international wirkmächtiges Phänomen ist. Dazu kommt, dass der Einsatz der Technologie sich oft schneller verbreitet und etabliert, als es Regularien und Forschungen einfangen können. Es scheint daher umso relevanter,

8 Ein reflektierter Umgang mit den Systemen sollte idealerweise bedeuten, mit Blick auf den großen Ressourcenbedarf oder die Fehleranfälligkeit der Systeme in Frage zu stellen, ob für die konkreten Handlung, die Nutzung von großen Sprachmodellen tatsächlich die beste oder einzige Lösung ist.

9 Jan Distelmeyer verweist in seinem Beitrag (Distelmeyer 2025) zurecht darauf, dass dieser Begriff die Unsichtbarmachung, die kritisiert werden soll, rekonstituiert und zugleich nicht für alle Arbeiter:innen zutreffend ist.

dass die entsprechenden Entwicklungen kontinuierlich kritisch perspektiviert werden.[10]

3. Systeminhärente Logiken

Neben den gesamtgesellschaftlichen Herausforderungen ist insbesondere die fehlerhafte Ausgabe von Informationen Fokus kritischer Untersuchungen. Die von Sprachmodellen generierten Inhalte können aus unterschiedlichen Gründen problematisch sein. Zu den größten Herausforderungen gehören die Halluzinationen der Systeme sowie die zum Teil in ihnen verankerten Diskriminierungen. Als Halluzination werden von KI-Systemen generierte Inhalte bezeichnet, die plausibel und wahrheitsgemäß erscheinen, aber de facto aus einer zufälligen statistischen Verteilung entstanden sind und daher weder wahr noch durch konkrete Quellen belegt sind. Einzelne aus den systeminhärenten Logiken der Systeme entstehende Halluzinationen werden regelmäßig durch viral gehende Fallbeispiele auch medial verhandelt.[11] Ist es vielleicht noch amüsant, wenn ChatGPT selbstsicher simple Additionen falsch beantwortet oder die Zugspitze zum höchsten Berg der Welt erklärt, dann sieht es schon anders aus, wenn das System souverän wissenschaftliche Quellen und Thesen für diskriminierende oder sexistische Inhalte ausgibt. Solche Halluzinationen der Systeme können in Form von glaubwürdig aussehenden Desinformationen weitreichende Folgen haben, wenn sie zum Beispiel intendiert generiert und von einzelnen Personengruppen anschließend zu manipulativen Zwecken weiterverbreitet werden.[12] Oder aber, wenn Personengruppen mit solch halluzinierten Inhalten konfrontiert werden, denen es entweder nicht möglich ist, den Wahrheitsgehalt zu überprüfen oder denen nicht bewusst ist, dass sie dies bei der Nutzung Großer Sprachmodelle tun sollten. Obgleich dies für

10 Ein Konzept hierfür sind Living Guidelines, die in regelmäßigen Abständen aktualisiert und den Entwicklungen folgend angepasst werden (vgl. ERA Forum/DG RTD 2024).

11 Viral ging so unter anderem das Video eines amerikanischen Anwalts, der sich von ChatGPT juristische Vergleichsfälle für seinen aktuellen Prozess ausgeben ließ, die sich schlussendlich als fabriziert herausstellten (vgl. Bohannon 2023).

12 Das Institut für strategischen Dialog hat im Jahr 2025 einen Bericht veröffentlicht, der darlegt, auf welche Weise rechtsextreme Akteur:sinnen in Deutschland generative KI gezielt einsetzen, um ihre Narrative online zu verbreiten (vgl. Hiller/Maristany de las Casas 2025).

alle Nutzer:innen gilt, scheint die Thematik noch relevanter, wenn man spezifische Bevölkerungsgruppen in den Blick nimmt. Vulnerable Gruppen wie Kinder, Senioren oder gesellschaftliche Minderheiten können aus unterschiedlichen Gründen besonders anfällig für die negativen Folgen von Sprachmodellen sein – insbesondere, wenn die Nutzung mit fehlender *literacy* verbunden ist, also einem mangelnden Verständnis der Funktionslogik der Systeme. Zu den Problemen, die Halluzination und Desinformation mit sich bringen, kommt hinzu, dass die Systeme möglicherweise sexualisierte, gewaltvolle, diskriminierende oder Copyright- und Persönlichkeitsrechte missachtende Inhalte ausgeben. Wenngleich dies in den Richtlinien der jeweiligen Unternehmen ausgeschlossen wird und ChatGPT bei konkreten Worten und Anfragen auch ausgibt „Dieser Inhalt verstößt möglicherweise gegen unsere Nutzungsrichtlinien", können entsprechenden Grenzen und Beschränkungen durch abweichendes Prompting gezielt (*jailbreaking*) sowie unabsichtlich umgangen werden (vgl. Liu et al. 2023). Obgleich also Filter und Grenzen für gewaltvolle, diskriminierende oder sexuelle Inhalte eingerichtet worden sind, können diese absichtlich sowie auch unabsichtlich umgangen werden und auf diese Weise entsprechende Inhalte auch an Kinder und Jugendliche ausgegeben werden, die davor geschützt werden sollten.[13] Auch andere Personengruppen können gefährdende oder sich negativ auswirkende Inhalte ausgegeben bekommen, beispielsweise wenn aufgrund traumatischer Erfahrungen konkrete ‚Triggerpunkte' existieren.

Die Herausforderungen, die sich aus der konkreten und individuellen Nutzung in Kombination mit den inhärenten Logiken der Systeme ergeben, können auf verschiedene Systemlogiken und Anwendungsprozesse zurückgeführt werden. Ein konkretes Scharnier der Funktionslogik sind die jeweils genutzten Daten. Sie bilden die Trainingsgrundlage der Systeme und damit das Fundament der Modellierung, sie sind Input (Prompt) sowie Output, aber auch elementarer Bestandteil der Anwendungslogik. Sie sind nicht nur notwendiges technisches Element der auf maschinellem Lernen basierenden Modellierungsverfahren, sie bilden zugleich die Grenzen und den Möglichkeitsraum der daraus entstehenden Sprachmodelle. Wenngleich nicht alle mit Großen Sprachmodellen verknüpften Probleme und Herausforderungen durch einen Blick auf die Trainingsdaten erklärt oder durch den Einsatz von entsprechenden ethischen Standards gelöst

13 Verstärkend kommt an dieser Stelle hinzu, dass es für ChatGPT zwar ein Mindestalter gibt, dieses bei der Nutzung allerdings nicht abgefragt wird.

werden können, lassen sich doch einige der angeführten Probleme auf die konkreten Datenmengen zurückführen.

4. Datenauswertung als entscheidendes Scharnier von Sprachmodellen

Derzeitige Entwicklungen des maschinellen Lernens hängen entscheidend an den jeweils vorhandenen und in den Einsatz gebrachten Trainingsdaten. Ohne zu tief in die technischen Funktionalitäten einzusteigen, sei darauf verwiesen, dass für die Entwicklung generativer KI sowie Großer Sprachmodelle sogenannte Deep Learning-Verfahren angewandt werden. Hierbei werden künstliche neuronale Netze mit einer großen Menge gelabelter oder auch ungelabelter Daten gespeist und trainiert. Im Laufe der Trainingsphase werden die Gewichtungen der Systeme kontinuierlich angepasst, sodass am Ende das Modell die ihm aufgetragene Aufgabe erfolgreich durchführen kann. Im Falle von Sprachmodellen ist dies das Identifizieren von konkreten Worten und ihren Relationen zueinander sowie das Antizipieren der darauf basierenden gewünschten Antwort mittels Mustererkennung. Daten bilden auf diese Weise einerseits die Grundlage für das Training der Modelle – ohne sie kann kein Sprachmodell entstehen –, setzen dadurch aber auch die Grenzen und Möglichkeitsbedingungen fest. Inhalte, die nicht Teil der Datenmenge sind, können nicht ausgegeben werden. Daneben schreiben sich auch implizit in den Daten vorhandene Relationen in die Modellierungen ein, dies kann im Extremfall dazu führen, dass ein wie auch immer gearteter Bias in den Datenmengen sich auch in den Ergebnissen der Systeme widerspiegelt. Dies allein zeigt bereits, welche Bedeutung der Implementierung der ‚richtigen' Daten für die Funktionalität der jeweiligen Systeme zukommt. So hat das Bundesamt für Sicherheit in der Informationstechnik (BSI) Anfang Juli 2025 einen methodischen Leitfaden zur Datenqualität in KI-Systemen vorgestellt und auch in der KI-Verordnung werden Daten und Daten-Governance als entscheidende Ebene adressiert und reguliert.

Neben der zum Zwecke des Datenschutzes im Jahr 2016 bereits in Kraft getretenen Datenschutz-Grundverordnung (DSGVO), deren Regulierungen auch für seitdem entstandene Technologien Gültigkeit besitzt, listet auch die KI-Verordnung explizite Kriterien und Regulierungen für den Einsatz von Daten im Diskursfeld rund um KI. Zu den spezifischen Regulierungen, die sich auf Große Sprachmodelle anwenden lassen, gehören neben der Datenqualität und der Daten-Governance auch Transparenzforde-

rungen, Risikobewertung und -management sowie die Pflicht zur Überwachung und Berichterstattung der Systeme. Mit der Transparenzforderung ist festgehalten, dass Anbieter von Sprachmodellen klar kommunizieren müssen, wenn die jeweiligen Inhalte von KI-Systemen generiert sind. Auf diese Weise soll verhindert werden, dass Nutzer:innen fälschlicherweise davon ausgehen, dass die Inhalte von menschlichen Akteur:innen erstellt wurden und entsprechende Zuschreibungen vornehmen. Unter dem Aspekt der Risikobewertung und dem Risikomanagement wird wiederum festgehalten, dass Sprachmodelle einer entsprechenden Bewertung unterzogen werden müssen, durch die Gefahren identifiziert und minimiert werden. Dieser Aspekt verweist auch bereits auf die Einteilung von Technologien basierend auf ihrem Risiko-Level. Große Sprachmodelle als Teilbereich generativer KI können je nach Anwendungskontext als Hoch-Risiko-System eingestuft werden. Aus diesem Grund sind Anbieter von Sprachmodellen wie OpenAI auch angehalten, die Leistung und Sicherheit ihrer Systeme kontinuierlich zu kontrollieren und entsprechende Berichte über die jeweilige Nutzung und die damit verbundenen Risiken zu erstellen. Nicht zuletzt hält die KI-Verordnung fest, dass es eine gewisse Form der Datenqualitätssicherung und Daten-Governance braucht. Dies bedeutet, dass Firmen Maßnahmen ergreifen sollten, um sicherzustellen, dass die Daten, die für das Training von Sprachmodellen eingesetzt werden, entsprechende Qualitätsmerkmale erfüllen und insbesondere frei von diskriminierenden oder schädlichen Inhalten sind. Damit gesetzliche Rahmungen oder ethische Standards für Trainingsdaten in Anwendung gebracht werden können, braucht es konkrete Trainingsdaten, die an diesen Standards und Richtlinien entlang bewertet werden können.

Eine erste wichtige Frage vor der Generierung Großer Sprachmodelle ist daher, ob es für den entsprechenden Anwendungsfall bereits entsprechende Datenmengen gibt, die die gewünschten Inhalte abbilden, oder ob erst Daten erhoben werden müssen. Für den Fall, dass es noch keine (adäquaten) Daten gibt und diese beschaffen oder erfasst werden müssen, ist es relevant zu überprüfen, ob für den konkreten Anwendungsfall eine datafizielle Erfassung überhaupt möglich ist, einerseits mit Blick auf technische Möglichkeitsbeschränkungen wie unzureichender Sensorik oder aber bei nur schwer (objektiv) zu dokumentierenden Phänomenen.

Darüber hinaus sollte in jedem Fall entschieden werden, ob eine Datenerfassung die möglichen Risiken des jeweils konkreten Anwendungsfall Wert ist. Fallen die entsprechenden Bereiche beispielsweise in den Schutz

der Privatsphäre und das Recht der informationellen Selbstbestimmung? Werden vulnerable Gruppen einem potentiell diskriminierenden Blick geöffnet? Ist bereits absehbar, dass die erfassten Daten durch die Einbindung in ein Sprachmodell möglicherweise problematischen Nutzungslogiken ausgeliefert sein werden? Wenngleich nicht all diese Fragen vorab beantwortet werden können, sollten sie für eine ethisch angeleitete Generierung von Modellierungen mitbedacht werden. Gesetzt den Fall, dass es sowohl möglich ist, Daten zu erheben oder auf vorhandene Datensätze zurückzugreifen und es auch keine ethischen oder gesetzlichen Einwände gegen die Erhebung oder Auswertung der jeweiligen Datensätze gibt, kann es dennoch sein, dass aus den scheinbar unproblematischen Daten, problematische Modellierungen entstehen.[14] Es sollte daher auch gefragt werden: Welche Qualität haben die genutzten Datensätze? Und auf welche Weise können Daten kritisch betrachtet und ethischen Standards folgend erhoben und eingesetzt werden?

5. Trainingsdatenqualität als Grundbedingung für ethische Sprachmodelle

Das Forschungsprojekt KITQAR, an dem die Universität Tübingen (Internationales Zentrum für Ethik in den Wissenschaften), das Hasso-Plattner-Institut (HPI), die Universität zu Köln und der Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE) beteiligt sind, hat in seiner Laufzeit die Qualität von KI-Test- und Trainingsdaten in der digitalen Arbeitsgesellschaft erforscht. Ein zentrales Ergebnis dieses Projekts ist die Generierung und Veröffentlichung eines Glossars zur Datenqualität (vgl. Mohammed et al. 2023). Datenqualität wird dabei definiert als die „Eignung von Daten für einen bestimmten Anwendungszweck" (Mohammed et al. 2023: 1). Die Qualität selbst kann anhand ganz unterschiedlicher Kriterien bewertet werden.

Einige dieser Kriterien sind automatisch messbar oder überprüfbar, während andere nur von Expert:innen bewertet werden können. Manche der Kriterien sind nicht nur voneinander abhängig, sondern beeinflussen oder nivellieren sich. Wenngleich für den Einsatz von Trainingsdaten in der digitalen Arbeitsgesellschaft entwickelt, lassen sich die entsprechenden Kriterien auch für die Entwicklung von Großen Sprachmodellen in Anwendung

14 Beispielsweise dann, wenn die Zusammenführung zuvor unabhängiger Daten und Datensätze in einem Modell zu neuen Auswertungslogiken führt.

bringen. In dem vom Bundesamt für Sicherheit in der Informationstechnik (BSI) veröffentlichten methodischen Leitfaden zur Datenqualität in KI-Systemen werden wiederum zehn Kriterien aufgelistet, die sich direkt sowie indirekt auch im Glossar von KITQAR wiederfinden lassen. Für das BSI sind bei der Entwicklung generativer KI folgender Kriterien entscheidend: Repräsentativität, Vollständigkeit, Genauigkeit, Konsistenz, Korrektheit, Einheitlichkeit, Gültigkeit, Eindeutigkeit, sichere Quellen sowie die Überprüfung auf Personenbezug bei den jeweils genutzten Daten und damit auch die Einhaltung des Datenschutzes. Die zehn Kriterien werden in dem Leitfaden durch konkrete Bausteine wie Vielfalt, Ausgewogenheit, Konsistenzsicherung oder Expertenanalyse ergänzt.[15] KITQAR listet 29 Kriterien zur Bewertung der Datenqualität auf, die im Kontext für KI-Test- und Trainingsdatenqualität in der digitalen Arbeitsgesellschaft von Bedeutung sind. Von der *Aktualität* der Daten bis hin zu Fragen der *Zugänglichkeit* werden die verschiedenen Kriterien nicht nur erklärt und in Relation zueinander gesetzt, sondern auch mit der DSGVO und dem KI-Act untermauert.

Sowohl KITQAR als auch das BSI inkludieren in ihren Kriterien verschiedene Ebenen der Trainingsdatenqualität. Einige beziehen sich auf die Generierung oder die jeweilige Herkunft der Datenmenge, andere fokussieren den konkreten Inhalt der einzelnen Daten sowie ihre Relation zueinander und wieder andere adressieren den kontinuierlichen Umgang mit bereits eingebundenen Datenmengen. Für alle Kriterien gilt jedoch, dass in der Regel nicht alle von ihnen zugleich in Anwendung gebracht werden können. Oftmals braucht es Abwägungen, die zu einem Trade-Off von ethischen Werten und Risiken führen können. Um die Wirkmacht, aber auch die Grenzen von ethischen Kriterien für Trainingsdatenqualität aufzuzeigen, sollen nachfolgend exemplarisch einige der Kriterien am Beispiel Großer Sprachmodelle genauer ausgeführt werden. Trainingsdaten bilden einerseits die notwendige Grundbedingung für die Modellbildung von Großen Sprachmodellen, müssen aber selbst oftmals erst produziert, zusammengeführt und nach konkreten Kriterien aufgearbeitet werden, bevor sie nutzbar sind. Eine erste entscheidende Perspektive von Kriterien für Trainingsdatenqualität richtet sich daher auf die Herstellung sowie die Akquirierung der jeweiligen Trainingsdaten. KITQAR listet so *Ansehen* als eines ihrer Kriterien und referiert damit auf die Vertrauenswürdigkeit der Datenquelle. Dieses Kriterium anzuwenden, bedeutet einerseits, dass

15 Insgesamt listet der Leitfaden 15 Bausteine auf.

Datenquellen, zu denen bereits (gute) Erfahrungswerte vorliegen, jenen vorzuziehen sind, bei denen man nicht einschätzen kann, welche Qualität die jeweiligen Daten haben werden. Der Leitfaden des BSI wiederum verweist darauf, dass stets auf sichere Quellen zurückgegriffen werden muss, so zum Beispiel im wissenschaftlichen Kontext auf Veröffentlichungen, die mit Peer Review überprüft worden sind. Die Realität sieht jedoch oft anders aus. ChatGPT stand bereits mehrfach in der Kritik dafür, dass seine Trainingsdaten durch effizientes Scraping aus dem Netz gezogen wurden, ohne stark zu differenzieren, welche Quellen sie dabei anzapfen.[16] Die meisten der sowohl von KITQAR als auch in den Leitlinien aufgelisteter Kriterien widmen sich der inhaltlichen Qualität der Trainingsdaten. Von den 29 Kriterien für Trainingsdatenqualität lassen sich für die ethische Betrachtung Großer Sprachmodelle vor allem die Kriterien Ausgewogenheit, Datenschutz, Diversität, Fairness, Korrektheit, Privatsphäre und Repräsentativität benennen.

Datensätze sind, gemäß dem Glossar, dann *ausgewogen*, wenn die Datenpunkte innerhalb des repräsentierten Wertebereichs im Verhältnis zueinander gleich verteilt sind. Für den Fall, dass für die Erfassung von Kund:innendaten beispielsweise nach Altersgruppen sortiert werden soll, braucht es für eine ausgewogene Datenmenge, pro in das Training inkludierter Altersgruppe, eine entsprechend gleich große Datenmenge. Auch das BSI führt Ausgewogenheit als Kriterium an und ordnet diesem die Funktion zu, Verzerrungen in Form von Unter- oder Überrepräsentation entgegenwirken zu können. Auf Große Sprachmodelle übertragen, braucht es ein alternatives Beispiel. Exemplarisch davon ausgehend, dass ein Sprachmodell dazu generiert und trainiert werden soll, um in einem Bildungsprogramm für Migrant:innen, den Schüler:innen aus verschiedenen Ländern personalisierte Lerninhalte bereitzustellen, müsste für einen ausgewogenen Datensatz sichergestellt werden, dass das Modell eine gleich verteilte und damit ausgewogene Menge an Daten zu jedem relevanten Migrationshintergrund erhält. Zu beachten ist, dass Ausgewogenheit hier lediglich bedeutet, alle relevanten Gruppen des Trainingsdatensatzes im Verhältnis gleich einzubeziehen, um so die Über- oder Unterrepräsentation einzelner Gruppen zu vermeiden. Das bedeutet nicht zwingend gleichermaßen eine allumfassende bzw. die Gesellschaft vollständig abbildende Datenerfassung.

16 Als erstes Medienunternehmen hat die New York Times daher sowohl OpenAI als auch Microsoft auf Grund von Urheberrechtsverletzung verklagt (vgl. Freeman et al. 2024).

Entscheidend ist, welches Ziel die jeweilige Modellierung hat und welche weiteren Kriterien in den Einsatz gebracht werden. Um bei dem Beispiel des Glossars zu bleiben: Wenn ein Unternehmen ermitteln möchte, welche Produkte Kund:innen, die über 30 Jahre alt sind, präferieren, ist es wenig zielführend, die Altersgruppe unter 30 in die Modellierung einzubeziehen. Auf Große Sprachmodelle übertragen hieße das, wenn die Zielgruppe des Sprachmodells ausschließlich aus französischsprachigen Herkunftsländern stammt, ist es vermutlich weniger zielführend, die spanische Sprache mittels entsprechender Trainingsdaten in das System einzupflegen.

Damit ein Datensatz *divers* ist, muss, gemäß dem KITQAR-Glossar, „jede Entität der Domäne in der Datenmenge repräsentiert" sein, also jeder Entitätstyp der Gesamtmenge mindestens einmal vorkommen (Mohammed et al. 2023: 3). Es geht bei diesem Kriterium nicht um eine möglichst ausgewogene Verteilung, sondern darum, alle Einzelentitäten der vorhandenen Gesamtmenge zu inkludieren. Wenn also Mitarbeitende einer Firma anhand ihres Alters erfasst werden sollen, dann muss jedes Alter, zu dem es mindestens einen Mitarbeitenden gibt, Teil der Trainingsdatenmenge sein. Das BSI wiederum fasst diese Ebene unter dem Begriff der Vielfalt und zielt darauf ab, die Varianz der Datenmengen zu maximieren. Der Leitfaden verweist an dieser Stelle sogar explizit auf Sprachmodelle und argumentiert, dass die Vielfalt der menschlichen Sprache – explizit im Sinne von Dialekten, Jargons und Akzenten – in Sprachmodellen implementiert sein sollten. Die beiden Kriterien *Ausgewogenheit* und *Diversität* können deckungsgleich sein, müssen es aber nicht. Eines von beiden an die jeweiligen Trainingsdaten anzulegen, bedeutet daher nicht unbedingt, dass auch das andere erfüllt ist.

Eng verwoben mit den beiden Kriterien Diversität und Ausgewogenheit listen sowohl KITQAR als auch das BSI *Repräsentativität*. Dieses meint, dass jede Entität der zu repräsentierenden Gesamtmenge die gleiche Chance hat, in der jeweiligen Datenmenge repräsentiert zu sein (vgl. Mohammed et al. 2023). Ein repräsentativer Datensatz spiegelt also die statistischen Verteilungsverhältnisse der realen Gesamtmenge wider. Dabei sollte allerdings kritisch mitbedacht werden, dass es vorkommen kann, dass durch einen statistisch repräsentativen Datensatz das daraufhin trainierte Modell für einzelne Entitäten weniger geeignet ist, da nicht ausreichend repräsentative Daten vorhanden sind. Wenn beispielsweise ein Sprachmodell Studierenden Fragen zu typischen Herausforderungen im Universitätskontext beantworten können soll, kann es durchaus sein, dass ein auf repräsentativen Daten trainiertes Modell weder für Personen im Rollstuhl

noch für Personen unter 16 Jahren nützliche Antworten ausgeben kann, da diese nur einen geringen Anteil der Studierenden ausmachen und daher in Relation weniger Einfluss auf die Modellbildung genommen haben. Das BSI definiert den Begriff wiederum mit Rückbezug auf und in Abgrenzung zu Konzepten wie Ausgewogenheit, Vielfalt oder Gültigkeit. Der Leitfaden nutzt das Beispiel eines Systems, das dafür eingesetzt wird, die Kreditwürdigkeit einzelner Kund:innen zu errechnen. Mit Blick auf die Trainingsdatenqualität wird exemplifiziert: „Um die Repräsentativität zu gewährleisten, müssen die Trainingsdaten eine breite und faire Strichprobe der gesamten Bevölkerung enthalten" (BSI 2025: 7).

Fairness befasst sich im Diskursfeld rund um KI-Anwendungen in der Regel mit der Identifizierung, Analyse und Quantifizierung von Verzerrungen (Bias), die Individuen und Gruppen nach bestimmten Merkmalen wie Geschlecht, Ethnie oder Behinderung unzulässig diskriminieren (vgl. AIEIG 2020; Heesen et al. 2021; Mehrabi et al. 2021). Systeme können beispielsweise dann als fair bezeichnet werden, wenn ihre Trainingsdaten die Standards zur Freiheit von diskriminierenden Verzerrungen erfüllen oder die darauf basierenden Systeme keine diskriminierenden Inhalte ausgeben. Fairness ist als Begriff jedoch durchaus heterogen konzipiert. Von einem fairen Sprachmodell zu sprechen, kann auch meinen, dass keine ethischen Werte missachtet oder Personen in der Generierung oder Nutzung diskriminiert oder ausgebeutet wurden. Fairness kann also auf die Qualität der Daten, die Bedingungen ihrer Produktion oder die Bedingungen ihrer Anwendung oder aber auf die Fairness der schlussendlichen Modellierungen abzielen.

Ähnlich breit anwendbar ist die Ebene von Privatsphäre und Datenschutz. Als entscheidende Grundlage für die Generierung der Modellierungen existiert ein großes Interesse daran, möglichst viele Daten über möglichst viele Bereiche zu haben. Dabei gleichermaßen Datenschutz und Privatsphäre zu wahren – insbesondere, wenn globale Perspektiven und länderspezifisch unterschiedliche Regulierungen und Verständnisse zu Bedenken sind – birgt eigene Herausforderungen. Dies beginnt bei den eingangs aufgeworfenen Fragen danach, an welcher Stelle und in welchen Zusammenhängen welche Daten erhoben werden sollen oder dürfen. Und auch wenn die Erhebung datenschutzkonform geschieht, existieren weitere Herausforderungen. So kann unzureichender Schutz der Daten zu Missbrauch oder Verlust der Privatsphäre führen. Neben den Diskussionen darum, welche Inhalte von KI-Systemen aus dem Netz oder auch während

der konkreten Nutzung abgegriffen und für das Training genutzt werden dürfen, besteht auch die Gefahr, dass gespeicherte Inhalte beispielsweise aufgrund von Datenlecks durch externe Parteien zugegriffen und weiter genutzt werden können. Wenn Große Sprachmodelle für medizinische Diagnosen oder für Formen der Gesprächstherapie genutzt werden, fließen in die Systeme höchst private, persönliche und vor allem je nach Auswertungslogik vulnerable Informationen ein, die anschließend abgegriffen werden können. Dies führt wiederum zu neuen Formen der mangelnden Kontrolle über die eigenen Daten. Dieser Aspekt gewinnt an Relevanz, wenn man die Nutzung der Systeme durch vulnerable Gruppen in den Blick nimmt, denen aus diversen Gründen vielleicht nicht möglich ist, zu erfassen, welche Folgen die Preisgabe intimer Daten haben kann. OpenAI ist sich dieser Problematik durchaus bewusst. Fragt man ChatGPT selbst, welche Gefahren darin liegen können, intime Informationen mit dem System zu teilen, gibt dieses an, die Probleme reichten von unzureichendem Datenschutz oder nicht ausreichend geschützter Privatsphäre über das möglicherweise überhöhte Vertrauen in die Fähigkeiten des Systems bis hin zu emotionaler Abhängigkeit seitens der Nutzer:innen, gekoppelt mit fehlender Empathie seitens des Systems.[17] Obgleich das System seinen Nutzer:innen also auf Nachfrage durchaus erklärt, dass es problematisch sein kann, persönliche oder intime Daten einzugeben, hat OpenAI nur begrenzt Kontrolle darüber, welche Inhalte Nutzer:innen in das System einspeisen. ChatGPT reagiert, sobald die Eingabe gemacht wurde, die Daten von dem System also bereits verarbeitet und damit sowohl digital übermittelt als auch in das System eingespeist worden sind. Dazu kommt, dass auch hier ein Trade-Off zwischen einzelnen Kriterien adressiert werden kann. Denn der (Daten)Schutz vulnerabler Gruppen kann zur Folge haben, dass die Daten einzelner vulnerabler Gruppen nicht in das Training der jeweiligen Modelle mit einbezogen werden (können). Dies wiederum vermindert möglicherweise die Diversität des Modells und kann auch bedeuten, dass die durch das Systeme ermöglichten Unterstützungsangebote – wie das Übersetzen von Inhalten – nicht für alle Bevölkerungsgruppen gleichermaßen effizient ausgearbeitet werden können.

Mögliche Maßnahmen, um diesen Daten-Problemen entgegenzuwirken, sind die Implementierung von Sicherheitsmaßnahmen wie Verschlüsselungen und Anonymisierungen, DSGVO-konforme Datenverarbeitung sowie

17 Diese Informationen sind die Antwort von ChatGPT auf Basis des Prompts „Welche Konsequenzen kann es haben, dir zu intime Informationen zu teilen?“.

transparente Datenschutzerklärungen und Opt-Out Optionen, oder auch die Reduktion der Datensammlung auf ein Minimum. Ebenfalls dazu beitragen kann die Einhaltung und Überprüfung der bis hierhin exemplarisch aufgelisteten Kriterien für Trainingsdatenqualität. Der exemplarische Blick auf einzelne Kriterien hat bereits gezeigt, dass einzelne ethische Standards im Widerspruch zueinanderstehen, sowie – je nach Anwendungskontext – in ihrer Relevanz variieren können. Ein erstes Fazit kann daher sein, dass es nicht nur entscheidend ist, die Kriterien stets bedacht und reflektiert in den Einsatz zu bringen, sondern es durchaus auch interdisziplinäre und anwendungsfallbezogene Diskussionen darum geben muss, welche ethischen Standards jeweils in Einsatz gebracht werden können oder müssen. Das wiederum zeigt auch, dass es nur schwer eine Automatisierung solcher Standards geben kann.

6. Fazit

Die bis hierhin schlaglichtartige Betrachtung ethischer Perspektiven hat deutlich gemacht, dass ethische Standards für Trainingsdatenqualität ein wichtiger Ansatz sein können, um konkrete Risiken von Sprachmodellen abzuschwächen. Nichtsdestotrotz bleibt es dabei, dass Sprachmodelle in ihrer Nutzung auch bei Beachtung von Trainingsdatenqualität sowohl in ihrer systeminhärenten Logik, ihren konkreten Anwendungsmomenten als auch in ihrer allgemeinen gesamtgesellschaftlichen Bedeutung durchaus negative Wirkmacht entfalten können. Die Nutzung sowie die Generierung der Systeme sollte daher stets unter Berücksichtigung der mit dieser Medientechnologie verbundenen Konsequenzen durchgeführt werden. Dies bedeutet einerseits, dass weiterhin kritische Untersuchungen der Nachhaltigkeit der Systeme aus ethischer Sicht von Relevanz sein sollten, ebenso wie der Blick auf die hinter den Systemen liegenden menschlichen Arbeit oder mit Blick auf die gesellschaftlichen Veränderungen, die sie auslösen. Darüber hinaus scheint es ebenfalls notwendig, die konkreten Nutzungspraktiken zu adressieren, die Große Sprachmodelle ermöglichen. Insbesondere, wenn diese möglicherweise negative Folgen für die einzelnen Nutzenden haben können. Wichtig zu untersuchen wäre darüber hinaus auch, inwieweit sich mit Blick auf die zum Teil abweichende Realität des Trainings von Großen Sprachmodellen, die jeweiligen Kriterien der Trainingsdatenqualität überhaupt produktiv anwenden lassen können. Wenn Systeme wie ChatGPT ihre Daten über Scraping abgreifen und mit riesigen

Datenmengen arbeiten, lassen sich dann Kriterien wie Diversität, Fairness und Ausgewogenheit noch adäquat anbringen? An dieser Stelle ließe sich auch fragen, wo die Grenzen einer ethischen Betrachtung liegen oder liegen können und ob es vielleicht Bereiche gibt, in denen es sinnvoll ist, entsprechende Systeme gar nicht erst in Anwendung zu bringen.

Literatur

Afshar, Melissa Fleur (2024): People Are Using ChatGPT to Help Them Achieve Their 'Dream Life', in: Newsweek, 05. November 2024 (online unter: https://www.newsweek.-com/people-using-chatgpt-help-achieve-dream-life-1979801 – letzter Zugriff: 5.11.2025).

AI Ethics Impact Group (AIEIG) (2020): From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, Gütersloh.

Ayre, Julie / Cvejic, Erin / McCaffery, Kirsten J. (2025): Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey, in: The Medical Journal of Australia 222 (4/2025), S. 210–212.

Bohannon, Molly (2023): Lawyer Used ChatGPT in Court and Cited Fake Cases. A Judge Is Considering Sanctions, in: Forbes, 08. Juni 2023 (online unter: https://www.forbes.com-/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ – letzter Zugriff: 5.11.2025).

Bundesamt für Sicherheit in der Informationstechnik (2025): QUAI-DAL. Teildokument B: „02-Qualitätskriterien & Bausteine", in: Bundesamt für Sicherheit in der Informationstechnik, 01. Juli 2025 (online unter: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/-QUAI-DAL_B_Qualitaetskriterien.pdf?__blob=publicationFile&v=4 – letzter Zugriff: 5.11.2025).

Chun, Wendy Hui Kyong (2021): Discriminating Data. Correlation, Neighborhoods, and the New Politics of Recognition, Cambridge, MA.

Crawford, Kate (2021): The Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven, London.

Distelmeyer, Jan (2025): Mit KI zu tun bekommen – Daten, Arbeit und Interfaces: Was können wir von Plattformen wie ChatGPT (und sie von uns) wissen?, in: cargo 65 (3/2025), S. 5.

European Research Area Forum & Directorate General for Research and Innovation (2024): Living guidelines on the responsible use of generative AI in research, in: Europäische Kommission, 15. April 2024 (online unter: https://european-research-area.ec.europa.eu/news-/living-guidelines-responsible-use-generative-ai-research-published – letzter Zugriff: 5.11.2025).

Europäische Union (2024): Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 über künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union, in: Amtsblatt der Europäischen Union, L 206, S. 1–161 (online unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32024R1689 – letzter Zugriff: 5.11.2025).

Freeman, Joshua et al. (2024): Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 Lawsuit, in: arXiv, 09. Dezember 2024 (online unter: https://doi.org/10.48550/arXiv.2412.06370 – letzter Zugriff: 4.8.2025).

Gallotta, Roberto et al. (2024): Large Language Models and Games: A Survey and Roadmap, in: arXiv, 09. Dezember 2024 (online unter: https://doi.org/10.48550/-arXiv.2402.18659.18659v1 – letzter Zugriff: 4.8.2025).

Gray, Mary L. / Suri, Siddharth (2019): Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass, Boston, New York.

Heesen, Jessica / Reinhardt, Karoline / Schelenz, Laura (2021): Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine digitale demokratische Gesellschaft, in: Gero Bauer et al. (Hg.), Diskriminierung und Antidiskriminierung. Beiträge aus Wissenschaft und Praxis, Bielefeld, S. 129–148.

Hiller, Anna / Maristany de las Casas, Pablo (2025): Generative KI und die deutsche extreme Rechte. Narrative, Taktiken und digitale Strategien (online unter: https://isdgermany.org/-generative-ki-und-die-deutsche-extreme-rechte-narrative-taktiken-und-digitale-strategien/ – letzter Zugriff: 5.11.2025).

Lee, Ian (2024): 4 Theses on Boyfriend Dan GPT, in: The Last Organizer, 31. März 2024 (online unter: https://medium.com/thelastorganizer/4-theses-on-dan-gpt-98bb2a682b5b – letzter Zugriff: 5.11.2025).

Liu, Yi et al. (2023): Jailbreaking ChatGPT via Prompt Engineering: An empirical study, in: arXiv, 23. Mai 2023 (online unter: https://arxiv.org/abs/-2305.13860 – letzter Zugriff: 5.11.2025).

Loh, Wulf (2024): Generative KI, digitale Teilhabe und epistemische Ungerechtigkeit, in: RphZ – Zeitschrift für Religion, Politik und Gesellschaft 10 (2/2024), S. 215–233.

Mehrabi, Ninareh et al. (2021): A survey on Bias and Fairness in Machine Learning, in: ACM Computing Surveys 54 (6/2021), Article 115.

OECD (2022): Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint (= OECD Digital Economy Papers, No. 341), Paris.

OpenAI (2025a): AI boyfriend (online unter: https://chatgpt.com/g/g-RwZG2pDs2-ai-boyfriend – letzter Zugriff: 5.11.2025).

OpenAI (2025b): TherapyAI (online unter: https://chatgpt.com/g/g-8yHB0UD8j-therapyai – letzter Zugriff: 5.11.2025).

Paris, Martine (2025): ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO, in: Forbes, 13. April 2025 (online unter: https://www.forbes.com/sites/martineparis/-2025/04/12/chatgpt-hits-1-billion-users-openai-ceo-says-doubled-in-weeks/– letzter Zugriff: 5.11.2025).

Heesen, Jessica et al. (2023): Künstliche Intelligenz im Journalismus. Potenziale und Herausforderungen für Medienschaffende. Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2023-1

Raile, Paolo (2024): The usefulness of ChatGPT for psychotherapists and patients, in: Humanities and Social Sciences Communications 11, Article 47.

Reyes, Marta (2025): Why You Shouldn't Say "Thank You" and "Please" to ChatGPT, in: Medium, 24. April 2025 (online unter: https://medium.com/@martareyessuarez25/why-you-shouldnt-say-thank-you-and-please-to-chatgpt-2910da23b3f3 – letzter Zugriff: 5.11.2025).

Scheiter, Katharina et al. (2025): Künstliche Intelligenz in der Schule. Eine Handreichung zum Stand in Wissenschaft und Praxis, hrsg. im Rahmen des KI-Begleitprozesses im Rahmenprogramm empirische Bildungsforschung, Bonn (online unter: https://www.empirische-bildungsforschung-bmbfsfj.de/img/KI_Review.pdf – letzter Zugriff: 5.11.2025).

Mohammed, Sedir et al. (2023): Ein Glossar zur Datenqualität (1.2), in: Zenodo, 06. März 2023 (online unter: https://doi.org/10.5281/zenodo.7702426 – letzter Zugriff: 5.11.2025).

Shaker, Noor / Togelius, Julian / Nelson, Mark J. (2016): Procedural Content Generation in Games, Cham.

TED (2025): OpenAI's Sam Altman talks ChatGPT, AI agents and superintelligence – Live at TED2025, in: YouTube, 11. April 2025 (online unter: https://www.youtube.com/watch?v-=5MWT_doo68k – letzter Zugriff: 5.11.2025).

Thompson, Tommy (2024): The Changing Landscape of AI for Game Development, in: Paul Roberts (Hg.), Game AI Uncovered, Boca Raton, S. 1–11.

Yannakakis, Georgios N. / Togelius, Julian (2024): Artificial Intelligence and Games, 2. Aufl., Cham.

Imagining Fair(er) Datasets for GenAI: Lessons from the Arts

Theresa Krampe[1]

Abstract

As generative artificial intelligence (GenAI) is rapidly inserting itself into many domains of everyday life, there is also a growing awareness of its ethical implications. Several systems, among them chatbots and image generators, have been shown to reiterate gendered, racial, or ableist stereotypes and to contribute to the erasure of marginalised voices and perspectives. In machine learning and AI ethics, concepts such as fairness and algorithmic bias have become instrumental in recognising and mitigating these issues. The task of addressing covert, structural, and multilayered forms of discrimination, however, remains challenging. In this chapter, I argue that the arts and culture, as domains that tend to be overlooked in mainstream discussions around GenAI, offer valuable inspiration for envisaging more diverse and inclusive datasets for fair(er) AI systems. With the help of two case studies—*The Zizi Project* by Jake Elwes and *Not the Only One* by Stephanie Dinkins—I show how AI artworks can draw attention to the risks of GenAI to unfairly discriminate against so-called vulnerable groups, challenge the values and assumptions underlying hegemonic visions of technology, and draft alternative AI futures.

1. Introduction: Imagining Fair(er) AI

Machine learning (ML) technologies are rapidly insinuating themselves into everyday life. The use of advanced search engines, digital assistants, or even automated decision-making software has already become quite natural to many of us. With the emergence of high functioning and highly visible Large Language Models (LLMs) such as GPT as well as image generators such as Midjourney or DALL-E, the same is increasingly true for generative artificial intelligence (GenAI). Over the past few years, the ability of

1 https://orcid.org/0009-0001-9416-4676

these systems to create an astonishing variety of texts, images, or sounds reinvigorated discussions around the potential of AI for innovation and creativity (see Cooper 2020; Eapen et al. 2023; Marr 2023), but also raised concerns about privacy (see Hagendorff 2019), authorship and copyright (see El Atillah 2023), misinformation (see Hsu and Thompson 2023), and job losses in the creative industries and other areas that formerly seemed automation-proof (see Ellingrud et al. 2023; Fleming 2024; Verma and De Vynck 2023). No less importantly, numerous studies have indicated that AI systems can be, and often are, subject to systematic biases that unjustly discriminate against groups of people on the grounds of race, class, gender, disability, or other protected attributes (see Akter et al. 2021; Barocas et al. 2023: 19–20; Hagendorff 2019; Heesen et al. 2021; Mehrabi et al. 2021a). Even though chatbots and image generators seldom act as immediate gatekeepers in such high-stakes scenarios as criminal punishment, medical diagnoses, or granting loans (see Angwin et al. 2016; Barocas and Selbst 2016), they can nevertheless discriminate against social groups if they perpetuate stereotypes, privilege certain types of knowledge, create disrespectful or demeaning output, and reinforce unfair regimes of in/visibility and marginalisation along the lines of identity (see Barocas et al. 2023; Gautam et al. 2024; Loh 2024; Mehrabi et al. 2021a). Words and images, after all, shape the way we make sense of the world and our place within it.

This chapter examines the discourse around fairness and/as discriminatory bias in AI ethics, as well as how these understandings are currently being renegotiated in the arts and culture. Artworks, activist interventions, and community-led projects have not yet received sufficient scholarly recognition in the field of ML. As a result, their potential to interrogate the values and assumptions underlying ML and their unique capacity to imagine fair(er) datasets and fair(er) uses of GenAI has remained largely untapped. By considering artworks in the context of AI ethics, I thus seek to shift the discussion around fairness and bias to forms of intervention that appropriate GenAI in unconventional, creative, and often subversive ways. Art, I argue, can draw attention to the risks of AI to unjustly discriminate against so-called vulnerable groups, challenge the values and assumptions underlying hegemonic visions of AI, and draft alternative AI futures, thereby offering valuable prompts for creators, users, and critics of GenAI.

The main part of this chapter is divided into two sections. Following immediately after this introduction, Section 2 offers an in-depth discussion of discriminatory bias in GenAI from an interdisciplinary ethical perspective, focussing on the idea(l) of fairness as freedom from discriminatory

bias as an important measure of training data quality, while also examining how these biases may impact social groups and power structures. Section 3 then shifts the perspective to interventions and potential solutions as they are envisioned in contemporary art. To do so, I discuss two case studies: *Zizi—Queering the Dataset* (2019) from the *Zizi Project* by London-based conceptual artist and researcher Jake Elwes and *Not the Only One* (2018–ongoing) by the US-American transdisciplinary artist Stephanie Dinkins. Both works use GenAI trained on diversified and carefully curated datasets as part of an artistic practice that challenges hegemonic formations of gendered and racialised identity. To analyse them, I integrate AI ethics with queer/feminist, intersectional, and media-analytical approaches.

2. Bias and Discrimination in GenAI

2.1 Fairness as Freedom from Discriminatory Bias

Fairness, in a general sense, can be defined as the impartial and just treatment of people. In Western societies, this is usually associated with the idea of equal chances for self-advancement, for example, that a person's position in society should be proportionate to their contribution, rather than to factors they have no control over (see Feuerriegel et al. 2020: 380–381). It is thus important to note that the un/fairness of AI potentially relates to a much wider range of pressing topical issues than what is addressed in this chapter, including AI's involvement in neo-colonial relations of subjugation and extractivism; practices of datafication and (racialised) surveillance; the exploitation of natural resources; or precarious labour in the Global South, among others (see Tacheva and Ramasubramanian 2023). In AI ethics, however, fairness has become strongly associated with the notion of discriminatory bias, understood as systematic distortions in an AI system that result in the unfair differential treatment of people based on socially relevant groups, usually on the grounds of protected attributes such as race, class, gender, ability, or sexual orientation (see Akter et al. 2021; Barocas et al. 2023; Hagendorff 2019; Kim 2022; Leavy et al. 2020; Mehrabi et al. 2021a). This focus on bias, and on datasets as sources of bias but also, crucially, points of intervention, is also echoed in the case studies analysed in this chapter, and hence I will mainly limit my discussion to fairness in this latter sense.

That ML can entail discriminatory biases first came to widespread attention around ten years ago, in a series of highly publicised incidents.

When Google's photo app assigned the label "gorilla" to People of Colour (PoCs) in 2015, this not only revealed inaccuracies in the algorithm, but also invoked rather nasty racist discourses. A few years later, an influential study by Joy Buolamwini and Timnit Gebru (2018) found biases regarding skin colour and gender in commercial classification software, which reinforces highly problematic forms of intersectional oppression and prolongs a history of marginalisation and disenfranchisement of Black women. Other well-known examples include translation software associating certain professions with gendered stereotypes (see Prates et al. 2020) or reiterating societal prejudices towards sensitive attributes such as gender, race, and sexual orientation (see Lin et al. 2023).

In academia, the occurrence of discriminatory bias in AI has motivated the development of an entire field dedicated to "uncovering and rectifying [...] biases in statistical and machine learning models" (Mitchell et al. 2021: 142).[2] In the corresponding publications, fairness is typically defined *ex negativo* and with reference to anti-discrimination: A model is fair if it does not show "prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al. 2021a: 1; see also Sun et al. 2025).[3] Scholars have furthermore pointed out that AI can discriminate against people even if it treats all groups equally, for instance if it employs variables that are correlated with group membership such as height for gender, language for nationality, and zip code for race or class (see Mehrabi et al. 2021a; Sun et al. 2025). This is why "fairness through unawareness"—the somewhat naïve view that discrimination can be avoided by excluding sensitive attributes from the ML process – is rightly deemed unsuitable for addressing discriminatory bias in most cases. Not only can removing information about race or gender diminish the model's performance,[4] but

2 This is also evidenced by the fact that fairness has become a mainstay for international conferences such as FAcct, ICML, or AAAI (Wang et al. 2022). Recurring labels applied to the research field include fair AI, fair ML, or algorithmic fairness. To my knowledge, there are no established definitions of or clear delimitations between these labels, so I will use them roughly synonymously in this chapter.

3 Note that discrimination can be justified under certain circumstances, for instance, if it is unavoidable or if the information is relevant for the decision (e.g., car insurance rates that vary with the age of the holder; see Barocas et al. 2023). Another interesting example is algorithmic affirmative action, i.e., the idea that algorithms could be trained in such a way as to counterbalance structural disadvantages for marginalised groups (see Kim 2021 for a critical legal perspective; Segev 2025 for an ethical one).

4 Though, by now, there are several technical fixes for this particular problem, as summarised by Hagendorff: "Advanced machine learning methods can learn from small

it is also ineffective as these attributes tend to re-enter the model by proxy. In the worst case, such colour or gender-blind approaches even exacerbate the problem because they make the model's biases harder to detect (see Brandner and Hirsbrunner 2023: 27; Feuerriegel et al. 2020: 379; see also Kim 2022; Mehrabi et al. 2021a).

2.2 The Role of Training Data Quality

A crucial factor for the formation (and hence the prevention) of bias in ML is the quality of the data used to train the model—though it is worth noting that training data quality is not the only source of bias.[5] More often than not, AI systems learn prejudices prevalent in society from the training dataset, which typically consist of large amounts of texts or images scraped from the internet. Pre-existing human prejudices can become encoded into AI applications because certain groups are over or underrepresented in the dataset, or because the training data replicates stereotypes. Discriminatory aspects can furthermore be introduced via the annotation of data by service providers, or through user feedback (see Barocas et al. 2023: 248–252; Brandner and Hirsbrunner 2023: 26; Feuerriegel et al. 2020: 381; Hagendorff and Wezel 2020: 358). As ML processes rely on generalisations and pattern recognition, such stereotypes are not only adopted but are in fact reinforced by GenAI. If a system defaults to male pronouns for "lawyers" or "doctors," for instance, this further increases the frequency with which these professions are associated with men in the dataset. Considering the importance of data to make or break this vicious circle, it is not coincidental that the art projects analysed below engage specifically with questions of training data quality, pointing to gaps in the dataset that lead to discriminatory output, but also imagining ways of using data in such a

datasets via data augmentation, can generate synthetic data via GANs or variational autoencoders to artificially increase the amount of training stimuli, use transfer learning to use knowledge from an already learned task, utilize few shot learning mechanisms, etc." (2021: 567).

5 In a comprehensive survey, Mehrabi and colleagues (2021a) identify multiple sources of bias along the entire ML process, from data collection via the algorithm to the interaction with the users, paying attention to how, among other things, a company's hiring practices, the design of the algorithm, or the presentation of information via the graphical user interface may influence a model's biases (see also Chowdhury and Mulani 2018; Zou and Schiebinger 2018).

way as to promote epistemic justice or, in any case, interrogate established epistemic hierarchies.

While in-depth scholarly engagements with training data quality and discrimination are still comparatively rare, by now, there is some agreement on criteria that are deemed conducive to increasing the quality of training data.[6] Some of these seem particularly relevant for reducing the risk of discriminatory outputs. The completeness of the dataset, as a case in point, touches upon matters of exclusion, as when certain groups are absent from the dataset due to persistent social dynamics of marginalisation. Inclusion in the dataset can be a double-edged sword, however. On the one hand, some AI applications show significantly poorer performance for some groups of people because the available training data is incomplete or insufficient. On the other hand, collecting additional data may put a burden on the groups in question if they are being "overresearched" or if their increased visibility also increases their vulnerability (see Benjamin 2019; Eubanks 2018). The diversity of the dataset, too, seems worth considering. If diversity is taken to mean that the dataset contains at least one type of each entity present in the overall group, this would ensure the representation of minority groups even though they only comprise a very small percentage of the relevant population (see Mohammed et al. 2023). However, this also means that there may be trade-offs between diversity and representativeness if the latter is taken to mean that the dataset should accurately reflect the population represented. What is more, diversity comes with its own set of challenges as it entails tricky questions of social categorisation. While it may be desirable for a dataset to distinguish between multiple subgroups, for example, to promote intersectional fairness (see below), the same strategies are also prone to pigeonholing and masking the contingency and constructedness of social groups as such.

As even such a selective discussion shows, there are several different dimensions of training data quality that significantly impact bias and dis-

6 Meta studies (e.g., Hagendorff 2019; 2021) as well as data quality tools such as the data quality glossary (Mohammed et al. 2023; see also Brandner et al. 2023) offer useful overviews of relevant criteria. Recurring criteria include correctness (does the dataset contain errors?); transparency (what information is available about the data, e.g., its origin, purpose, or the quality assurance measures employed?); timeliness (is the data up to date?); relevance (is the data and metadata relevant to the purpose?) as well as more complex criteria such as explainability and credibility. From a normative perspective, Hagendorff furthermore explores the idea of "'good' behavioral datasets" (2021: 564), i.e., training data that is chosen from a subset of the population whose behaviour is deemed both competent and morally sound.

crimination. Not all of these can necessarily be satisfied in equal measure, making it impossible to create a one-fits-all solution. What is more, most approaches to fairness in ML do not (or not sufficiently) account for the complexity of the socio-cultural contexts from which AI systems emerge and within which they operate. The underrepresentation or stereotyping of specific groups in the training data is often the result of complex structural forms of marginalisation and historical injustice, not all of which are easy to detect, let alone remedy (see Costanza-Chock 2020). Therefore, most fairness tools are ill-prepared to address the root causes of discrimination or to consider the indirect and long-term effects of predictions and decisions (see Thomsen 2024). As Anna Lauren Hoffmann puts it, there is no "easy fix" (Hoffmann 2019: 910) for structural discrimination that can simply be applied at the level of code. Quite the opposite, fairness tools could actually detract from the need to interrogate social power structures (see Leavy et al. 2020: n.pag.). What is more, if we take seriously the idea of fairness as equal opportunity, we may also want to think about correcting for systemic structural disadvantages and historical injustices (see Barocas et al. 2023; Binns 2018; Crawford 2017). In the field of GenAI, this could mean finding ways of reinserting the voices of women, PoCs, and those who have been systematically erased from the archive into the dataset, even at the expense of values like representativeness or accuracy.

In closing this section, it is worth stressing once more that data quality and the absence of bias are not the only understandings of fairness that are potentially relevant to the question of discrimination. Fairness may also concern matters of legitimacy (see Mitchell et al. 2021: 143). Indeed, the question of whether or not, in a given situation, it is justified to use ML in the first place ought to precede concerns about algorithmic bias (see Barocas et al. 2023: 23–24). Writing from an intersectional queer/feminist perspective, Katrin Köppert maintains that even methods like fair and explainable AI often fail question the assumptions behind the AI imaginaries they advance. In other words: Making AI fair(er), more transparent, more trustworthy, etc. does not automatically ensure a project's usefulness or ethical merit. This is especially true when considering AI "in the bigger picture of the climate crisis, extractivism, and machismo" (Köppert 2024: n.pag.); or in its complicity with those hegemonic power relations and mechanisms that Jasmina Tacheva and Srividya Ramasubramanian (2023) aptly subsume under "AI Empire."

2.3 Setting the Terms of In/Exclusion

Understanding the moral stakes of algorithmic bias requires us to further unpack how GenAI is implicated in harmful and unjust regimes of in/visibility and in/exclusion. From an ethical perspective, the notion of representational harms (see Crawford 2017) seems helpful when it comes to articulating when and why biases in GenAI are harmful. In contrast to allocative harms, which concern the distribution of resources and opportunities (see Barocas et al. 2023: 19–20), representational harms can be understood as "harms [that] arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognise their existence altogether" (Blodgett et al. 2020: 5455). Hence, representational harms are concerned with the relations between real-world social hierarchies and (verbal, audiovisual, etc.) representations of groups of people. In particular, this concerns positive or negative stereotypes, the use of denigrating language or imagery, the erasure of minority identities, Othering and dehumanisation, as well as the naturalisation of social categories more generally (see Barocas et al. 2023: 19–20; Crawford 2017: n.pag.; Dev et al.: n.pag.; Mehrabi et al. 2021b: 5017).

To further specify the nature of the harm done, it is also worth considering forms of epistemic injustice: the wrongs "done to someone specifically in their capacity as a knower" (Fricker 2007: 1).[7] Examples may include the misrepresentation, dismissal, or silencing of a person's knowledge and contribution to discourse. Analysing the consequences of ML for digital participation, Wulf Loh cautions that GenAI runs the risk of creating a new kind of digital divide that exacerbates injustice and discrimination. Models trained by conventional means, i.e., by scraping data from existing archives such as the internet, are not only conservative by default but also likely to "overlook speech acts, knowledge domains or cultural artifacts underrepresented on the web, such as minority languages or art collections" (Loh 2024: 215). The resulting misrepresentation or erasure of marginalised

7 With Miranda Fricker, we may further distinguish between testimonial injustices that occur when a speaker is given less credibility because of their social identity, and hermeneutical injustices that occur if a social group is excluded from the construction and negotiation of social values and meanings (2007: 1, 6). Both occur in the context of GenAI, for instance if the system reproduces identity prejudices ("women are irrational") that undermine their credibility (testimonial injustice) or if it prevents people from articulating their identity and experience and having that experience be part of the dataset (hermeneutical injustice). See Loh (2024) for a detailed discussion.

groups constitutes an instance of epistemic injustice because it disregards the identity of the members of said groups, discounts them as competent knowers, and ultimately renders their experience unintelligible. At the same time, this dynamic of in/exclusion is also detrimental for the majority group. Biased GenAI typically privileges culturally dominant knowledges while withholding alternative ways of knowing, thus preventing new ideas from entering the discourse (see Fricker 2007: 43–44). In these cases, GenAI not only fails to redress but in fact exacerbates social bias and epistemic injustice by further marginalising minority voices and barring their experiences from the dataset.

Similar arguments have long been articulated within the field of Postcolonial Studies. The concept of subaltern articulation, theorised by Gayatri Chakravorty Spivak in her influential essay "Can the Subaltern Speak" (1988), chimes particularly well with concerns around the epistemic in/justices produced by GenAI in that it highlights asymmetries in who is granted voice and intelligibility within systems of knowledge and power. Occupying a position of radical marginalisation, the subaltern is structurally excluded from hegemonic discourse and unrepresentable within dominant epistemic frameworks. Spivak's conclusion that the subaltern cannot speak—arguably less an expression of fatalism than a critique of "Western" attempts to speak on behalf of the subaltern—also offers an important problematisation of contemporary practices around AI ethics: Considering the structural, historical, and political conditions that render certain voices in/audible or un/intelligible, it is insufficient for powerful tech companies to introduce ethics reviews or to diversify their datasets. Quite the contrary, such gestures risk reproducing the very hierarchies they claim to challenge, thus perpetuating forms of epistemic injustice. Similar criticisms can also be levelled against the field of AI ethics itself as long as the discussion remains biased towards Western values and epistemic traditions, sidelining scholarship from the so-called Global South as well as minority voices in the Global North (see Roche et al. 2021; 2023; Segun 2021).

Finally, any analysis of discriminatory bias in AI would do well to pay attention to the complex ways in which racism, sexism, ableism, etc. overlap and combine to produce very specific forms of social inequality. What makes these forms of intersectional discrimination particularly insidious is that their cumulative effects amount to more than the sum of their parts: As Kimberlé Crenshaw (1989) has famously shown, even if a person is neither discriminated against on the grounds of her gender, nor on the

grounds of her race, she may nevertheless experience discrimination as a *Black woman.* Consequently, anti-discrimination laws based on single-axis thinking cannot protect individuals from intersectional forms of injustice and oppression—and neither can single-axis approaches to bias in AI (see Ciston 2019; Collins 2017; Costanza-Chock 2020; Hoffmann 2019; Kong 2022; Noble 2018).[8] Recognising the importance of intersectionality, ML researchers have since proposed criteria for an intersectional approach to fairness, demanding that fair AI should consider multiple attributes and protect minority groups by not only protecting each attribute value but also all intersecting values (see Foulds et al. 2020; Kearns et al. 2018) Modelling these criteria, however, remains challenging, not least because the uniqueness, complexity, and sheer number of intersecting forms of discrimination make them difficult to compute (see Fosch-Villaronga and Malgieri 2024; Kong 2022).

Clearly, neither fairness guidelines nor statistical debiasing and other technological "fixes" are sufficient to properly recognise, let alone eradicate, historically and structurally anchored intersectional forms of discrimination that are perpetuated by GenAI. To reiterate, medial representations, including AI-generated texts, images, and other outputs, transport and produce "controlling images" (see Collins 2002) and narratives that shape how socially relevant groups are imagined, spoken about, or encountered in a given society. They set the terms of who is visible, who can speak, and who is listened to; terms that are typically skewed towards the normalisation of a dichotomy between an unmarked, universal, and empowered (*white,* male) subject position and its various marked (gendered, racialised, etc.) Others (see Haitz 2022: 235). But this does not mean that resistance is futile, or that exercising meaningful epistemic agency in, through, or around GenAI would be impossible. Emerging fields such as queer, decolonial, or indigenous AI have already come up with promising critical approaches that bridge theory and practice, move across activist and academic venues, and combine scholarly scrutiny with political action. Under the banner of design justice (Costanza-Chock 2020), for instance, we find important initia-

8 That an intersectional perspective is not only helpful but essential for fair AI can easily be demonstrated by returning to the groundbreaking study on automatic facial recognition by Buolamwini and Gebru (2018). Instead of focusing on individual metrics, the authors compared the performance of Microsoft, IBM, and Face++ in relation to four subgroups that take into account both skin colour and gender and were thus able to show that the error rate for Black women was significantly higher in all three systems studied than for all other groups (see also Mehrabi et al. 2021a: 9).

tives towards community-led ways of creating and using AI that are based on principles of sustainability, accountability, or accessibility and which are overall attuned to the needs (and vulnerabilities) of those affected by the AI system. What tends to unite ethical, queer-feminist, intersectional, and decolonial approaches is their aim to break with entrenched institutional logics and naturalised patterns of thought, and their impetus to imagine AI otherwise. As I explain in more detail in the upcoming section, the arts can become vital resources for such interventions as they challenge hegemonic narratives and reinsert marginalised voices, knowledges, experiences, and imaginaries into the dataset.

3. The Intersectional Datasets of Zizi *and* Not The Only One

3.1 *Zizi—Queer(ing) the Dataset* by Jake Elwes

To create fair(er) AI, we need to pay keen critical attention to the processes by which new technologies normalise and reify socio-cultural forms of exclusion and violence, as well as the real-world consequences this has for different people (see Eubanks 2018; Hoffmann 2019). This also means paying attention to AI imaginaries, i.e., the totality of culturally dominant narratives, ideas, beliefs, and values associated with AI (see Ernst et al. 2019; Jasanoff and Kim 2015; Mager and Katzenbach 2021; Natale and Ballatore 2020). They influence not only how AI systems are understood and legitimised in the present, but also what kinds of technologies become imaginable for the foreseeable future. With a view to the state of the discourse around fair AI outlined above, the situation seems comparatively dire at first glance. To quote Köppert again, technical approaches, scholarly discourses, and fantasies about AI "cement a very specific concept of technology and, in this respect, are a tactic of concealing what AI could also be" (Köppert 2024: n.pag.). There is still hope for GenAI, though, and perhaps surprisingly to some, it comes in the form of AI art. As artists, curators, and media scholars have begun to point out, and as I will demonstrate shortly, the arts and culture can pinpoint ethical, ecological, and empowering ways of creating and using AI (see Köppert 2024: n.pag.).[9] In the words of Sandra Ciston of the Creative Code Collective, these projects

9 By now, instructive examples abound, including very visible ones such as ImageNet Roulette by Trevor Paglen and Kate Crawford, or the Algorithmic Justice League. ImageNet Roulette is a largescale installation that targets the automated interpretation

"bring intersectional thinking into tech spaces, helping shift an entrenched mindset with creative and helpful, playful and interventionist tools alike" (Ciston 2019: 5). The arts could also offer productive critical interventions in the discourse around AI fairness as such, by reflexively engaging with the inherent contradictions, complexities, and shortcomings of fairness, rather than trying to simplify or mask them. For AI ethics as a research field, it thus certainly seems worthwhile to take them seriously as contributions to the conceptualisation and implementation of fairness, and to look for productive intersections between scholarship and artistic practice.

A particularly interesting example in this regard is *The Zizi Project* (2019–ongoing), a collection of works by London-based conceptual visual artist-researcher Jake Elwes that emerged from a partnership between the artist and the Experiential AI research group at the University of Edinburgh. *The Zizi Project* takes a decidedly queer approach to the training data by combining GenAI with drag performance. Here, I would like to focus on the first part of the project, *Zizi—Queering the Dataset* (2019; hereafter: *Zizi*), which consists of a digital video that was presented at different sites as an installation with several video channels (see Elwes 2019). The video shows a series of faces, often ambiguous in terms of race and gender, slowly morphing into one another in ever-shifting constellations. Bold makeup accentuates lips and eyes, noses and other features pop in and out of existence, creating strange one-eyed creatures. Before long, a new face emerges, and is temporarily thrown into almost startling relief, before shifting yet again, never static, and never staying long enough to be pinned down by the observer's gaze. The effect is unsettling, yet strangely beautiful, and invites audiences to question their assumptions regarding the supposedly natural facts of race or gender and to instead adopt a more fluid perspective.

With regard to *Zizi*'s contribution to the discussion around fair datasets in GenAI, the processes of data curation and training, as outlined by the promotional material and the project website, are revealing (see Elwes 2019; Watling 2021). The video was created by feeding 1.000 images of drag per-

and labelling of images. Users can upload their webcam image and subsequently observe how it is being labelled by a neural network trained on the ImageNet database (Crawford and Paglen 2019; see also Ciston 2019). The Algorithmic Justice League is an organisation combining research (a.o. by prominent scholars such as Joy Buolamwini and Sasha Costanza-Chock) with activism and art, including visual arts, creative writing, and poetry. It also features strongly in the 2020 Netflix documentary *Coded Bias*.

formers into StyleGan, a generative adversarial network that was originally trained on 70.000 images of human faces contained in the Flickr-Faces-HQ Dataset, and then using it to generate new faces. As per the blurb on the artist's website, *Zizi* seeks to intervene in the myths and power structures around AI: The disruption and re-training of models "causes the weights inside the neural network to shift away from the normative identities it was originally trained on and into a space of queerness" (Elwes 2019: n.pag.). True to its "mission statement" to queer the *dataset,* the work is thus centrally concerned with the impact of training data on the output, exploring possibilities of diversifying the dataset, discarding the criterium of representativeness as an expression of a heteronormative order, and thereby finding a means of overcoming the conservatism of GenAI and facilitating the emergence of new and unexpected results.

"Queering," in this context, can be understood in a twofold sense: as an umbrella for the identities behind and beyond the acronym LGBTQIA*, and as an intervention in normative discourses and binary distinctions *per se*. Drawing on a Butlerian framework of gender performativity, *Zizi* shows that sex and gender are, indeed, performative, produced by discourse, constructed from repeated acts and moment-to-moment gestures, and interpreted by and through cultural meanings. Yet, even though this performative construction happens within the "rigid regulatory frame" (Butler 1990: 43) of normative assumptions and pressures, it is not unchangeable; an aspect that *Zizi* emphasises strongly. On the one hand, the exaggerated features of *Zizi*'s morphing images serve as a means of parody, which has value in itself by exposing the constructedness of gender as a powerful social construct. On the other hand, the ever-shifting faces highlight the fact that gender is fluid and unstable; that it must be constantly expressed and interpreted in order for the body to become legible. *Zizi* refuses to offer the audience any sort of respite or sense of reassurance that might come with a temporary halt in the flow of images where gender identity could be fixed, instead privileging the flux, ambiguity, and open-endedness of queer play. In this sense, the work demonstrates how AI-generated images need not offer simplified answers or create a false sense of objectivity, but can also express ambiguity while remaining very much attuned to complex and shifting sociocultural contexts.

The capacity of art for showing, rather than telling, holds promise when it comes to promoting critical data literacy and increasing public awareness of bias in AI. Drew Hemment and colleagues (of the Experiential AI research group that commissioned and collaborated on *The Zizi Project*)

propose the term experiential AI to capture the potential of AI art for offering new, human-centred perspectives on explainable AI (see Hemment et al. 2023; 2024). Analysing the *Zizi Show* (Elwes 2021–ongoing), a deepfake drag show that is also part of the Zizi collection, the authors conclude that "*Zizi* is an explanation of bias in ML and the power of the dataset through experiential means. *Zizi* highlights the way data and design choices shape what ML does" (Elwes 2019: n.pag.; see also Hemment et al. 2024). Data-driven art projects such as *Zizi* could offer more visceral and engaging learning processes, and thereby enhance public understanding of AI, including a sort of critical data literacy based on an awareness of the gaps and biases in the dataset and how these relate to social norms and power structures (see Hemment et al. 2023; 2024). It seems safe to assume that the different approach to explaining AI and the different type of intellectual access offered by the arts will also appeal to different kinds of audiences, notably, audiences beyond the "classic" stakeholders reached by scholarly papers and ethics guidelines. What is more, by creating productive interfaces between art, queer activism, and AI technology, these projects could also help dissolve so-called second-order divisions, i.e., the ostensibly self-imposed exclusions that hinder individuals or groups from participating in the digital realm (see Loh 2024). Upon closer inspection, these "self"-exclusions often turn out to be the result of structural factors and oppressive gendered, racialised, or ableist hierarchies, which once more points to the importance of community-led practices when it comes to re-imagining AI (see Costanza-Chock 2020).

To summarise, *Zizi—Queering the Dataset* marries the unmasking of the purportedly neutral fact of gender to the unmasking of the purported neutrality of the dataset, to the effect that both are exposed as powerful myths. In more concrete terms, *Zizi* combines its critique of sex and gender with a critique of white heteronormative bias in the datasets used to train influential AI systems. By inserting images of faces that do not conform to cis/white/heterosexual norms of gender expression, Elwes quite literally "queers" the dataset. That this also causes the algorithm to generate vastly different output clearly shows the close relation between the selection of training data and the representational politics of the output. To return to the discussion of fairness presented in Section 2, *Zizi* draws attention to the interdependence between (seemingly objective and representative) training data and societal values and hegemonic cultural formations and demonstrates the value of diversified datasets when it comes to prompting GenAI to create more inclusive, even subversive content. Through its rather

spectacular example of an AI-driven interrogation of gender norms, *Zizi* furthermore exemplifies the potential of art for increasing AI literacy, promoting critical reflection, and encouraging different groups of people to experiment with forms of AI that better suit their epistemic needs.

3.2 *Not the Only One* by Stephanie Dinkins

N'TOO, short for *Not the Only One* (Dinkins 2018–ongoing), is described on its webpage as "an ongoing experiment" and "an attempt to narrate a multigenerational memoir of a black American family" from the first-person perspective of an evolving AI (see Dinkins 2018: n.pag.). *N'TOO* was exhibited in various locations in the form of a sculpture resembling a seashell, with the faces of three Black women protruding from its surface. The bot inside the sculpture is voice-interactive, meaning that visitors can converse with it and listen to answers generated from the system's database. When curating the dataset for *N'TOO*, great care was taken to avoid importing biases and hegemonic epistemic hierarchies, focussing instead on creating a markedly intersectional database. *N'TOO* was trained on two types of data: the oral histories contributed by three Black women, representing three generations of the artist's family, and on Black diasporic literatures, films, and TV that were central to these women's experience. All of these data sources are particularly attuned to *Black women*'s experiences, and to the complex and overlapping forms of discrimination they have been facing across generations. As an evolving system, the AI also learns from its interactions with humans and expands its vocabulary, thus adding additional voices to the archive. By contrast, no open-source, ready-made, or large scraped datasets were used to avoid importing racist bias and epistemic violence that would "taint" (see Dinkins 2018: n.pag.) the self-narratives and self-understandings encoded in and communicated by the bot (see Cooper 2020; Klassen and Aceves Sepúlveda 2022; Paul 2024). As a downside, the data used to train the bot is insufficient to support seamless conversations with visitors. However, according to Dinkins, this is not so much a flaw as a crucial feature of *N'TOO*, since the gaps in the model create teachable moments about "the limitations of big data and possibilities of small data," the value of data sovereignty, and the role of the community (Dinkins 2018: n.pag.).

Not dissimilar to *Zizi*, *N'TOO* deconstructs the myth of data neutrality and encourages critical reflections about the quality of training data and its relation to systemic oppression. It does so with a strong decolonial, intersectional, and future-oriented impetus. At the core of *N'TOO*, we find a rather hopeful vision of a future in which marginalised communities would take control over the development, training, and use of GenAI, infusing it with community-specific values and tailoring it to community-specific goals. In turn, GenAI could enable the self-expression of these communities and help recover archives that seemed lost. As a first-person storyteller, *N'TOO* speaks with one voice, yet this voice remains polyphonous, representing both shared and profoundly personal experiences. As a living, Black female and multigenerational memoir, *N'TOO* shows how GenAI could help amplify voices, communicate experiences, and reconstruct genealogies that seemed lost because they were historically written out; erased from the official record. Within the framework of epistemic justice, *N'TOO* thus signifies a sort of collective "exercise of epistemic agency" by and for Black women that challenges "prevailing practices of epistemic injustice" (Collins 2017: 117). Such situated, intersectional, and community-led practices for training and using GenAI could then become effective means of writing (or rather speaking) back,[10] as well as an effective first step towards decolonising the archives upon which we train future AI systems (see Adams 2021; Hakopian 2024; Murphy and Largacha-Martínez 2022).

In its interest in drawing connections between humans and AI, *N'TOO* also chimes with posthumanist thinking and alternative models of kinship (see Haraway 1991; Nakamura 2023). Treated as a mind, interpreter, and witness in its (or rather her?) own right, the bot becomes something in-between an archive and a fourth-generation family member (see Dinkins 2024; Klassen and Aceves Sepúlveda 2022). Observing visitors engage with *N'TOO*, Dinkins takes particular note of the tentative bonds of nurture and care that emerge between humans and AI, and which seem to hint at the possibility for new forms of kinship and relationality. Dinkins is undoubtedly quite optimistic about the potential of AI to "make kin" and

10 Once again, I take inspiration from influential concepts and approaches in postcolonial (literary) studies, where the "writing back" paradigm refers to concerted and often successful efforts of postcolonial authors to contest and subvert dominant imperialist discourses through narrative practices that respond to canonical literary works (Ashcroft et al. 1989). Writing and artistic expression, in this context, become acts of reclaiming authority and demanding voice; a means of asserting control over the narrative and to re-establish interpretive sovereignty.

to erode boundaries between subject and object, self and other, human and machine. As she argues in her essay "Afro-now-ism," the present moment of rapid technological development is one of profound opportunity for imagining, and ultimately achieving, a better future beyond systemic oppression and oppositional thinking. Radical technological and cultural changes around AI, she writes, challenge us to "understan[d] ourselves as participants in an expanding continuum of intelligences" (Dinkins 2024: 5–6) and to recognise that "[t]he boundaries between sovereign consciousness, nature, valued knowledge, biotechnologies, power and social reality are optical illusions" (Dinkins 2024: 7). Rather than cautioning against the human tendency towards anthropomorphism, *N'TOO* embraces the blurring of boundaries between human and machine as a way of overcoming binary thinking and, possibly, as an alternative to technological fixes for our present condition of ontological and epistemological uncertainty.

Both *Zizi* and *N'TOO* are characterised by a transgressive approach to training data in the sense that they free themselves from the hegemonic power structures that govern our data and our mythmaking while showcasing how technology might be repurposed to empower marginalised communities. Importantly, neither project stops at recognising and criticising the systematic exclusion of non-*white* and non-heteronormative voices from the dataset. Rather, the agency, creativity, and knowledge(s) of these communities are understood as essential for thinking and building better technological futures (see Nakamura 2024). To close the circle to the theoretical and practical approaches to fairness in GenAI surveyed in the beginning, artistic interventions such as *Zizi* and *N'TOO* show how the fairness of AI could be improved by listening to marginalised communities and taking them seriously as knowers. On the one hand, this has very immediate practical implications since designers of GenAI must learn about the needs and desires of everyone affected, and especially of vulnerable groups, and accommodate this information in the design of the system (see Fosch-Villaronga and Malgieri 2024). Approaches such as Design Justice provide good-practice examples (see Costanza-Chock 2020). On the other hand, the inclusion and wider familiarity with intersectional approaches stands to affect the scholarly discourse, ideally leading to a more contextualised and epistemically just understanding of fairness in AI.

4. Conclusion

Drawing on an extensive and fast-growing body of research on fairness in ML, this chapter proposed an understanding of fairness as freedom from discriminatory bias, where biases are understood as systematic errors that lead to the unfair differential treatment of socially relevant groups of people. In the context of GenAI, discriminatory biases can lead to representational harms by perpetuating stereotypes, generating demeaning content, or by erasing specific identities, typically those that are already disadvantaged in "real life." Remedying unfair discriminatory biases in GenAI is therefore imperative from an ethical point of view, but is currently hindered by tricky challenges that indicate the need for contextual, community-led, and intersectional approaches. I have moreover argued that the arts and culture can point the way towards possible solutions to the dilemmas and paradoxes of fair AI, not least because they provide access to a more diverse set of epistemic resources. On the basis of these theoretical considerations, the chapter then discussed *Zizi—Queering the Dataset* by Jake Elwes and *Not the Only One* by Stephanie Dinkins as two examples of art projects that use diverse datasets in order to re-imagine fairness in GenAI from queer and decolonial perspectives. In emphasising intersectional thinking and the agency of racialised and marginalised groups, these works highlight the untapped potential of playful, creative, and community-led approaches for infusing GenAI with new ideas and values in the present, and for imagining fairer AI futures.

Funding Declaration: This work was funded by the German Ministry of Education and Research within the project "Privacy, Democracy, and Self-Determination in Times of AI and Globalization" (PRIDS). Funding no.: 16KIS1380.

Acknowledgements: I wish to thank my colleagues Jana Hecktor and Lisa Koeritz for their thorough reading and knowledgeable feedback on an early draft of this paper. Any remaining errors are my own.

References

Adams, Rachel (2021): Can Artificial Intelligence Be Decolonized?, in: Interdisciplinary Science Reviews 46 (1–2), pp. 176–197.

Akter, Shahriar et al. (2021): Algorithmic Bias in Data-Driven Innovation in the Age of AI, in: International Journal of Information Management 60, article 102387.

Angwin, Julia et al. (2016): Machine Bias, in: ProPublica, 23 May 2016 (online available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing – accessed 18 October 2024).

Ashcroft, Bill / Griffiths, Gareth / Tiffin, Helen (1989): The Empire Writes Back: Theory and Practice in Post-Colonial Literatures, New York.

Barocas, Solon / Hardt, Moritz / Narayanan, Arvind (2023): Fairness and Machine Learning: Limitations and Opportunities, Cambridge.

Barocas, Solon / Selbst, Andrew D. (2016): Big Data's Disparate Impact, in: California Law Review 104 (3), pp. 671–732.

Benjamin, Ruha (2019): Race After Technology: Abolitionist Tools for the New Jim Code, Cambridge.

Binns, Reuben (2018): Fairness in Machine Learning: Lessons from Political Philosophy, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, pp. 149–159.

Blodgett, Su Lin et al. (2020): Language (Technology) Is Power: A Critical Survey of "Bias" in NLP, in: Dan Jurafsky et al. (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476.

Brandner, Lou Therese / Hirsbrunner, Simon David (2023): Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung, in: TATuP 32 (1), pp. 24–29.

Brandner, Lou Therese et al. (2023): How Data Quality Determines AI Fairness: The Case of Automated Interviewing, in: EWAF'23: European Workshop on Algorithmic Fairness, Winterthur, Switzerland (online available at: https://ceur-ws.org/Vol-3442/paper-25.pdf – accessed 18 October 2024).

Buolamwini, Joy / Gebru, Timnit (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, pp. 77–91.

Butler, Judith (1990): Gender Trouble: Feminism and the Subversion of Identity, New York.

Chowdhury, Rumman / Mulani, Narendra (2018): Auditing Algorithms for Bias, in: Harvard Business Review (online available at: https://hbr.org/2018/10/auditing-algorithms-for-bias – accessed 18 October 2024).

Ciston, Sarah (2019): Intersectional AI Is Essential: Polyvocal, Multimodal, Experimental Methods to Save Artificial Intelligence, in: Journal of Science, Technology and Arts 11 (2), pp. 3–8.

Collins, Patricia Hill (2002): Mammies, Matriarchs, and Other Controlling Images, in: Patricia Hill Collins, Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment, New York, pp. 69–96.

Collins, Patricia Hill (2017): Intersectionality and Epistemic Injustice, in: Ian James Kidd / José Medina, Gaile Pohlhaus Jr. (eds.), The Routledge Handbook of Epistemic Injustice, New York, pp. 115–124.

Cooper, Imani (2020): Inheritance: Ode to N'TOO, in: Absinthe 26 (1), n.pag.

Costanza-Chock, Sasha (2020): Design Justice: Community-Led Practices to Build the Worlds We Need, Cambridge.

Crawford, Kate (2017): The Trouble with Bias, in: AI Now Institute (online available at: https://ainowinstitute.org/news/the-trouble-with-bias – accessed 16 December 2025).

Crawford, Kate / Paglen, Trevor (2019): Excavating AI: The Politics of Images in Machine Learning Training Sets (online available at: https://www.excavating.ai – accessed 18 October 2024).

Crenshaw, Kimberlé (1989): Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, in: University of Chicago Legal Forum, pp. 139–167.

Dev, Sunipa et al. (2022): On Measures of Biases and Harms in NLP, arXiv. https://doi.org/10.48550/arXiv.2108.03362

Dinkins, Stephanie (2018): Not The Only One: Project Website (online available at: https://www.stephaniedinkins.com/ntoo.html – accessed 18 October 2024).

Dinkins, Stephanie (2024): Afro-Now-Ism: The Unencumbered Black Mind Is a Wellspring of Possibility, in: Srimoyee Mitra (ed.), Stephanie Dinkins on Love and Data, Ann Arbor, pp. 4–15.

Eapen, Tojin T. et al. (2023): How Generative AI Can Augment Human Creativity, in: Harvard Business Review (online available at: https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity – accessed 18 October 2024).

El Atillah, Imane (2023): Copyright Challenges in the Age of AI: Who Owns AI-Generated Content?, in: Euronews, 10 July (online available at: https://www.euronews.com/next/2023/07/10/copyright-challenges-in-the-age-of-ai-who-owns-ai-generated-content – accessed 18 October 2024).

Ellingrud, Kweilin et al. (2023): Generative AI and the Future of Work in America, McKinsey Global Institute (online available at: https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america – accessed 18 October 2024).

Elwes, Jake (2019): Project Website for Queering the Dataset (online available at: https://www.jakeelwes.com/project-zizi-2019.html – accessed 18 October 2024).

Ernst, Christoph / Schröter, Jens / Sudmann, Andreas (2019): AI and the Imagination to Overcome Difference, in: spheres: Journal for Digital Cultures 5, pp. 1–12.

Eubanks, Virginia (2018): Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, New York.

Feuerriegel, Stefan / Dolata, Mateusz / Schwabe, Gerhard (2020): Fair AI: Challenges and Opportunities, in: Business & Information Systems Engineering 62 (4), pp. 379–384.

Fleming, Sam (2024): Generative Artificial Intelligence Will Lead to Job Cuts This Year, CEOs Say, in: Financial Times, 15 January (online available at: https://www.ft.com/content/908e5465-0bc4-4de5-89cd-8d5349645dda – accessed 18 October 2024).

Fosch-Villaronga, Eduard / Malgieri, Gianclaudio (2024): Queering the Ethics of AI, in: David J. Gunkel (ed.), Handbook on the Ethics of Artificial Intelligence, Cheltenham, pp. 301–315.

Foulds, James R. et al. (2020): An Intersectional Definition of Fairness, in: 36th International Conference on Data Engineering (ICDE), pp. 1918–1921.

Fricker, Miranda (2007): Epistemic Injustice: Power and the Ethics of Knowing, Oxford.

Gautam, Sanjana / Venkit, Pranav N. / Ghosh, Sourojit (2024): From Melting Pots to Misrepresentations: Exploring Harms in Generative AI, arXiv. https://doi.org/10.48550/arXiv.2403.10776

Hagendorff, Thilo (2019): From Privacy to Anti-Discrimination in Times of Machine Learning, in: Ethics and Information Technology 21, pp. 331–343.

Hagendorff, Thilo (2021): Linking Human and Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning, in: Minds and Machines 31 (4), pp. 563–593.

Hagendorff, Thilo / Wezel, Katharina (2020): 15 Challenges for AI: Or What AI (Currently) Can't Do, in: AI & Society 35 (2), pp. 355–365.

Haitz, Louise (2022): Medienwissenschaft und Intersektionalität, in: Astrid Biele Mefebue / Andrea D. Bührmann / Sabine Grenz (eds.), Handbuch Intersektionalitätsforschung, Wiesbaden, pp. 229–242.

Hakopian, Mashinka Firunts (2024): Art Histories from Nowhere: On the Coloniality of Experiments in Art and Artificial Intelligence, in: AI & Society 39 (1), pp. 29–41.

Haraway, Donna (1991): A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century, in: Donna Haraway, Simians, Cyborgs, and Women: The Reinvention of Nature, New York, pp. 149–181.

Heesen, Jessica / Reinhardt, Karoline / Schelenz, Laura (2021): Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine digitale demokratische Gesellschaft, in: Gero Bauer et al. (eds.), Diskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis, Bielefeld, pp. 129–148.

Hemment, Drew et al. (2023): AI in the Public Eye: Investigating Public AI Literacy through AI Art, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York, pp. 931–942.

Hemment, Drew et al. (2024): Experiential AI: Enhancing Explainability in Artificial Intelligence through Artistic Practice, in: Leonardo 57 (3), pp. 298–306.

Hoffmann, Anna Lauren (2019): Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse, in: Information, Communication & Society 22, pp. 900–915.

Hsu, Tiffany / Thompson, Stuart A. (2023): Disinformation Researchers Raise Alarms about A.I. Chatbots, in: The New York Times, 20 June (online available at: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html – accessed 18 October 2024).

Jasanoff, Sheila / Kim, Sang-Hyun (2015): Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power, Chicago.

Kearns, Michael et al. (2018): Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, in: PMLR 80, pp. 2564–2572.

Kim, Pauline T. (2022): Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action, in: California Law Review 110 (5), pp. 1539–96.

Klassen, Lois / Aceves Sepúlveda, Gabriela (2022): Amplified Listening to Race and Gender in Fiamma Montezemolo's "Echo" and Stephanie Dinkins's "N'TOO", in: Media-N 18 (1), pp. 102–120.

Kong, Youjin (2022): Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, pp. 485–494.

Köppert, Katrin (2024): Queersplaining AI, in: Boell, 28 May (online available at: https://eu.boell.org/en/2024/05/28/queersplaining-ai – accessed 18 October 2024).

Leavy, Susan / O'Sullivan, Barry / Siapera, Eugenia (2020): Data, Power and Bias in Artificial Intelligence, arXiv. https://doi.org/10.48550/arXiv.2008.07341

Lin, Cong et al. (2023): Trapped in the Search Box: An Examination of Algorithmic Bias in Search Engine Autocomplete Predictions, in: Telematics and Informatics 85.

Loh, Wulf H. (2024): Generative KI, digitale Teilhabe und epistemische Ungerechtigkeit, in: Rechtsphilosophie – Zeitschrift für Grundlagen des Rechts 10 (2), pp. 215–233.

Mager, Astrid / Katzenbach, Christian (2021): Future Imaginaries in the Making and Governing of Digital Technology: Multiple, Contested, Commodified, in: New Media & Society 23 (2), pp. 223–236.

Marr, Bernard (2023): The Intersection of AI and Human Creativity: Can Machines Really Be Creative?, in: Forbes, 27 March (online available at: https://www.forbes.com/sites/bernardmarr/2023/03/27/the-intersection-of-ai-and-human-creativity-can-machines-really-be-creative – accessed 18 October 2024).

Mehrabi, Ninareh et al. (2021a): A Survey on Bias and Fairness in Machine Learning, in: ACM Comput. Surv. 54 (6), Article 115, pp. 1–35.

Mehrabi, Ninareh et al. (2021b): Lawyers Are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources, in: Marie-Francine Moens et al. (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, pp. 5016–5033.

Mitchell, Shira et al. (2021): Algorithmic Fairness: Choices, Assumptions, and Definitions, in: Annual Reviews of Statistics and Its Applications 8 (1), pp. 141–163.

Mohammed, Sedir et al. (2023): Ein Glossar zur Datenqualität (1.2), Zenodo. https://doi.org/10.5281/zenodo.7702426

Murphy, John W. / Largacha-Martínez, Carlos (2022): Decolonization of AI: A Crucial Blind Spot, in: Philosophy & Technology 35 (4), pp. 1–13.

Nakamura, Lisa (2023): "Who Are Your People?": Stephanie Dinkins's Afro-Now-ism as Algorithmic Abundance, in: Mitra, Srimoyee (ed.), Stephanie Dinkins on Love and Data, Ann Arbor, pp. 52–59.

Natale, Simone / Ballatore, Andrea (2020): Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence, in: Convergence 26 (1), pp. 3–18.

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

Paul, Christiane (2024): The Data You Give, in: Mitra, Srimoyee (ed.), Stephanie Dinkins on Love and Data, pp. 28–35, Ann Arbor.

Prates, Marcelo O.R. / Avelar, Pedro H. / Lamb, Luís C. (2020): Assessing Gender Bias in Machine Translation: A Case Study with Google Translate, in: Neural Computing and Applications 32, pp. 6363–6381.

Roche, Cathy / Lewis, Dave / Wall, P.J. (2021): Artificial Intelligence Ethics: An Inclusive Global Discourse?, Proceedings of the 1st Virtual Conference on Implications of Information and Digital Technologies for Development, arXiv. https://doi.org/10.48550/arXiv.2108.09959

Roche, Cathy / Wall, P.J. / Lewis, Dave (2023): Ethics and Diversity in Artificial Intelligence Policies, Strategies and Initiatives, in: AI and Ethics 3 (4), pp. 1095–1115.

Segev, Re'em (2025): The Moral Status of Input and Output Discrimination, in: AI and Ethics 5 (1), pp. 323–332.

Segun, Samuel T. (2021): Critically Engaging the Ethics of AI for a Global Audience, in: Ethics and Information Technology 23 (2), pp. 99–105.

Spivak, Gayatri Chakravorty (1988): Can the Subaltern Speak?, in Cary Nelson / Lawrence Grossberg (eds.), Marxism and the Interpretation of Culture, London, pp. 24–28.

Sun, Xiao-yu / Ye, Bin / Xia, Bao-hua (2025): The Problem of Fairness in Tools for Algorithmic Fairness, in: AI and Ethics 5, pp. 1847–1857.

Tacheva, Jasmina / Ramasubramanian, Srividya (2023): AI Empire: Unraveling the Interlocking Systems of Oppression in Generative AI's Global Order, in: Big Data & Society 10 (2), pp. 1–13.

Thomsen, Frej Klem (2024): Algorithmic Indirect Discrimination, Fairness and Harm, in: AI and Ethics 4, pp. 1023–1037.

Verma, Pranshu / De Vynck, Gerrit (2023): ChatGPT Took Their Jobs. Now They Walk Dogs and Fix Air Conditioners, in: The Washington Post, 2 June (online available at: https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/ – accessed 18 October 2024).

Wang, Xiaomen / Zhang, Yishi / Zhu, Ruilin (2022): A Brief Review on Algorithmic Fairness, in: Management System Engineering 1, Article 7.

Watling, Eve (2021): Meet the Artist Queering AI Technology, in: The Independent, 26 July (online available at: https://www.independent.co.uk/arts-entertainment/photography/zizi-queering-dataset-ai-drag-jake-elwes-b1876396.html – accessed 18 October 2024).

The Zizi Project by Jake Elwes (n.d.): The New Real (online available at: https://www.newreal.cc/artworks/the-zizi-project – accessed 18 October 2024).

Zou, James / Schiebinger, Londa (2018): AI Can Be Sexist and Racist — It's Time to Make It Fair, in: Nature 559, pp. 324–326.

Der Algorithmus macht, was er soll, oder? – Eine technikethische Reflexion automatisierter Detektion von Desinformationen im Internet

Mario Anastasiadis und Hektor Haarkötter

Zusammenfassung

Desinformation hat sich zu einer zentralen Problemlage für die gesellschaftliche Selbstverständigung entwickelt, wobei insbesondere soziale Online-Medien im alltäglichen Medienhandeln eine herausragende Bedeutung einnehmen. Unwahre oder irreführende Inhalte finden sich in nahezu allen Themenbereichen – etwa in Politik, Gesundheit, Medizin oder Kultur – und treten auf sämtlichen gesellschaftlichen Ebenen auf, häufig verbunden mit erheblichen Herausforderungen und problematischen Folgen für die jeweils betroffenen Menschen. Der vorliegende Beitrag widmet sich der Frage nach algorithmisch gestützten, automatisierten Detektionssystemen, die der Identifikation von Desinformation dienen und zugleich auf die Förderung kritischer und resilienter Medienkompetenzen abzielen. Grundlage ist das Forschungsprojekt NEBULA, in dessen Rahmen ein entsprechendes Detektionssystem als Demonstrator einer mobilen App entwickelt wird. Nach einer begrifflichen Einordnung, kurzen Einblicken in die empirische Datenlage zu Desinformation im Medienalltag sowie einer Darstellung des Zusammenhangs mit sozialen Online-Medien werden ausgewählte Ergebnisse der die Entwicklung begleitenden qualitativen Forschung vorgestellt. Im Fokus steht dabei die Frage, inwiefern technikethische Kriterien wie Validität (zuverlässige Klassifikation), Adaptivität (Anpassungsfähigkeit an neue Kontexte), Transparenz und Verständlichkeit der Ausgaben, Reziprozität sowie die Förderung von Medienkompetenz in den Entwicklungsprozess integriert wurden.

1. Einleitung

Desinformationen sind zu einer großen Herausforderung für die gesellschaftliche Selbstverständigung geworden, wobei insbesondere Sozialen

Medien in der alltäglichen Mediennutzung vieler Menschen eine wesentliche Rolle zukommt. Desinformationen sind in nahezu allen Themenfeldern, etwa Politik, Gesundheit und Medizin oder Kultur, sowie auf allen gesellschaftlichen Ebenen virulent – nicht selten mit problematischen Konsequenzen für die betroffenen Akteure.

Laut der dem britischen Innenministerium zugeordneten *Accelerated Capability Environment* (ACE) werden im Jahr 2025 zudem schätzungsweise 8 Millionen Deepfakes veröffentlicht. Bemerkenswert ist daran die Steigerungsrate: Zwei Jahre zuvor waren es „nur" etwa 500.000 (vgl. Henzler 2025). Auf Makroebene sind Staaten herausgefordert, da Desinformationen als Mittel politischer Konflikte Hochkonjunktur haben – derzeit etwa im Kontext russischer Propagandaaktivitäten (vgl. Sato/Wiebrecht 2024: 1011f.). An dieser Stelle sind es staatliche Akteure, die Desinformationen gezielt, planvoll und in großer Menge herstellen und über klassische sowie digitale Medien verbreiten. Auf der Mesoebene haben Desinformationen Konsequenzen für institutionelles Handeln, da sie die Integrität etwa von Medien, Wissenschaft oder Politik schädigen können (vgl. McIntosh/White/Vitale 2023: 8). Sie sind zudem auf der Mikroebene der täglichen, individuellen Lebenswelten hoch präsent (vgl. Bernhard et al. 2024).

Dieser Beitrag diskutiert algorithmisch unterstützte, automatisierte Detektionssysteme zur Identifikation von Desinformation sowie zur Stärkung kritischer und resilienter Medienkompetenzen. Grundlage dafür bildet ein Forschungsprojekt[1], in dem ein solches Detektionssystem in Form eines Demonstrators für eine mobile App entwickelt wird. Nach einer begrifflichen Einordnung, kurzen Hinweisen zur empirischen Datenlage in Bezug auf Desinformationen im Medienalltag der Menschen sowie Ausführungen zum Zusammenhang mit Sozialen Medien werden ausgewählte Ergebnisse aus der die Entwicklung flankierenden qualitativen Begleitforschung präsentiert. Dabei wird diskutiert, inwiefern die technikethischen Parameter der Validität (verlässliche Klassifikation), Adaptivität (Anpassung eines Systems an neue Kontexte), Transparenz und Verständlichkeit des Outputs, Reziprozität sowie die Stärkung von Medienkompetenz in die Entwicklung eingeflossen sind.

1 Das Projekt „NEBULA – Nutzerzentrierte KI-basierte Erkennung von Fake News und Fehlinformationen" wurde im Rahmen des Programms „Forschung für die zivile Sicherheit 2018 – 2023" der Bundesregierung durch das Bundesministerium für Forschung, Technologie und Raumfahrt gefördert.

2. Politische Desinformation und digitale Öffentlichkeit

Neben der genaueren begrifflichen Konturierung ist nachfolgend eine Skizze des kommunikativen Umfelds digitaler Öffentlichkeit in sozialen Online-Medien hinsichtlich ihrer Relevanz für die Existenz und Verbreitung von Desinformationen notwendig.

2.1. Desinformation – Abgrenzung und Begriffsschärfung

Eine terminologische Annäherung an den Gegenstandsbereich offenbart eine Vielzahl von Begriffen, wie etwa „Desinformation", „Fehlinformation" oder „Misinformation", die in der akademischen Debatte sowie im Alltagsdiskurs nicht selten synonym gebraucht werden, jedoch wichtige Unterschiede aufweisen. Um den Begriff der Desinformation genauer zu konturieren, bedarf es vor allem einer Abgrenzung zu den Begriffen der Fehl- und Misinformation. Während diese Begriffe auch Inhalte bezeichnen, die nicht intentional falsch sind, zum Beispiel redaktionelle Fehler oder falsche Informationen wider besseres Wissen (vgl. Zimmermann/Kohring 2018), stellt das in Anlehnung an Allcott und Gentzkow (2017) hier zugrundegelegte Verständnis von Desinformationen demgegenüber neben der nachweisbaren Falschheit der Inhalte ihre Intentionalität, also die konkrete Täuschungsabsicht ins Zentrum (vgl. Haarkötter 2021). Desinformationen im engeren Sinne sind also intentional falsche Inhalte, als solche hergestellt und mit einer dezidierten Täuschungsabsicht verbunden, die darauf abzielt, das gesellschaftliche Meinungsklima zu verändern.

2.2. Digitale Öffentlichkeit und Desinformation

Öffentlichkeit, verstanden als Forum, in dem Bürger:innen, zivilgesellschaftliche und politische Akteure deliberative Auseinandersetzungen austragen (vgl. Habermas 1990), ist durch die Digitalisierung einem tiefgreifenden Strukturwandel unterworfen (vgl. Habermas 2022; Seeliger/Sevignani 2021). Digitale Öffentlichkeit in Social Media zeichnet sich durch ambivalente Charakteristika aus, die auch für den Bereich der Verbreitung, Aneignung und Wirkung von Desinformationen zentral sind. Einerseits haben sie erhebliche partizipative und mobilisierende Potenziale. Andererseits wird die deliberative Güte der digitalen Öffentlichkeiten mittlerweile meist kri-

tisch gesehen (vgl. Seeliger/Sevignani 2021). Für den vorliegenden Kontext ist zunächst die Zunahme der Präsenz von Desinformation in der täglichen Social Media-Nutzung vieler Menschen zu konstatieren. So haben allein im Jahr 2023 etwa 89 Prozent der Menschen in Deutschland Desinformationen im Internet wahrgenommen (vgl. Bernhard et al. 2024). Ein genauerer Blick auf einzelne Social Media-Kanäle zeigt, dass Desinformationen den Nutzer:innen besonders auf TikTok, X und Facebook begegnen. 88 Prozent der befragten Nutzer:innen nahmen auf TikTok Desinformationen wahr. Auf X waren es 90 Prozent und auf Facebook gar 94 Prozent (ebd.). Social Media spielen für die Verbreitung von Desinformationen somit eine erhebliche Rolle. Nachfolgend werden einige der zentralen Gründe für diese Entwicklung skizziert, und zwar hinsichtlich kommunikativer Ermächtigungseffekte sozialer Medien, der ökonomischen und technologischen Ausrichtungen der großen Plattformen, ihrer für Desinformationen relevanten unternehmensstrategischen Entscheidungen im Umgang mit Fact Checking, in Bezug zur medienpolitischen Ausrichtung der US-Administration sowie hinsichtlich eines libertären Kommunikationsverständnisses der Plattformbetreiber.

Social Media hat kommunikative Ermächtigungseffekte auch für solche Akteure, die alternativen Wissensformen (vgl. Fries 2021) zuneigen und Desinformationen oder gar digitale Propaganda verbreiten (vgl. Broschart 2024). Dies konkretisiert sich etwa in der Entstehung und Konsolidierung ‚alternativer' Nachrichtenwelten, die offen für die Verbreitung von Desinformationen und vornehmlich über Soziale Medien erreichbar sind. Diese alternativen Nachrichtenmedien verstehen sich nicht selten als Opposition zu Informationsjournalismus und angeblich hegemonialer Öffentlichkeit aus Politik und Medien (vgl. Schwaiger 2022). Des Weiteren gilt es, den Blick auf die Funktionsweisen der Plattformen, wie algorithmische Filterung sowie ihre unternehmerischen Ausrichtungen selbst zu werfen (vgl. van Dijck/Poell 2013). Für den Bereich der politischen Desinformation ist relevant, dass Plattformen auf Grundlage verschiedener algorithmisch unterstützter Formen der Informationskuratierung arbeiten (vgl. Gillespie 2014).

Für den Bereich der politischen Kommunikation können dabei inhaltliche Homogenisierungs- sowie Fragmentierungseffekte relevant sein. Dabei wird vielfach eine mögliche Aufsplitterung in kleinteilige Teilöffentlichkeiten diskutiert, die deliberative Aushandlungsprozesse erschweren kann (vgl. Habermas 2022). Dies bedeutet jedoch nicht, dass Nutzer:innen informationell gänzlich isoliert von Inhalten, Perspektiven oder Argumenten

sind, die sich von ihren Haltungen unterscheiden. An dieser Stelle haben Vorstellungen von hermetisch geschlossenen Echokammern (vgl. Sunstein 2001) oder Filterblasen (vgl. Pariser 2017) allenfalls heuristischen Wert (vgl. Bruns 2019). Gleichwohl kann algorithmische Informationskuratierung eine deutliche vereinheitlichende Tendenz haben, die im Sinne der Nutzerbindung ökonomisch legitim sein mag, im Sinne eines pluralistisch informierten Subjekts und einer deliberativ orientierten Informiertheit und Debatte jedoch – zumindest mit Blick auf eben diese normativen Prinzipien – deutlich nachteilig sein kann. Dieser Effekt kann im Bereich der politischen Desinformation verstärkt werden, da gerade diese Inhalte nicht selten im Sinne einer hohen *clickability* gestaltet sind und dadurch mitunter algorithmisch präferiert werden.

Die hohe Präsenz von politischen Deformationen erklärt sich jedoch nicht nur ökonomisch oder hinsichtlich der Rolle von Algorithmen. Vielmehr stehen auch konkrete unternehmerische Entscheidungen hinter dieser Entwicklung. Dies lässt sich etwa an X (vormals Twitter) sowie Facebook (Meta) veranschaulichen, denn in beiden Fällen wurden Systeme zum Fact Checking und zur Reduktion von Desinformation konkret zurückgefahren. Nach der Übernahme von Twitter durch Elon Musk im Oktober 2022 wurde das interne System zur Identifikation und Entfernung von Desinformationen grundlegend umgebaut. Zunächst wurde die redaktionell unterstützte Moderation von Kommunikation weitgehend abgeschafft (vgl. Delcker 2022). Außerdem beendete Musk die Zusammenarbeit mit professionellen, unabhängigen Faktenprüfern, etwa Reuters oder Associated Press (vgl. Becker 2025). Musk begründete dies auch mit seinem Misstrauen gegenüber Faktenprüfern, da diese politisch voreingenommen seien (ebd.). An ihre Stelle traten sogenannte *community notes*, bei denen Nutzer:innen selbst Korrekturen zu fragwürdigen Tweets vorschlagen und einreichen können. An dieser Stelle wird also einem Crowdsourcing-Prinzip einer Expertenprüfung gegenüber der Vorzug gegeben. Parallel dazu kündigte Meta nach der Rückkehr Trumps ins Weiße Haus im Jahr 2024 einen Strategiewechsel an, der sich inhaltlich stark mit Musks Ansatz deckt. Meta beendet weitgehend seine Kooperation mit schätzungsweise 100 unabhängigen internationalen Fact-Checking-Organisationen, darunter auch Correctiv in Deutschland, und verlagert Moderationsteams vom Kalifornien in konservativere US-Regionen wie Texas, um angebliche politische Voreingenommenheit dieser Teams auszugleichen (vgl. Duffy 2025; Weatherbed 2025). Auch Marc Zuckerberg begründet den Rückzug aus der Faktenprüfung damit, dass deren Arbeit zunehmend als politisch zu links und

eher konservativen Positionen gegenüber zensierend wahrgenommen werde (vgl. Laaff 2025). Die bisherige Moderation soll nun ebenfalls über eine an *community notes* angelehnte Form der Nutzerbeteiligung erfolgen, und nicht mehr im Kern auf einer Expertenprüfung basieren (vgl. Weatherbed 2025). Wie auch Musk beschreibt Zuckerberg unabhängige Fact-Checker als zu voreingenommen gegen konservative Positionen und kritisiert, dass moderierende Maßnahmen zu viele harmlose Nutzer:innen zensierten. Diese Maßnahmen gelten zwar primär für den US-Markt. Es ist fraglich, ob Meta ähnliche Prinzipien in anderen Ländern übernehmen wird. Die möglichen Auswirkungen dieser Strategiewechsel sind mit Blick auf die hier in Rede stehenden Fragen eindeutig. Beide Beispiele illustrieren einen Trend hin zu einer schwächeren Infrastruktur der inhaltlichen Qualitätssicherung zugunsten eines Modells, das auf Crowdsourcing und freier Rede basiert. Diese Entwicklung ist zentral, da sie eine erhöhte Reichweite und Persistenz auch von Desinformationen fördert und den Zugang zu zeitnahen und fundierten Korrekturen im Rahmen der Plattformen erschwert. Somit ist eine starke Zunahme von Desinformation auf den entsprechenden Social Media-Plattformen sehr wahrscheinlich. Insgesamt markieren die Strategiewechsel der Plattformbetreiber eine signifikante Verschiebung in der Verantwortung sozialer Medien von professionellem Fact-Checking hin zu einem unregulierten, nutzergesteuerten Faktenmonitoring, mit womöglich weitreichenden Folgen für die demokratische Diskurskultur und die Bekämpfung von Desinformationen – nicht nur in den USA.

Nicht wenige Stimmen aus Journalismus, Politik und Zivilgesellschaft werfen X und Meta vor, eine kapitulative Haltung gegenüber rechten politischen Strömungen – und konkret gegenüber der US-Administration unter Trump – einzunehmen (vgl. Weatherbed 2025). Dabei ist es offensichtlich, dass die unternehmerischen Entscheidungen von X und Meta zumindest in den USA von politisch höchster Stelle legitimiert werden und sich mit den medienpolitischen Positionen der US-Administration decken. Dies wurde in jüngerer Vergangenheit vor allem an Äußerungen des US-Vizepräsidenten JD Vance deutlich, etwa im Rahmen seiner Rede auf der Münchener Sicherheitskonferenz 2025, in der er Einblicke in seine Einschätzungen zum Verhältnis von vermeintlich „freier Rede" und ihrer regulatorischen Steuerung gewährt.

> „In […] Europe, free speech, I fear, is in retreat. […] I will admit that sometimes the loudest voices for censorship have come not from within Europe but from within my own country, where the prior administration

> threatened and bullied social media companies to censor so-called misinformation – misinformation like, for example, the idea that coronavirus had likely leaked from a laboratory in China. Our own government encouraged private companies to silence people who dared to utter what turned out to be an obvious truth. So, I come here today not just with an observation but with an offer. And just as the Biden administration seemed desperate to silence people for speaking their minds, so the Trump administration will do precisely the opposite, and I hope that we can work together on that. In Washington, there is a new sheriff in town. And under Donald Trump's leadership, we may disagree with your views, but we will fight to defend your right to offer them in the public square, agree or disagree."

Die Ausführungen adressieren und kritisieren die Bemühungen der Plattformbetreiber zur Reduktion von Desinformationen direkt und stellen sie in den Kontext von Zensur. Wie skizziert, haben die großen Plattformen entsprechende Konsequenzen gezogen, um der von JD Vance markierten Linie besser zu entsprechen. Mit dieser Entwicklung korrespondiert auch ein bei vielen Vertreter:innen der Silicon Valley-Tech-Elite vorhandenes techno-libertäres Verständnis von Kommunikations- und Meinungsfreiheit (vgl. Golumbia 2024), das sich am US-Konzept der *free speech* orientiert, welches sich vom deutschen Konzept der Meinungsfreiheit deutlich unterscheidet. In den USA garantiert das *First Amendment* eine nahezu absolute Meinungsfreiheit, wobei der Staat nur in extremen Ausnahmefällen eingreifen darf. Einschränkungen gelten im Wesentlichen nur für sehr spezifische Fälle wie etwa Aufrufe zu unmittelbarer Gewalt oder Betrug. Selbst extremistische, beleidigende oder unwahre Aussagen fallen meist unter den Schutz der *free speech*. In Deutschland ist die Meinungsfreiheit durch Artikel 5 des Grundgesetzes geschützt. Allerdings steht sie unter dem Vorbehalt allgemeiner Gesetze und kann zum Beispiel beim Schutz der Jugend, der persönlichen Ehre oder bei Volksverhetzung erheblich eingeschränkt werden, da historische Erfahrungen, insbesondere mit dem Nationalsozialismus, strengere Grenzen nahelegen (vgl. Wissenschaftliche Dienste des Deutschen Bundestages 2018). Die Anlehnung an das Konzept der *free speech* wird in Aussagen Musks deutlich, in denen er sich als Verteidiger absoluter Meinungsfreiheit geriert (vgl. Pao 2022).

Das libertäre Grundverständnis von Kommunikation dieser Prägung basiert auf einem Dualismus aus einerseits dem Diktum maximaler uneingeschränkter freier Rede sowie andererseits der Haltung, dass jedes staatli-

che Eingreifen in die Selbstverständigung der Gesellschaft, sei es durch Fakten-Checks, etwaige externe Steuerung und Moderation von Diskursen oder gar Löschung von Inhalten, nichts anderes sei als Zensur. Mit der Digitalisierung erleben libertäre Prinzipien eine ideologische und praktische Renaissance. Zu diesen Prinzipien gehört die Vorstellung maximaler individueller Redefreiheit. Jede Person hat – so das Diktum – das Recht, ihre Meinung frei von jedweder staatlichen Kontrolle oder Zensur zu äußern. Damit korrespondiert eine Ablehnung staatlicher Eingriffe in das mediale Geschehen. Medien haben die Aufgabe, Informationen und Meinungen frei und ohne staatliche Kontrolle zu verbreiten (vgl. Dahlberg 2017; Siebert/Peterson/Schramm 1963). Der Staat darf keine Kontrolle über Inhalte ausüben. Jedweder Eingriff wird als Zensur betrachtet.

Im Bereich des Internets fordern libertäre Positionen ein „hands off the internet" – also minimale Gesetzgebung, um die maximale Meinungsfreiheit im digitalen Raum nicht zu gefährden, selbst bei kontroversen oder problematischen Inhalten (vgl. Coe 2018; Dahlberg 2017; Thierer/Szoka 2009). Unregulierte Social Media entsprechen diesem Ideal einer möglichst freien Infrastruktur, in der staatliche Kontrolle kaum vorhanden ist. Im libertären Medienverständnis soll zudem ein freier Medienmarkt durch Angebot und Nachfrage Pluralität und Qualität sichern. Nutzer:innen entscheiden selbst, welche Inhalte sie lesen oder verbreiten (vgl. Coe 2018).

An dieser Stelle knüpft ein ebenso wichtiges wie kritisches Moment des libertären Medien- und Kommunikationsverständnisses an, nämlich die Annahme einer Rationalität der Öffentlichkeit. Die Öffentlichkeit und ihre Individuen werden als fähig angesehen, im offenen und ökonomisch gedachten ‚Markt der Ideen' aus verschiedenen Informationen auszuwählen und deren Güte, individuellen Nutzen und letztlich auch deren Wahrheitsgehalt selbst und ohne fremde Hilfe einschätzen zu können (vgl. Siebert et al. 1963). Die zeitgenössische Variante dieser Denkfigur mit Blick auf Social Media ist nun die, dass sich Plattformen und Communities, also digitale Öffentlichkeiten, selbst regeln. Der Staat soll bei diesem Prozess der Selbstregulierung prinzipiell keine lenkende Rolle einnehmen – auch dann nicht, wenn es sich, wie etwa in Deutschland, um ein System staatsferner Medienkontrolle handelt. Bezüglich der Rationalität der Öffentlichkeit ist jedoch erhebliche Skepsis angebracht, da nicht nur die weiter oben skizzierten Funktionsweisen der Plattform und medienpolitischen Ausrichtungen der Betreibenden *per se* bereits digitale Öffentlichkeit in ihren inhaltlichen Tendenzen deutlich vorprägen, sondern auch, weil die für die Subjekte einer kritischen und informierten digitalen Öffentlichkeit notwendigen Kompe-

tenzen nicht von selbst emergieren. Kritische Medienkompetenz sind an dieser Stelle wesentlicher Teil einer informationellen Selbstbestimmung, die erlernt werden will. Nutzer:innen benötigen umfassende Medienkompetenz, um selbstbestimmt und kritisch mit den vielfältigen Angeboten und Manipulationsmöglichkeiten auf sozialen Plattformen umzugehen (vgl. Richter-Boisen/Mertens 2023) – eine Forderung, die im Gegensatz zur libertären Idee der Rationalität der Masse steht. An dieser Stelle tritt die Frage nach kritischer Medienkompetenz als normativem Zielwert ebenso ins Blickfeld wie die Frage nach der Rolle technischer Assistenzsysteme im Bereich Fact Checking.

3. Technische Assistenzsysteme und Desinformation: Das Projekt NEBULA

Plattformen tun derzeit noch wenig, um Desinformationen effektiv einzudämmen. Wie skizziert, sind augenblicklich sogar gegenteilige Tendenzen sichtbar. Somit geraten technische Lösungen zur Detektion von Desinformationen stärker ins Blickfeld (vgl. Lahby et al. 2022), mit denen eine Erkennung, Löschung und Reduktion unwahrer Inhalte vorangetrieben werden kann. Ein weiterer Zielwert ist die Erhöhung von Medienkompetenz, damit Nutzende Informationen einordnen, Desinformationen besser erkennen, kontextualisieren, reduzieren helfen und somit selbstbestimmt, informiert und resilient an der Gesellschaft partizipieren können (vgl. Funiok 2020).

3.1. NEBULA – Ausgangspunkte und Ziele von Forschung und App-Entwicklung

Das Projekt NEBULA adressiert mit der wertbasierten Entwicklung eines Mobile App-Demonstrators zur KI-unterstützten Identifikation von Desinformationen konkret die Dimension der Erhöhung von Medienkompetenz. Um dem Anspruch gerechter zu werden, allen „sozialen Gruppen eine gleichberechtigte Teilhabe am Selbstverständigungsprozess der Gesellschaft zu ermöglichen“ (vgl. Röben 2013: 10), fokussiert NEBULA dabei auf Angehörige vulnerabler Gruppen, nämlich ältere Menschen (vgl. Shrestha/Spezzano 2019), Jugendliche (vgl. Seo et al. 2021) und Migrant:innen (vgl. Ruokolainen/Widén 2020), da diese in vielen Lebensbereichen und so auch in ihrer Medienrezeption strukturellen Benachteiligungen ausgesetzt sind

(vgl. Gomolla 2016). Für sie können sich auch die möglichen negativen Folgen der Nutzung von digitalen Medien besonders auswirken, also ebenso die negativen Effekte von Desinformationen in Social Media. NEBULA kombiniert dabei Technologieentwicklung und qualitative Begleitforschung in den drei vulnerablen Gruppen. Im Rahmen mehrerer iterativer Schleifen aus Konzeption, technischer Entwicklung und Begleitforschung im Sinne des *Value Sensitive Design* (VSD) sollen Kompetenzförderung gestärkt, Vertrauen in die Technologie erhöht sowie Reaktanzen reduziert werden. *Value Sensitive Design* (VSD) ist ein ethisch fundiertes Konzept, um menschliche Werte systematisch in Technologieentwicklungsprozesse zu integrieren und adressiert unter anderem die „vielschichtigen Benachteiligungen insbesondere von Minderheiten, die Technikentwicklungen wissentlich, willentlich oder völlig unbeabsichtigt nach sich ziehen" (vgl. Hillerbrand 2021: 469). Eine zentrale Prämisse ist dabei, dass Technologie nicht *per se* wertneutral ist, sondern stets bestimmte Werte transportiert, handlungsrahmende Affordanzen aufweist und somit die Nutzung auf bestimmte Weise zwar nicht determiniert, zwangsläufig aber vorprägt. Wie Hillerbrand (2021) weiter betont, ist es im VSD daher besonders relevant, Werte bereits früh in den Designprozess einzubringen. Dies umfasst universelle Mindestwerte, wie Wohlergehen, Gerechtigkeit und Würde als normative Basis sowie weitere fallbezogen zu konkretisierende Zielwerte (vgl. Friedman/Hendry 2019), wie zum Beispiel Nachhaltigkeit (vgl. Asikis et al. 2021) – auch wenn es bei einem Wertepluralismus auch miteinander in Konflikt stehende Werte geben kann (vgl. Jacobs/Huldtgren 2021). Da Werte sich mit Zeit und Technologie zudem verändern, ist VSD im Prinzip stets iterativ und reflexiv angelegt (Friedman et al. 2013). VSD ist grundsätzlich als dreistufiger Prozess konzeptualisiert, der wiederum mehrere iterative Schleifen beziehungsweise Zyklen durchlaufen kann. In diesen Zyklen werden in einem multimethodischen Programm drei Ebenen aufeinander bezogen:

a. Konzeptuelle Untersuchungen zum in Rede stehenden Phänomenbereich inkl. Identifikation von Stakeholdern (direkt und indirekt), einer Klärung, welche Werte im Kontext relevant sind sowie einer Analyse möglicher Zielkonflikte zwischen verschiedenen Werten;
b. Empirische Untersuchungen mit Methoden der qualitativen und quantitativen Sozialforschung zu Werten, Bedürfnissen und Erfahrungen der Stakeholder, etwa in Form von Interviews, Umfragen oder Workshops (Cruz-Martínez et al. 2021; Winkler/Spiekermann 2021)

c. Technikentwicklung von Prototypen beziehungsweise Demonstratoren unter Einbeziehung der Ergebnisse der Begleitforschung. Dieser Zyklus wird mehrmals durchlaufen, um Werte frühzeitig und fortlaufend in die Technikentwicklung zu integrieren.

Die Kernergebnisse der ersten Phase der Begleitforschung umfassen qualitative Ergebnisse aus einer Interview-Studie zu den Kategorien (K1) Wissen und Assoziationen zu Cybersicherheit und Desinformation, (K2) Eigene und vermittelte Erfahrungen, (K3) Medienrepertoire, Medienpraktiken, Medienkompetenzen, (K4) Einschätzungen zur gesellschaftlichen Relevanz von Cybersicherheit und Desinformation, (K5) Unterschiede und Gemeinsamkeiten Deutschland / Herkunftsland sowie (K6) Bedarfe und Erwartungen bezüglich technischer Assistenzsysteme, wobei im vorliegenden Kontext in erster Linie Ergebnisse zur letzten dieser Kategorien relevant sind (im Detail siehe Anastasiadis et al. 2025). Die Kategorie umfasst Aussagen zur grundlegenden Bedienbarkeit sowie zu Form und Inhalt des Output / Feedback technischer Assistenzsysteme. Aus diesen Aussagen ließ sich entnehmen, dass vor allem alle erklärenden Texte, die Grundfunktionalität, das Interface sowie der Output der App, also die Ergebnisdarstellung und Transparenz, nachvollziehbar und verständlich sein sollen. Diese Hinweise sind, wie weiter unter konkretisiert, unmittelbar in die weitere App-Entwicklung eingeflossen. In den Ergebnissen wurde zudem – und dies ist für die konkrete App-Entwicklung ebenfalls zentral – klar, welche Relevanz Vertrauen und Misstrauen in Technologie haben, weswegen eine induktive Querschnittskategorie zu diesem Themenfeld entwickelt werden konnte. Dabei wurden für die konkrete Systementwicklung wesentliche Aspekte deutlich. Zum einen wird gefordert, dass die Systeme vertrauenswürdig sein sollen; zum anderen präzisieren die Proband:innen, auf welche Weise aus ihrer Sicht dieses Vertrauen entsteht würde. Manche betonen, dass eine Anbindung an offizielle Institutionen – etwa staatliche Stellen – und damit eine institutionelle Legitimierung vertrauensfördernd wirken würde. Ebenso wird deutlich, dass Forschungseinrichtungen und Universitäten als vertrauenswürdige Akteure wahrgenommen werden. Zusätzlich wird erwartet, dass die Systeme ihre genutzten Quellen offenlegen. Neben dieser institutionellen Verankerung wird auch die Forderung erhoben, Vertrauenswürdigkeit durch eine klare Kennzeichnung sichtbar zu machen, beispielsweise in Form von an Gütesiegel erinnernde Markierungen von Inhalten. Während viele Teilnehmende die Anbindung an deutsche Institutionen als vertrauensbildend hervorheben, äußern einige gleichzeitig deutliche

Vorbehalte gegenüber staatlichen Strukturen in ihren Herkunftsländern. Staatliche Institutionen werden dort häufig direkt mit der Regierung gleichgesetzt, der wiederum Misstrauen entgegengebracht wird – insbesondere bei Proband:innen aus autoritären Regimen. Im folgenden Abschnitt wird konkretisiert, wie diese Ergebnisse in die weitere Entwicklung eingeflossen sind, um Vertrauen und Transparenz zu erhöhen und um Reaktanzen abzubauen beziehungsweise vorzubeugen.

3.2. Iterativer Ergebnistransfer von der Begleitforschung in die App-Entwicklung

Auf Basis der oben skizzierten Ergebnisse aus der ersten Phase der Begleitforschung wurden in der dann folgenden Entwicklungsphase die technikethisch relevanten Dimensionen der Validität, Aktualität / Adaption, Verständlichkeit, Transparenz, Reziprozität und Kompetenzvermittlung besonders priorisiert.

Validität, verstanden als verlässliche Erkennung von Desinformation durch technische Systeme, ist eine Grundvoraussetzung ethisch vertretbarer Detektionssysteme, denn unzuverlässige Detektion kann erhebliche negative Folgen haben, etwa indem sie legitime Beiträge als Desinformation klassifiziert (*false positive*) oder tatsächliche Desinformation nicht identifiziert (*false negative*). Im Fall von NEBULA werden bei der Messung der Validität drei Klassen von Aussagen unterschieden: (1) korrekte Aussagen, (2) falsche Aussagen und (3) Aussagen, für die nicht genug Informationen vorliegen, um eine Entscheidung zu treffen. In Tests, in denen die automatische Detektion die für die Entscheidung wichtigen Hintergrundinformationen erhält, erreicht NEBULA eine hohe Genauigkeit von ca. 90 %. In der Praxis stellt dieser vorgelagerte Schritt, also die Sammlung der richtigen Hintergrundinformationen zum Treffen einer Entscheidung, eine große Herausforderung dar. So kam die Detektion zu Beginn des Projekts mit 36 Prozent auf eine Genauigkeit, die nur ein wenig besser als die 33,3 Prozent einer zufälligen Antwort war (gemäß der drei genannten Aussagetypen ‚korrekt', ‚falsch' oder ‚nicht genug Informationen'). Dies konnte im Verlauf des Projekts gesteigert werden, so dass 75 Prozent der richtigen Aussagen als solche erkannt werden. Die beiden anderen Klassen können in ca. 80 Prozent der Fälle korrekt von richtigen Aussagen unterschieden werden. Die Unterscheidung zwischen falschen Aussagen und solchen, die vom System durch fehlende Informationen nicht beantwortet werden können, ist

jedoch schwierig und reduziert die Gesamtgenauigkeit auf ca. 53 Prozent. Dies liegt vor allem daran, dass in vielen Fällen keine Evidenz gefunden wird, um eine Aussage als falsch zu identifizieren.

Die Anpassung des Detektionsalgorithmus und die Implementierung weiterer Referenzkorpora sollen diese Werte jedoch weiter verbessern. Algorithmen zur Erkennung von Desinformation sollten sich idealerweise laufend aktuellen Entwicklungen anpassen, was aus technik- und medienethischer Sicht eine wesentliche Dimension der Systemgestaltung ist (vgl. Mittelstadt et al. 2016). Nur adaptive Systeme können der schnellen Evolution von Desinformationen gerecht werden, indem sie neue Akteure, Themenfelder und Strategien zu erkennen in der Lage sind. Im vorliegenden Kontext lässt sich die Aktualität durch die Einspeisung neuer Referenzkorpora und die Implementierung neuer Indikatoren realisieren. Die Punkte Verständlichkeit und Transparenz sind in Phase 2 des iterativen Prozesses von besonderer Bedeutung. Verständlichkeit ist ein zentrales Qualitätsmerkmal digitaler Kommunikationssysteme. Wenn Interface, Texte und Outputs nachvollziehbar, adressatengerecht und sprachlich verständlich sind, steigt die Wahrscheinlichkeit gelingender Nutzerorientierung und Kompetenzvermittlung.

Zugleich sollten die Grundlagen der Detektion und Gründe für die Klassifikation als Desinformation transparent nachvollziehbar sein. Transparenz – verstanden als Offenlegung der verwendeten Entscheidungsgrundlagen und Detektionskriterien – gehört zu den elementaren ethischen Anforderungen für automatisierte Systeme zur Desinformationsdetektion (vgl. Mittelstadt et al. 2016).

Verständlichkeit und Transparenz wurden im Rahmen des iterativen Prozesses als Kernpunkte identifiziert und in der Entwicklung priorisiert, um der in Phase 1 der Begleitforschung bestätigten hohe Relevanz von Fragen rund um das Vertrauen in die NEBULA-App zu begegnen. Im Rahmen der Weiterentwicklung der NEBULA-App wurde eine Reihe gezielter Maßnahmen implementiert, um die Benutzerfreundlichkeit (*usability*), Transparenz und Vertrauenswürdigkeit zu erhöhen und mögliche Reaktanzen zu verringern, denn diese Aspekte sind zentrale Qualitätskriterien für digitale Anwendungen im Kontext der Informationsvermittlung (vgl. Nielsen 2010):

a. Zunächst wurde das Onboarding, also der erklärende Einstieg in die erstmalige Nutzung der App, umfassend ausgebaut, um einen strukturierteren Einstieg in die Funktionalitäten der App zu ermöglichen.

b. Parallel hierzu erfolgte eine Anpassung und Vereinfachung sämtlicher Texte, um eine klare, nutzerorientierte Kommunikation sicherzustellen. Dies entspricht dem Prinzip der kognitiven Ergonomie und ist ein zentrales Element benutzerfreundlicher Systeme (vgl. ISO 2019).
c. Zur Förderung von Barrierefreiheit und Inklusion wurden zwei zentrale Optionen ergänzt, nämlich die Integration einer Multilingualitätsfunktion, um eine Nutzung in mehreren Sprachen zu ermöglichen, sowie die Implementierung einer Einstellung für einfache Sprache, um die Zugänglichkeit für Personen mit unterschiedlichen sprachlichen Kompetenzen sicherzustellen. Solche Maßnahmen können potenzielle Exklusionsrisiken minimieren.
d. Darüber hinaus wurde Funktionalität und Nutzungsanforderungen differenzierter erklärt. Auch wurde eine umfassende Erklärung der Funktionsweise und Prüfprozesse der zugrunde liegenden Algorithmen implementiert. So wird eine vertiefte Auseinandersetzung mit Desinformationen und den der App zugrunde liegenden algorithmischen Prüfprozessen ermöglicht.

Durch die Offenlegung der methodischen Grundlagen und Bewertungsverfahren wird dem Prinzip der algorithmischen Transparenz Rechnung getragen, das als entscheidend für die Akzeptanz algorithmischer Entscheidungssysteme gilt (vgl. Diakopoulos 2016). Die Darstellung der Faktencheck-Ergebnisse wurde sowohl grafisch als auch sprachlich überarbeitet. Ziel war eine visuell klarere und sprachlich verständlichere Präsentation. In Kombination mit der deutlichen Ausweitung erklärender Texte und Hintergrundinformationen soll dies zur Erhöhung der Transparenz und Verständlichkeit beitragen, die wiederum das Vertrauen in die App stärken sollen. Ein weiterer Schwerpunkt zur Erhöhung des Vertrauens, der sich ebenfalls unmittelbar aus den Ergebnissen der qualitativen Begleitforschung ableitet, lag auf der Transparenz bezüglich der institutionellen, finanziellen und politischen Rahmenbedingungen der App-Entwicklung. Dies wurde durch die Markierung der institutionellen Herkunft der App, also des Fördergebers sowie die Benennung der an der Entwicklung beteiligten Akteure aus Forschung und Entwicklung realisiert. Schließlich wird nun auch die Finanzierungsgrundlage der App sowie ihre politische Unabhängigkeit und Neutralität explizit ausgewiesen.

Eine weitere unter technikethischen Gesichtspunkten relevante Dimension ist die Reziprozität technischer Systeme (vgl. Rath 2019). Dies meint, dass Systeme Eingaben und Interaktionen von Nutzenden aufnehmen und

für ihre Weiterentwicklung nutzen. In medienethischer Perspektive wird dies als eine Möglichkeit für Vertrauensaufbau, Nutzerorientierung und gesellschaftliche Teilhabe gesehen – mit dem Ziel, Systeme gemeinschaftlich zu verbessern und eine Reflexion über technische wie soziale Nebenfolgen sicherzustellen (vgl. Mittelstadt et al. 2016). Gleichwohl sind Entwicklungskapazitäten auch im NEBULA-Kontext nicht unbegrenzt und in konkrete Technologie-Design-Entscheidungen eingerahmt. Da der Fokus auf der Erkennung von Desinformationen und in der Ausgabe möglichst valider Informationen für die weitere Entscheidungsfindung von Nutzer:innen liegt, sind keinerlei Crowd Sourcing-Funktionalitäten und somit keine im engeren Sinne auf Reziprozität zielenden Funktionen vorgesehen. Es wird auch keine Community-Features in der App geben.

Die NEBULA-App soll nicht nur Hinweise auf mögliche Desinformationen ausgeben, sondern auch die Fähigkeit stärken, Desinformation eigenständig zu erkennen und mit diesen resilient umzugehen, also Medienkompetenzen vermitteln. Dies entspricht dem Ziel, Nutzer:innen zu befähigen, kritisch, informiert und autonom mit digitalen Medien umzugehen (vgl. Funiok 2020). Inwieweit dies gelingt, muss im Wesentlichen in konkreten Aneignungsstudien nach Fertigstellung der App untersucht werden.

4. Fazit und Ausblick

Ausgangspunkt des Projekts NEBULA ist die normative Annahme, dass Desinformationen eine substanzielle Gefährdung für demokratische Prozesse und die informierte Partizipation an der Gesellschaft darstellen. Desinformation unterminiert die öffentliche Meinungsbildung, destabilisiert diskursive Räume und kann das Vertrauen in demokratische Institutionen sowie Willensbildung nachhaltig beschädigen. Vor diesem Hintergrund ist die Entwicklung von Abwehrstrategien gegen Desinformation nicht nur technisch möglich, sondern auch normativ geboten. Sie dient der Sicherung demokratischer Prinzipien politischer Partizipation in der (digitalen) Öffentlichkeit. Ziel des Projekts ist die Kompetenzsteigerung vulnerabler Gruppen im Umgang mit Desinformationen, da diese Gruppen aufgrund geringerer Medienkompetenz, eingeschränkter Zugänge oder sprachlicher Barrieren in besonderer Weise gefährdet sind, Desinformationen nicht adäquat erkennen oder verarbeiten zu können. Mit Blick auf die Rahmenbedingungen digitaler Öffentlichkeiten bleibt die Entwicklung von technischen Assistenzsystemen für valides Fact Checking eine wichtige Aufgabe

für Forschung und Entwicklung, eben weil jüngste politische Entwicklungen zumindest eine temporäre Renaissance libertärer Prinzipien erkennen beziehungsweise aus normativ-deliberativer Perspektive befürchten lassen. Trotz Limitationen, wie vor allem die Reduktion auf Texterkennung, können automatisierte Fact Checking-Tools wie NEBULA wichtige Unterstützung für eine informierte und resiliente Praxis im Umgang mit Desinformationen bieten.

Literatur

Allcott, Hunt / Gentzkow, Matthew (2017): Social Media and Fake News in the 2016 Election, in: Journal of Economic Perspectives 31 (2/2017). https://doi.org/10.1257/jep.31.2.211

Anastasiadis, Mario et al. (2025): Sicherung sozialer Nachhaltigkeit durch Technologie? Eine qualitative Studie zu technischen Assistenzsystemen im Bereich Cybersicherheit, in: Vanessa Kokoschka et al. (Hg.), Nachhaltigkeit in der Medienkommunikation, Baden-Baden, S. 305–320.

Asikis, Thomas et al. (2021): How Value-Sensitive Design Can Empower Sustainable Consumption, in: Royal Society Open Science 8 (1/2021). https://doi.org/10.1098/rsos.201418

Becker, Gunter (2025): Meta stoppt Fact Checking: Wer schützt die Wahrheit, wenn die Journalist:innen gehen?, in: Deutscher Fachjournalisten Verband, 19. Februar 2025 (online unter: https://dfjv.de/publikationen/fachjournalist/meta-stoppt-fact-checking-wer-schuetzt-die-wahrheit-wenn-die-journalistinnen-gehen – letzter Zugriff: 23.10.2025).

Bernhard, Lukas et al. (2024): Verunsicherte Öffentlichkeit. Superwahljahr 2024: Sorgen in Deutschland und den USA wegen Desinformationen, hrsg. von der Bertelsmann Stiftung, Gütersloh.

Broschart, Steven (2024): Putins digitale Front und die Wahrheit dahinter, Wiesbaden.

Bruns, Axel (2019): Are filter bubbles real? Digital futures, Cambridge.

Coe, Peter (2018): (Re)embracing social responsibility theory as a basis for media speech: shifting the normative paradigm for a modern media, in: Northern Ireland Legal Quarterly 69 (4/2018). https://doi.org/10.53386/nilq.v69i4.186

Cruz-Martínez, Roberto Rafael et al. (2021): Toward the Value Sensitive Design of eHealth Technologies to Support Self-Management of Cardiovascular Diseases: Content Analysis, in: JMIR Cardio 5 (2/2021). https://doi.org/10.2196/31985

Dahlberg, Lincoln (2017): Cyberlibertarianism, in: Oxford Research Encyclopedia of Communication, Oxford.

Delcker, Janosch (2022): Twitter: Sorge über Entlassungen von Content-Moderatoren, 17. November 2022 (online unter: https://www.dw.com/de/twitter-sorge-%C3%BCber-entlassungen-von-content-moderatoren/a-63792800 – letzter Zugriff: 23.10.2025).

Diakopoulos, Nicholas (2016): Accountability in Algorithmic Decision Making, in: Communications of the ACM 59 (2/2016), S. 56–62.

van Dijck, José / Poell, Thomas (2013): Understanding Social Media Logic, in: Media and Communication 1 (1/2013), S. 2–14.

Duffy, Clare (2025): Meta Gets Rid of Fact Checkers and Says It Will Reduce 'Censorship', in: CNN Business, 07. Januar 2025 (online unter: https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation – letzter Zugriff: 23.10.2025).

Friedman, Batya / Hendry, David (2019): Value Sensitive Design: Shaping Technology with Moral Imagination, Cambridge.

Friedman, Batya et al. (2013): Value Sensitive Design and Information Systems, in: Neelke Doorn et al. (Hg.), Dordrecht Early engagement and new technologies: Opening up the laboratory (= Philosophy of Engineering and Technology, Bd. 16), Dordrecht, S. 55–95.

Fries, Fabian (2021): Die Ränder der (Pseudo-)Wissenschaft: umstrittene Wissenskonzeptionen zwischen Avantgarde und Häresie, Basel.

Funiok, Rüdiger (2020): Verantwortliche Mediennutzung: Wünschenswerte Selbstverpflichtungen von Rezipient*innen und Nutzer*innen, in: Communicatio Socialis 53 (2/2020). https://doi.org/10.5771/0010-3497-2020-2-136

Gillespie, Tarleton (2014): The Relevance of Algorithms, in: Tarleton Gillespie / Pablo J. Boczkowski / Kirsten A. Foot (Hg.), Media Technologies, Cambridge, S. 167–194.

Golumbia, David (2024): Cyberlibertarianism: the right-wing politics of digital technology, Minneapolis.

Gomolla, Mechtild (2016): Direkte und indirekte, institutionelle und strukturelle Diskriminierung, in: Albert Scherr / Aladin El-Mafaalani / Emine Gökcen Yüksel (Hg.), Handbuch Diskriminierung, Wiesbaden, S. 1–23.

Haarkötter, Hektor (2021): Wahrheit und Lüge im (außer-)journalistischen Sinne, in: Christian Schicha / Ingrid Stapf / Saskia Sell (Hg.), Medien und Wahrheit, Baden-Baden, S. 317–340.

Habermas, Jürgen (1990): Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft, Frankfurt am Main.

Habermas, Jürgen (2022): Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Berlin.

Henzler, Piotr (2025): Wir leben in einer Welt der Fake News, in: Goethe-Institut (online unter: https://www.goethe.de/ins/pl/de/kul/med/rfn/zsfn.html – letzter Zugriff: 23.10.2025).

Hillerbrand, Rafaela (2021): Value Sensitive Design, in: Armin Grunwald / Rafaela Hillerbrand (Hg.), Handbuch Technikethik, Stuttgart, S. 466–471.

ISO (2019): Part 210: Human-Centred Design for Interactive Systems, ISO 9241–210:2019, in: Ergonomics of Human-System Interaction, Juli 2019 (online unter: https://www.iso.org/standard/77520.html#lifecycle – letzter Zugriff: 23.10.2025).

Jacobs, Naomi / Huldtgren, Alina (2021): Why Value Sensitive Design Needs Ethical Commitments, in: Ethics and Information Technology 23 (1/2021), S. 23–26. https://doi.org/10.1007/s10676-018-9467-3

Laaff, Meike (2025): Faktencheck bei Meta: Er spart sich das einfach, in: Die Zeit, 8. Januar 2025 (online unter: https://www.zeit.de/digital/2025-01/faktencheck-meta-mark-zuckerberg-moderation-instagram-facebook – letzter Zugriff: 23.10.2025).

Lahby, Mohamed et al. (Hg.) (2022): Combating Fake News with Computational Intelligence Techniques (= Studies in Computational Intelligence, Bd. 1001), Cham.

McIntosh, Leslie D. / White, William / Hudson Vitale, Cynthia (2023): Unveiling Deception: Establishing a Taxonomic Framework for Disinformation within Scientific Discourse, Ithaca.

Mittelstadt, Brent Daniel et al. (2016): The Ethics of Algorithms: Mapping the Debate, in: Big Data & Society 3 (2/2016). https://doi.org/10.1177/2053951716679679

Nielsen, Jakob (2010): Usability Engineering, Amsterdam, Heidelberg.

Pao, Ellen K. (2022): Elon Musk's Vision of 'Free Speech' Will Be Bad for Twitter, in: The Washington Post, 8. April 2022 (online unter: https://www.washingtonpost.com/outlook/2022/04/08/musk-twitter-equity-discrimination-speech/ – letzter Zugriff: 23.10.2025).

Pariser, Eli (2017): Filter Bubble: Wie wir im Internet entmündigt werden, München.

Rath, Matthias (2019): Zur Verantwortungsfähigkeit künstlicher ‚moralischer Akteure': Problemanzeige oder Ablenkungsmanöver?, in: Matthias Rath / Friedrich Krotz / Matthias Karmasin (Hg.), Maschinenethik: Ethik in mediatisierten Welten, Wiesbaden, S. 223–242.

Richter-Boisen, Anette / Mertens, Claudia (2023): Individuelle und kollektive Folgen von Social Media-Plattformen aus Sicht der Medienpädagogik, in: Medienimpulse 61 (4/2023). https://doi.org/10.21243/mi-04-23-19

Röben, Bärbel (2013): Medienethik und die „Anderen": Multiperspektivität als neue Schlüsselkompetenz, Wiesbaden.

Ruokolainen, Hilda / Widén, Gunilla (2020): Conceptualising Misinformation in the Context of Asylum Seekers, in: Information Processing & Management 57 (3/2020). https://doi.org/10.1016/j.ipm.2019.102127

Sato, Yuko / Wiebrecht, Felix (2024): Disinformation and Regime Survival, in: Political Research Quarterly 77 (3/2024). https://doi.org/10.1177/10659129241252811

Schwaiger, Lisa (2022): Gegen die Öffentlichkeit: Alternative Nachrichtenmedien im deutschsprachigen Raum, in: Publizistik 67 (4/2022). https://doi.org/10.1007/s11616-022-00752-w

Seeliger, Martin / Sevignani, Sebastian (Hg.) (2021): Ein neuer Strukturwandel der Öffentlichkeit? (= Sonderband Leviathan 37/2021), Baden-Baden.

Seo, Hyunjin et al. (2021): Vulnerable Populations and Misinformation: A Mixed-Methods Approach to Underserved Older Adults' Online Information Assessment, in: New Media & Society 23 (7/2021). https://doi.org/10.1177/1461444820925041

Shrestha, Anu / Spezzano, Francesca (2019): Online Misinformation: From the Deceiver to the Victim, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, Vancouver, S. 847–850.

Siebert, Fred S. / Peterson, Theodore / Schramm, Wilbur (1963): Four Theories of the Press: The Authoritarian, Libertarian, Social Responsibility, and Soviet Communist Concepts of What the Press Should Be and Do, Champaign, Urbana.

Sunstein, Cass R. (2001): Echo Chambers: Bush v. Gore, Impeachment, and Beyond, Princeton.

Thierer, Adam / Szoka, Berin (2009): Cyber-Libertarianism: The Case for Real Internet Freedom, 12. August 2009 (online unter: https://techliberation.com/2009/08/12/cyber-libertarianism-the-case-for-real-internet-freedom/ – letzter Zugriff: 23.10.2025).

Weatherbed, Jess (2025): Zuckerberg, inspired by Musk, ditches fact checking for Community Notes", in: The Verge, 7. Januar 2025 (online unter: https://www.theverge.com/2025/1/7/24338062/facebook-instagram-threads-meta-abandon-fact-checking – letzter Zugriff: 23.10.2025).

Winkler, Till / Spiekermann, Sarah (2021): Twenty Years of Value Sensitive Design: A Review of Methodological Practices in VSD Projects, in: Ethics and Information Technology 23 (1/2021). https://doi.org/10.1007/s10676-018-9476-2

Wissenschaftliche Dienste des Deutschen Bundestages (2018): Zum Schutz der Meinungsfreiheit in Deutschland und in den USA(= WD 3 – 3000 – 052/18) (online unter: https://www.bundestag.de/resource/blob/556742/b5134f621e8813c184fce1a82cb0df9e/wd-3-052-18-pdf-data.pdf – letzter Zugriff: 23.10.2025).

Zimmermann, Fabian / Kohring, Matthias (2018): ‚Fake News' als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs, in: Medien & Kommunikationswissenschaft 66 (4/2028). https://doi.org/10.5771/1615-634X-2018-4-526

III.
Öffentliche Sprache: Ethische Perspektiven auf demokratische Debattenkultur

Ausgewählte Perspektiven auf den Einsatz von Künstlicher Intelligenz in den Medien

Petra Grimm, Susanne Kuhnert und Marcel Schlegel

Zusammenfassung

Der Beitrag präsentiert Erkenntnisse aus den Forschungsprojekten „IKID" und „GEISST", die sich mit dem Einsatz von Stimmen, die mit Künstlicher Intelligenz generiert wurden, in den Medien befassen. Zum einen wird aufgezeigt, wie KI-Kompetenz im Rahmen einer interdisziplinären Ausbildung von Medienstudierenden gefördert werden kann. Zum anderen werden ethische Leitlinien für den verantwortungsvollen Umgang mit KI-Stimmen in Medienorganisationen vorgestellt.

1. Einleitung

Künstliche Intelligenz (KI) beschäftigt gegenwärtig vor allem Politik, Wirtschaft und Wissenschaft (vgl. Menzel/Winkler 2018). Gerade Generative KI (GenKI), mit der Texte, Bilder, Videos, Code und Daten auf synthetischem Wege erzeugt werden kann, verbreitet sich rasant. Ihre einfache und (partiell) kostenfreie Nutzung, die menschliche Kommunikation nachahmt, ermöglicht auch Laien einen Zugang zu KI-Technik. Dergestalt dringen GenKI-Tools in immer mehr Lebensbereiche ein. Aufgrund ihrer Funktionalität erreichen sie auch das Medien- und Bildungssystem und fordern dortige Strukturen heraus.

Im Folgenden werden zwei Forschungsprojekte aus diesen Bereichen vorgestellt, die am *Institut für Digitale Ethik (IDE)* umgesetzt wurden: „GEISST – Generator für emotional individualisierbare Synthetik-Stimmen" und „IKID – Interdisziplinäres KI-Exploratorium: Integrierte Lehre zur verantwortungsvollen Nutzung Künstlicher Intelligenz auf Basis physisch-virtueller Demonstratoren".[1] Beide eint die Frage, welche Auswirkun-

1 Beide Projekte wurden maßgeblich an der *Hochschule der Medien Stuttgart (HdM)* umgesetzt und vom *Bundesministerium für Forschung, Technologie und Raumfahrt*

gen GenKI auf das Mediensystem haben könnte und wie diese aus ethischer Sicht zu reflektieren sind. Auch konzentrieren sich die Projekte auf KI-Techniken zur artifiziellen Nachbildung menschlicher Stimmen, richten ihr Augenmerk jedoch auf unterschiedliche Zielgruppen: So zielte IKID darauf ab, Medienstudierende auf eine verantwortungsvolle KI-Nutzung vorzubereiten. Dafür arbeiteten Lehrende unterschiedlicher Disziplinen zusammen, um ein didaktisches Konzept zum Erwerb von interdisziplinär verstandener KI-Kompetenz[2] zu entwickeln (vgl. Braßler 2020: 11–14; Grimm 2025a). Aus diesem wurden Lehrveranstaltungen abgeleitet, die den Prinzipien der interdisziplinären Lehre folgen (vgl. Braßler 2020: 15). Anhand eines dieser Projektseminare, das die Nutzung von *voice clones* und *deep fakes* im Journalismus thematisiert, zeigt der Aufsatz, wie ethische Perspektiven einfließen können in die Ausbildung von Medienstudierenden – als jene künftigen Praktiker:innen, die die KI-Ausbreitung im Medien- und indirekt auch im Bildungssystem maßgeblich mitgestalten werden.

In GEISST standen hingegen Medienschaffende selbst im Mittelpunkt. Sie wurden zunächst befragt mit dem Ziel, KI-Folgen für Medienorganisationen aus der Perspektive der Betroffenen zu erfahren. Zuerst wurden Haltungen sowie bedrohte Standards eruiert, die anschließend in wertebasierte Leitlinien für einen ethisch sensiblen GenKI-Einsatz mit Fokus auf Sprachsynthese überführt werden konnten. Die Handlungsempfehlungen, die Medienschaffenden dabei helfen sollen, die ethische Legitimität eines KI-Einsatzes zu überprüfen, werden im Folgenden in komprimierter Form vorgestellt.[3]

2. KI und der neue Strukturwandel: Potenziale und Risiken

GenKI trifft das Mediensystem in einer Zeit, in der es ohnehin angeschlagen ist (vgl. Krämer 2021: 28–35). Wie Umfragen zeigen, hält sich das Ver-

(BMFTR) gefördert, IKID zudem vom *Ministerium für Wissenschaft, Forschung und Kunst BaWü (MWK)*.

2 Darunter wird ein Bündel an Befähigungen verstanden, das sich – abhängig vom Verwendungszusammenhang von KI-Technologie – aus unterschiedlichen Fähigkeiten, Fertigkeiten, Kenntnissen, Methoden- und Schlüsselkompetenzen zusammensetzt. In der konkreten Lehrpraxis sollen jene Kompetenzen aus den Disziplinen Informatik, Wirtschaft, Recht und Ethik vermittelt werden, die für eine ganzheitliche Bearbeitung eines spezifischen KI-basierten Problems notwendig sind (vgl. Schlegel et al. 2024).

3 Die Leitlinien von Grimm, Kuhnert und Schlegel (2025) sind in vollständiger Form einsehbar unter dem Link im Literaturverzeichnis.

trauen der deutschen Bevölkerung in Medien zwar auf einem konstanten Niveau und ist, relativ zu vergleichbaren Ländern, weiterhin hoch (vgl. Jakobs et al. 2022: 387–396). Dennoch vertraute im Schnitt der Jahre 2016 bis 2023 weniger als jede:r Zweite den Medien „eher" oder „voll und ganz"; jede:r Dritte gab an, medialer Berichterstattung nur „teils, teils" zu vertrauen (vgl. Quiring et al. 2024: 2). Da Vertrauen als fundamentaler Wert von menschlichen Beziehungen gelten kann und auch die Basis der Relation zwischen Journalismus und Publikum bildet (vgl. Bentele 2021: 73), kann diese Entwicklung zulasten der Glaubwürdigkeit der Medien gehen (vgl. Tschopp et al. 2022: 330–333). Überdies führt die wachsende Bedeutung von Online-Plattformen dazu, dass die *gatekeeper*- und *agenda-setter*-Rolle der Massenmedien graduell ausgehebelt wird, was eine abermalige Schwächung der funktionalen Legitimität klassischer Berichterstatter:innen – der Massenmedien als Hersteller von Öffentlichkeit, als Kontrolleure der Politik sowie als Werbe- und Wirtschaftsplattform – bedeuten kann (vgl. Jarren 2019: 164–170).

Weil Soziale Medien es Nutzenden ermöglichen, sich netzöffentlich zu äußern, gelangen zusehends auch Halb- und Unwahrheiten sowie, auch mit Kalkül gestreute, Desinformationen in eine politische Öffentlichkeit, die durch die Plattformöffentlichkeiten ohnehin fragmentiert wird. Habermas (2022) deutet diese Entwicklung als „neuen Strukturwandel der Öffentlichkeit" und erkennt darin eine Herausforderung für deliberative Demokratien. Bezüglich der öffentlichen Kommunikations-Infrastruktur problematisiert er, dass angesichts der Schwächung des traditionellen Journalismus die objektive Selektion von „relevanten und entscheidungsbedürftigen Themen" zunehmend ausbleibe und so die Erstellung „*qualitativ gefilterter* Meinungen nicht mehr [zu] gewährleisten" sei (ebd.: 65; Hervorhebungen im Original). Nicht allein Anzahl oder Frequenz, sondern die fehlende Wahrnehmung von Desinformation erzeuge eine „Deformation der Wahrnehmung der politischen Öffentlichkeit" (ebd.: 64). Schaden nehmen Demokratien dann, wenn freie Medien und der Qualitätsjournalismus, welche als „Vierte Gewalt" das politische System kontrollieren sollen, durch ein emergentes Zusammenwirken verschiedener Faktoren geschwächt werden: durch ökonomischen Druck auf die klassischen Medien; durch unfairen Wettbewerb mit den Plattformbetreibern; durch ein multifacettenreiches Aufkommen von (auch mit KI erzeugter und verbreiteter) Desinformation. Eine Dynamik, die den Vertrauensverlust der Bürger:innen in Medien und Politik beschleunigen könnte, weil Stabilität und Ordnung der liberalen Demokratie sowie die soziale Integration der Mitglieder einer Gesellschaft

wesentlich davon abhängen, dass sich Rezipient:innen darauf verlassen können, von öffentlichen Sprecher:innen gewissenhaft recherchierte Informationen zu erhalten. Weil Bürger:innen mediale Informationen für die eigene Meinungsbildung nutzen und diese gewissermaßen aus zweiter Hand erhalten, müssen sie sicher sein können, dass diese faktenbasiert und von Journalist:innen ohne bewusste Täuschungsabsicht verbreitet worden sind (vgl. Stapf 2021: 99–101). Deshalb stellen Wahrheit (als Wert) und Wahrhaftigkeit (als Eigenschaft von Kommunikator:innen) die zentralen Maximen des Journalismus dar (ebd. 2024: 315).

Angesichts der zügigen Verbreitung und Ausdifferenzierung von KI, deren Erzeugnisse auch über Soziale Medien gestreut werden können, sieht sich das Mediensystem erneut mit strukturellen Veränderungen konfrontiert, die auf der gesellschaftlichen Makro- und der institutionellen Mesoebene ebenfalls als systemische Risiken gelesen werden können. Auf der Mikroebene konkreter Praktiken eröffnet KI für Medienanbieter indes auch Potenziale, die nun exemplarisch referiert werden: So lassen sich mit textbasierter KI-Software journalistische Arbeitsroutinen effektiver gestalten (vgl. de Ruiter 2019: 1322f.). *Chatbots* können zum Beispiel verwendet werden, um automatisiert Texte zu erstellen, zu strukturieren, zu überprüfen, zu verbessern oder zusammenzufassen; um Kurz- und Funktionstexte wie Überschriften, Vorspänne, SEO-Texte und so weiter zu kreieren; um Layouts, Visualisierungen, Grafiken oder fotorealistische Abbildungen zu erzeugen (vgl. Prinzing 2023: 521–525). Weil sich mithilfe von KI auch riesige Datensätze oder digitale Archive auswerten lassen, können selbstlernende Systeme als Recherche-, Analyse- und Monitoring-Werkzeug Verwendung finden. Darunter fällt auch die Möglichkeit, Misinformation und Desinformation zu detektieren, was gerade für den investigativen Journalismus ertragreich sein kann (vgl. Kleemann 2023: 7).

Außerdem lassen sich mit KI mediale Beiträge kanal- und zielgruppengerecht aufbereiten und ausspielen, etwa entlang der Modalitäten von Sozialen Medien oder Suchmaschinen. Auch die hiermit beschriebene Personalisierung basiert auf der Erfassung, Aus- und Verwertung von Daten durch KI. Idealiter münden KI-Verfahren zur algorithmischen Automatisierung von Zuschnitt, Zustellung und Verteilung medialer Produkte in der Verbesserung der Publikumsbindung und hiernach in einem gesteigerten Medienvertrauen. Neben KI, deren Ein-/Ausgaben Texte darstellen, können im Journalismus auch KI-Techniken genutzt werden, die (audio-)visuelle Produkte generieren oder modifizieren. So lassen sich mit GenKI defizitäre Originalaufnahmen interpolieren oder Transkriptions- und Übersetzungs-

leistungen (Fremdsprachen, Dialekte, inklusive Sprache, etc.) vollziehen. Überdies lassen sich auf artifiziellem Wege Personas erstellen, die (existierenden) Menschen in Aussehen, Mimik, Gestik, Stimmen und so weiter gleichen (vgl. Pawelec/Bieß 2021). Mit solchen *deep fakes* oder *voice clones* können mediale Darstellungsformen und Formate umgesetzt werden, die bislang auf menschliche Moderator:innen oder Kommentator:innen angewiesen waren.

Zusammengefasst lassen sich mit KI primär Ressourcen-, Effizienz- und Produktivitätsgewinne erzielen, etwa durch Zeit-, Personal- und Kosteneinsparungen. Das kann die wirtschaftliche Rentabilität erhöhen, aber gleichzeitig Arbeitsplätze gefährden, Berufsrollen transformieren oder diese obsolet machen (vgl. Albrecht 2023; WEF 2025). Je nach Einsatzkontext, Verwendungsform oder betroffenen Akteur:innen können sich KI-bedingte Einsparungen folglich auch als negative Charakteristika erweisen. So können mit GenKI Desinformationen kostenarm hergestellt, Personen geschädigt, Debatten verfälscht, Diskurse polarisiert oder Meinungsbildungsprozesse gestört werden (vgl. Grimm 2020; Diakopoulous/Johnson 2020: 2073; Krämer 2021: 25, 35–37). Das kann demokratische Pfeiler destabilisieren, Gesellschaftssysteme überfordern, jedenfalls aus demokratietheoretischer Sicht in kaum wünschenswerter Weise verändern (vgl. Pawelec/Bieß 2021: 44–48). Dies besonders dann, wenn KI dafür eingesetzt wird, Leistungen von Journalist:innen zu substituieren und somit die Funktionen der Medien auszuhebeln, was wiederum die Genese von Öffentlichkeit stören sowie dazugehörige Variablen wie politische Stabilität oder sozialer Zusammenhalt sabotieren könnte (vgl. Tschopp et al. 2022: 330–340).

Übergeordnet dürfte sich der Druck auf den Journalismus durch KI also weiter erhöhen. Zum einen, weil es dessen Aufgabe ist, den öffentlichen Raum „reinzuhalten" von Falschem, Unwahrem und Irrelevantem und Medienschaffende zunehmend damit beschäftigt sein könnten, Desinformation zu entlarven. Zum anderen könnten Medienorganisationen durch die Konkurrenz aus den Sozialen Medien und von KI-Anbietern vermehrt in ökonomische Bedrängnis geraten, was sie aus Effizienz- und Kostengründen dazu zwingen könnte, selbst auf KI-Systeme zurückzugreifen. Dies dürfte wiederum zulasten eines menschengemachten Journalismus gehen und das Vertrauen von Rezipient:innen in Medien negativ beeinflussen, sofern deren Qualität nachlässt.

Den Journalismus betreffende Organisationen, gerade solche Zusammenschlüsse, die um den Erhalt von Gütestandards bemüht sind, raten deshalb dazu, institutionelle Rahmenbedingungen für den verantwortungsvol-

len KI-Einsatz zu schaffen. Der *Deutsche Journalistenverband* (DJV 2025) etwa stellt die folgenden Forderungen auf:

1. Journalist:innen sollen weiterhin die Oberhand in der Recherche und Produktion von Inhalten haben, der Automatisierungsgrad sollte klug und maßvoll eingesetzt werden;
2. Medienunternehmen sollen sich bei dem Einsatz von KI auf ethischen Leitlinien verständigen;
3. Inhalte mit KI-Generierung sind deutlich und verständlich zu kennzeichnen;
4. Medienunternehmen sollten sich an dem Aufbau von werteorientierten Datenbanken und zertifizierten KI-Tools beteiligen;
5. Medienunternehmen sollen für ihre Mitarbeitenden Aus- und Weiterbildungsangebote für KI-Kompetenzen schaffen;
6. Der Gesetzgeber soll gerechte Vergütungsprinzipien für Journalist:innen durchsetzen.

Wie beschrieben stellt Vertrauen ein zentrales Gut in der Relation von Medien und Publikum dar. Die Vorschläge des *DJV* zahlen auf diesen Wert ein, weil idealtypisch argumentiert werden kann, dass sich die Einhaltung von ethischen Normen positiv auf die Aufgaben des Journalismus und dessen Wahrnehmung auswirken dürften: Zum einen können ethische Vorgaben für handelnde Akteur:innen als Orientierungsmarker dienen, KI verantwortungsvoll einzusetzen und technologische Veränderungen selbst aktiv mitzugestalten. Zum anderen müsste sich die Vertrauenswürdigkeit von Medienvertreter:innen erhöhen, wenn sich deren Rezipient:innen aus guten Gründen darauf verlassen können, dass Medienschaffende eine KI-Nutzung nur dann in Erwägung ziehen, wenn sie Risiken für Stakeholder reflektiert haben und ihre KI-Verwendung auch offenlegen. Berufsethisch lässt sich eine solche KI-Kennzeichnung mit den für den öffentlichen Raum zentralen Werten Wahrheit und Transparenz begründen: Journalist:innen zeigen ihrem Publikum dann, dass sie auf KI ohne Täuschungsabsicht setzen und weisen so ihre Wahrhaftigkeit nach. Das kann deshalb angebracht sein, weil es sich bei KI im Grunde um eine Nachahmung des Menschlichen handelt und künstlich erzeugte Produkte somit die Tendenz zur Täuschung graduell eingeschrieben ist. Der bewusste Verzicht auf eine Markierung von KI *kann* hingegen als Manipulation betrachtet werden (vgl. Schicha 2021: 175).

Demgegenüber kann eine Abwägung über die Zweckmäßigkeit und Legitimität von KI in Medien auf Seiten der Adressat:innen als Nachweis da-

für gewertet werden, dass Medienschaffende der Verantwortung als öffentliche Sprecher:innen nachkommen und ihrem gesellschaftlichen Auftrag auch nachkommen *wollen*. Daran zeigt sich, dass Absichten und Pflichtbewusstsein handelnder Akteur:innen sowie deren Kapazität zum sittlichen Umgang mit KI in der Medienproduktion als entscheidend für eine sozialverträgliche KI-Implementierung betrachtet werden können. Wie diese konkret ausgestaltet sein muss, hängt, wie allgemein für Technikimplementierung geltend, von den jeweiligen Einsatzkontexten und Nutzungsweisen ab, welche damit die entscheidenden Indikatoren darstellen, wenn eine Einschätzung darüber getroffen werden soll, wo, ob und wie KI aus Sicht der Ethik eingesetzt werden soll (vgl. Kuhnert/Grimm 2020: 253).

Um eine solche Abwägung vornehmen zu können, müssen die Beteiligten Funktionsweise und Folgen von KI verstehen beziehungsweise reflektieren und KI-Anwendungen selbst souverän anwenden können. KI-Mündigkeit kann dann mitunter auch bedeuten, den KI-Einsatz aus moralischen oder anderen legitimen Gründen abzulehnen. Schließlich gehört es zur Kernaufgabe von Journalist:innen, kritisch, sorgsam und objektiv zu sein – relevanten Dritten, aber auch der eigenen Rolle, dazugehörigen Pflichten und (technologischen) Phänomenen gegenüber. Am Ende begründet sich darin der Vertrauensvorschuss, den eine Bevölkerung ihren Medien entgegenbringt: Medien fungieren als Korrektive der Gesellschaft; ihr Dienst entfaltet sich an der und zum Wohle der Gesellschaft. Entsprechende Leistungen weisen sie durch eine redliche Arbeitsweise und eine durchdachte Technikverwendung nach. Der *DJV* bezieht sich in seinem Vorschlag zum Aufbau von KI-Kompetenz ebenfalls darauf. Die Frage, wie Medienstudierende zu einem souveränen KI-Einsatz im Sinne des Gemeinwohls befähigt werden könnten, steht im Zentrum von IKID.

3. Ansätze zur Vermittlung von interdisziplinärer KI-Kompetenz an Hochschulen

3.1 Das Lehrforschungsprojekt IKID – Vorgehen, Ziele und ethischer Schwerpunkt

Das Lehrforschungsprojekt IKID hatte zum Ziel, Studierende auf eine Medien- und Arbeitswelt vorzubereiten, die von KI geprägt sein wird. Dazu entstanden interdisziplinäre Lehrangebote, deren Ziel es war, KI-Kompetenz zu vermitteln. Die Kapazität zur kontextsensiblen Perspektivenerweite-

rung, welche an die Konzepte „Future Skills“ oder „Data Literacy“ (vgl. Ridsdale et al. 2015; Schüller 2019: 298ff.; Bandtel/Gläser 2021: 52, 60) anschließt, kann als konzeptionelle Erweiterung von Medien- und Digitalkompetenz gelesen werden. Darunter fallen Fähigkeiten, wie jene zur kritischen Reflexion, zur verständigungsorientierten Kommunikation, zur übergreifenden Folgenanalyse und zur konsensorientierten Aushandlung von Stakeholdern. Übergeordnet sollen jene Fähigkeiten, Eigenschaften und Identitätsdispositionen gestärkt werden, die Menschen von Maschinen unterscheiden – auch, um so typischen KI-Gefahren präventiv zu begegnen. Entlang des humanistischen Bildungsideals, das im Hinblick auf KI-Lehre auch der Ethikrat (vgl. 2023: 163–167) starkmacht, lassen diese als Befähigungen zum freien, vernünftigen, selbstbestimmten und kreativen Handeln, Urteilen und Entscheiden beschreiben. Da KI die Wissensarbeit zu substituieren droht, geht es überdies darum, bei Lernenden bestehende Kompetenzen zu erhalten und einen Fähigkeiten-Abbau (*deskilling*) zu verhindern (vgl. Reinmann 2023: 5–8).

Bezogen auf den Bildungskontext muss dabei speziell die Kapazität zur interdisziplinären Problem-bearbeitung erst noch aufgebaut werden, bei Studierenden wie Lehrenden (vgl. Albrecht 2023: 78; Ethikrat 2023: 167). Die IKID-Seminare wurden deshalb von Forschenden aus Informatik, Wirtschaft, Ethik und Recht zusammen entwickelt.[4] Kern des Moduls, dem eine Propädeutik vorausging, stellten die „Integrierten KI-Projekte” dar. In diesen wurden Lernende heterogener Studienherkünfte mit dem realitätsnahen Fall eines Medienunternehmens konfrontiert. Konkret wurde die *case study* eines Fernsehanbieters vorgegeben, der aus wirtschaftlichen Gründen KI-Klone einführen möchte, um Moderator:innen zu ersetzen und neue Formate zu entwickeln. Die studentischen Gruppen nahmen die Rolle eines interdisziplinären Consulting-Teams ein, das die Geschäftsführung beraten soll. Angestrebt wurde eine integrierte Lösung, die eine KI-Implementierung verfolgt, die inhaltlich innovativ, technisch umsetzbar, dem wirtschaftlichen Unternehmensziel zuträglich ist und rechtliche Vorgaben wie ethische Anforderungen bedenkt. Weil Kompetenzen in der Praxis erworben werden, sollten sich die Lernenden im geschützten Raum selbst

4 Im IKID-Projekt wurde in einer Serie von Whitepapers das didaktische Konzept skizziert sowie Lehrinhalte, Lehrmethoden und Kompetenzziele aller beteiligten Disziplinen dargelegt und gezeigt, wie durch den Einsatz einer „Sandbox“ und von KI-Demonstratoren eine anwendungsbezogene Lehre ermöglicht wird, den Link dazu ist im Literaturverzeichnis zu finden.

an KI-Software ausprobieren, Erfahrungen sammeln und Fehler machen dürfen (vgl. Schüller 2019: 303).

3.2 Ethische Anliegen

Aus Sicht der Lehrenden der Ethik ging es darum, den Studierenden einen ethisch reflektierten Umgang mit KI zu vermitteln, der die konzeptionellen Vorgaben der Technikfolgenabschätzung befolgt und übergeordnet „eine frühzeitige Auseinandersetzung mit den gesellschaftlichen Konsequenzen, aber auch Rahmenbedingungen der [KI-initiierten] Entgrenzungsdynamik […]" (Kehl 2021: 157) einsteuert. Dafür wurde ein verkürzter *ethics-by-design*-Prozess (vgl. Grimm 2025b) nachgestellt – mit den typischen Schritten (vgl. Manders-Huits 2011: 275)

- einer philosophisch-konzeptionellen Problemdiagnose;
- einer (teils auch empirisch ermittelten) Stakeholder-Auswertung;
- einer technischen Analyse von KI-Systemen;
- der Sensibilisierung für und Abschätzung von (möglichen) KI-Folgen für jeweilige Stakeholder und die Gesellschaft;
- der Ableitung ethischer Prinzipien und deren Berücksichtigung im Gestaltungsprozess von Technologien wie KI, von Konzepten wie Business-Plänen oder in Bezug auf KI-basierte Darstellungsformen.

In der Propädeutik wurden zunächst Grundlagen der Digital- und KI-Ethik vermittelt. Das umfasste passende Methoden, etwa Reflexions- und Analysewerkzeuge zur Stakeholder-Ermittlung und Folgenabschätzung. Außerdem wurden bestehende ethische Kodizes ausgewertet, deren Maximen auch in der *case study* berücksichtigt werden sollten.[5] In den KI-Seminaren kamen Diskussionsformate wie Rollenspiele zur Anwendung, in denen sich Teilnehmenden die Fähigkeit zur Perspektivübernahme aneignen, über ethische Fragen zu KI austauschen, soziale Folgen von KI besprechen, Haltungen, Bedarfe und Einstellungen artikulieren und ihre Erkenntnisse auf die Fallstudie übertragen sollten. Teil der Prüfungsleistung war es, eine

5 Konkret wurde das *Aktantenmodell*, die *Wertematrix*, das *Futures Wheels* und weitere Methoden gelehrt. Bezüglich ethischer Standards waren die *TARES*-Prinzipien (vgl. Baker/Martinson 2001), der Pressekodex (vgl. Deutscher Presserat 2025) und die *Pariser Charta* zu KI und Journalismus (vgl. Reporter ohne Grenzen 2023) mitsamt bedeutsamer Werte und Prinzipien des Journalismus (Vertrauen, Wahrheit, Unabhängigkeit, Neutralität, Glaubwürdigkeit, Objektivität, etc.) Gegenstand der Seminare.

begründete Werte-Topografie für das Medienunternehmen herzuleiten. Daraus wiederum sollten die Studierenden ethische Leitlinien für ihren formatspezifischen GenKI-Einsatz entwickeln.

3.3 Ausgewählte Implikationen von KI für die Bildung

Lehrende stehen vor der Aufgabe, sich und ihre Studierenden auf KI vorzubereiten und damit auf einen Gegenstand, der sich simultan zur individuellen Aneignung und didaktischen Vermittlung verändert (vgl. Ethikrat 2023: 163–182 und Schlegel et. al 2024). Von Lehrenden wird entsprechend erwartet, anpassungsfähig und experimentierfreudig zu sein (vgl. Pancratz et al. 2023: 88). KI zwingt sie daher dazu, Lehr-/Lernansätze und Prüfungskulturen zu modifizieren (vgl. Albrecht 2023: 78). Aus gutem Grund denkt man nur an die vielfältigen Einsatzmöglichkeiten von GenKI im Hochschulbereich: als Recherche-Werkzeug; zum Organisieren von Wissen; zum Strukturieren, Verfassen, Überprüfen und so weiter von Texten; zum Erstellen von Präsentationen oder Visualisierungen sowie für weitere Zwecke. (vgl. Reinmann 2023).

Grundlegend zahlen die IKID-Seminare auf diese Anforderungen ein, zumal sie im Rahmen der Lehrforschung evaluiert und angepasst wurden. Zudem adressiert das überfachliche Vorgehen eine strukturell bedingte Herausforderung von Bildungseinrichtungen: die domänenspezifische Fokussierung, die dazu führt, dass Studierende in nur einer Disziplin ausgebildet werden und neue Phänomene daher primär aus Sicht ihres Fachs einzuschätzen lernen. Weil KI jedoch als gesellschaftliches Querschnittsphänomen einzuordnen ist, bedarf es einer integrierten Lehre (vgl. Schlegel et al. 2024). Um tragfähige Lösungen und (neue) Methoden für künftige Herausforderungen entwickeln zu können, müssen vormals getrennte Fachgrenzen graduell überwunden werden, zumindest braucht es die Bereitschaft dazu (vgl. Grimm 2025a; Braßler 2020). Das erfordert experimentelle Lehrkonzepte, Lehr- und Lernumgebungen, in deren Entwicklung die Anspruchsgruppen – entlang der Maximen von *ethics by design* – miteinbezogen werden müssen. Lehrinnovation entwickelt sich dann als Ko-Konstruktion zwischen Forschenden, Lehrenden und Studierenden, die auch als Nutzende und damit als Expert:innen in eigener Sache verstanden werden müssen. So gestalten sie die KI-Integration im Bildungsbereich in indirekter Weise mit. Da Kompetenzen in der Praxis erworben werden, eignen sich Lehrende und Lernende die notwendigen Fähigkeiten im Umgang

mit KI sukzessive an und bilden idealerweise eine Haltung in Technikfragen heraus. Darüber hinaus definiert sich Kompetenz als die Kapazität, in komplexen Situationen, die sich stetig verändern, kontingente Verläufe annehmen und offene Ausgänge zeitigen, kreative Lösungen zu entwickeln – eine bislang genuin menschliche Gabe (vgl. Lenbet 2024: 223–225). Wie gezeigt, kann KI als ein solches emergentes Phänomen verstanden werden.

Dabei adressiert das kompetenzorientierte Vorgehen von IKID erneut den Wert des Vertrauens: Studierende, die zur Nutzung von KI und zur Abschätzung von KI-Folgen befähigt wurden, entwickeln zum einen Vertrauen in ihren Umgang mit KI. Sie sind also fähig, einzuschätzen, in welchen *use cases* der KI-Einsatz gerechtfertigt sein kann und wann derlei Systemen misstrauisch zu begegnen ist – zum Beispiel, wenn Werte des sozialen Zusammenlebens bedroht sind. KI-Mündigkeit kann zum anderen auch bedeuten, zu erkennen, welche menschlichen Dispositionen gegenüber einer Maschine zu schützen sind. Um die KI-Transformation unter Beteiligung der Bevölkerung gelingend zu gestalten, braucht es einerseits technische Kompetenz und andererseits Akzeptanz gegenüber neuen Technologien. Wie Studien zeigen, korreliert das eine mit dem anderen: So weisen Arnold et al. (2020: 24, 34–41) nach, dass Menschen KI-Systemen weniger misstrauisch begegnen oder sich diesen seltener verweigern, wenn sie deren Auswirkungen für beherrschbar halten. Einschätzungen zur Kontrollierbarkeit von KI sind dann höher, wenn Befragte solche Tools selbst nutzen. Übergeordnet sichert KI-Kompetenz daher die menschliche Selbstbestimmung und Autonomie; (Selbst-)Vertrauen in menschliche Fähigkeiten also.

4. Ethisch reflektierter Umgang mit künstlichen Stimmen in der Medienpraxis

4.1 Begründung, Beschreibung und Ziele des Forschungsprojekts GEISST

Das Forschungsprojekt GEISST beschäftigte sich mit dem Einsatz von synthetischen und geklonten Stimmen im Journalismus. Ausgegangen wurde von einem *use case*, bei dem mit KI, abgestimmt auf verschiedene Ausspielungskanäle, variierende Versionen eines audiovisuellen Nachrichtenbeitrags generiert werden. Der ethische Projektschwerpunkt begründet sich zum einen darin, dass KI-erstellte Stimmen in einem grundlegenden Bereich menschlicher Kommunikation eingreifen: die gesprochene Spra-

che. Sprache ist nicht nur Träger von Information, sondern auch Ausdruck von Identität. GenKI kann Sprache und Stimmen technisch reproduzieren. Damit verschiebt sich die Produktion gesprochener Sprache von einem menschlich-körperlichen Ausdruck hin zu einem technisch vermittelten Prozess. KI-Stimmen können dadurch das Erleben von Sprache verändern. In journalistischen Kontexten berührt dies grundlegende Fragen von Glaubwürdigkeit, Transparenz und Verantwortung. Sowohl Sprache als auch Stimmen stellen Identitätsmerkmale und damit ein zentrales Instrument journalistischer Integrität dar.

Zum anderen begründet sich der ethische Fokus darin, dass Medien ein Kulturgut und aufgrund ihrer systemerhaltenden Funktion in Demokratien qua Verfassung geschützt sind. Um Qualitätsstandards zu garantieren und um Sorge dafür zu tragen, dass journalistische Gütekriterien nicht ökonomischen Zielen zum Opfer fallen, braucht es rechtliche Rahmensetzungen und ethische Normen. Letztere wurden in GEISST in Form von Leitlinien umgesetzt, welche aus einer Interviewstudie mit Medienschaffenden abgeleitet wurden (vgl. Grimm/Kuhnert/Schlegel 2025). Mit dieser sollte ermittelt werden, welchen Werte die Befragten beim Einsatz von GenKI zur Nachahmung von Stimmen im Journalismus eine exponierte Stellung zusprechen. Durch den Stakeholder- und Anwendungsbezug, der durch die Befragung von Medienpraktiker:innen sichergestellt wurde, sollte erreicht werden, dass die Handreichungen realitätsnah und entsprechend wirkungsvoll sind.[6]

4.2 Ablauf und Methode der Befragung

Im Rahmen der qualitativen, nicht-repräsentativen Interviewstudie wurden zwölf Personen im Alter von Mitte 20 bis Anfang 60 Jahren befragt. Die Redakteur-, Journalist- und Sprecher:innen sind in die Nachrichtenproduktion von privaten Medienunternehmen oder öffentlich-rechtlichen Einrichtungen in den Bereichen Rundfunk, Print oder TV eingebunden. Die teilstrukturierten Einzelinterviews wurden in Präsenz oder via Online-Vi-

6 Die Beteiligung von Stakeholdern, umgesetzt in Form von empirischen Methoden wie Befragungen, Interviews und so weiter, stellt eine zentrale Maxime von ethischen *by-design*-Ansätzen dar. Gleichzeitig ist zu betonen, dass empirische Befunde allein nie als Grundlage für normative Empfehlungen ausreichen, aber zur Überprüfung der „Realitätsadäquatheit" einer Medienethik dienen können (vgl. Rath 2013: 296).

deotelefonie durchgeführt. Für die Auswertung wurden die 50- bis 90-minütigen Gespräche transkribiert und anonymisiert. Ziel der Befragung war es, begründete Einschätzungen darüber zu erlangen, unter welchen Bedingungen Medienschaffende einen KI-Einsatz in der Nachrichtenproduktion akzeptieren (würden) (F1); welche Werte sie in diesem Verwendungs-Zusammenhang betonen (F2); zudem wurde Bezug genommen auf bestehende ethische Rahmenwerke und die dort hervorgehobenen Normative von Wahrheit und Wahrhaftigkeit, die in maßgeblichen Kodizes prominent platziert sind.[7] Im Zentrum der Befragung standen diese Forschungsfragen:

F1: Wie steht es um die Akzeptanz von Generativer KI im Medienbereich und dem Nachrichtenjournalismus, speziell bei Journalist:innen und Redakteur:innen?
F2: Was verstehen die Befragten unter „gutem Journalismus"? Welche Standards sollen in Zukunft beim Einsatz von Generativer KI gelten?
F3: Welche Bedeutung hat der Wert der Wahrheit für Journalist:innen und Redakteur:innen und wie definieren diese Wahrheit in Bezug auf den Einsatz von Generativer KI?

Das Vorgehen wurde an die Vorgaben des *value-sensitive-design*-Ansatzes angelehnt, der sich anbot, weil er den Fokus auf die Beteiligung von Stakeholdern bei der Entwicklung ethischen Standards richtet. Hier wird ein Wert recht weit gefasst: „What is important to people in their lives, with a focus on ethics and morality" (Friedman/Hendry 2019: 24). Die Befragten sollten in ein relativ freies Erzählen gebracht werden, während Interviewende gleichzeitig dafür sorgen mussten, die Gespräche auf das Thema KI im Journalismus zu konzentrieren (vgl. Müller/Grimm 2016). Zunächst wurden den Befragten deshalb verschiedene KI-Anwendungsfälle präsentiert, zu denen sie assoziative Eindrücke schildern sollten. Durch diesen offenen Initialimpuls sollte ein lockerer Einstieg ins Thema erreicht werden. Die Spanne ausgewählter *use cases* war dabei bewusst breit gefächert und beschränkte sich nicht nur auf einen KI-unterstützten Journalismus. Im Anschluss wurde die Befragung auf die Möglichkeiten der KI-Nutzung in der täglichen Arbeit der Befragten und damit auf deren persönliche Erfahrungswelt gelenkt. Auf diese Weise konnten Praxisnähe und Kontextbezug

7 So heißt es beispielsweise in Ziffer 1 des „Pressekodex" des *Deutschen Presserates*: „Die Achtung vor der Wahrheit, die Wahrung der Menschenwürde und die wahrhaftige Unterrichtung der Öffentlichkeit sind oberste Gebote der Presse. Jede in der Presse tätige Person wahrt auf dieser Grundlage das Ansehen und die Glaubwürdigkeit der Medien."

gewährleistet und gleichsam explizit wie implizit vorliegende Wertvorstellungen identifiziert werden.

4.3 Ausgewählte Ergebnisse

Ein zentraler Befund ist, dass die im Journalismus bestehenden normativen Vorgaben weiterhin als axiomatischer Bezugspunkt fungieren: Alle Befragten verweisen auf bekannte Ethik-Kodizes, vor allem den „Pressekodex". Bekannte Professionsstandards betrachten sie als Maßstab für guten Journalismus und bestätigen somit deren Gültigkeit und Aktualität. Zudem verdeutlichen die Ergebnisse, dass Medienschaffende journalistische Qualität an die Einhaltung von grundlegenden Prinzipien koppeln, speziell an redaktionelle Sorgfaltspflichten und institutionelle Verantwortlichkeiten. Die tragende Rolle bewährter Praktiken wird dabei immer wieder hervorgehoben und speziell mit der Gewährleistung von Perspektivenvielfalt, dem journalistischen Neutralitäts- und Transparenzgebot, redaktionellen Prüfungs- und Kontrollschleifen und inhaltlichen Einordnungen verbunden, beispielsweise der Trennung von Fakten und Meinungen – exemplarisch hierzu die Äußerung einer Journalistin (TN 9):

> „[…] es gibt dieses Sich-nicht-mit-einer-Sache-gemein-Machen: also auch mal die Perspektiven zu wechseln. Mal ganz bewusst auch eine Distanz aufzubauen, weil man natürlich schnell […] auch subjektiv wird […]. Ich glaube, guter Journalismus entsteht im Team, durch Diskussionen, in Redaktionskonferenzen, Vier-Augen-Prinzip […]."

Der Einsatz von GenKI im Journalismus wird keineswegs pauschal abgelehnt, allerdings wird er an zentrale Forderungen geknüpft: Für alle Befragten steht fest, dass technologische Mittel die Redaktionen nicht von deren Verantwortungsübernahme für die Publikationen entbinden und Maschinen somit den Menschen nie ersetzen sollten. Vielmehr sollen technische Mittel, auch KI, unter menschlicher (Letzt-)Kontrolle verbleiben. Dies ist vor allem jenen Personen wichtig, die direkt betroffen wären, wenn ihre Stimme für einen KI-erstellten Beitrag genutzt wird. Die Aussagen aller Befragten verweisen implizit auf eine professionsethische Grundeinstellung gegenüber Technologien und einer auch politisch interpretierbaren Erwartungshaltung, die so lauten könnte: Technologische Innovationen sollen Arbeitsweisen erleichtern, dürfen gleichwohl nicht dazu führen, dass ethische Grundwerte, grundlegende redaktionelle Aufgaben oder gesellschaft-

liche Leistungen des Journalismus relativiert werden. Neue Technologien werden damit als unterstützendes Werkzeug betrachtet. TN 11 skizziert dies sogar in einem historischen Zusammenhang:

> „Das war nie so, dass wir ohne Technik ausgekommen sind […]. Und unter dem Aspekt […] sind diese Algorithmen [und] KI-Tools ja im Grunde […] nichts anderes als jedes andere Tool, das wir in den vergangenen Jahrzehnten und Jahrhunderten genutzt haben – von der Schreibmaschine über die Rechtschreibprüfung bis zum Ganzseitenumbruch bis zur automatisierten Sendeabwicklung."

Die Interviews offenbaren ein subtil vorliegendes Misstrauen gegenüber KI-Stimmen in den Medien, das gleichsam ambivalent wahrgenommen wird: Einerseits werden technische Innovationen akzeptiert und die KI-Integration in Medienorganisationen als quasi unvermeidlich hingenommen. Andererseits äußern einige Befragte Bedenken hinsichtlich der Authentizität synthetischer Stimmen und deren Einfluss auf Wahrnehmungen von Wahrhaftig-, Vertrauens- und Glaubwürdigkeit, speziell von Sprecher:innen, so etwa TN 3:

> „[…] Ich bin zurückhaltend mit künstlichen Stimmen, weil ich es nach wie vor wichtig finde für die Authentizität des Themas [oder] Beitrags, dass die Menschen selbst […] zu Wort kommen, weil es vermittelt auch […] Glaubwürdigkeit, wenn der Mensch selber in seiner Originalstimme sprechen kann."

Ein Teil der Befragten äußert explizit – und ohne dazu befragt worden zu sein – den Bedarf nach mehr Qualifizierung: Dies betrifft sowohl die technischen Aspekte von KI als auch den professionellen Umgang mit und Fähigkeiten zur Detektion von Desinformation. In einem ähnlichen Zusammenhang stellen alle Befragten Transparenz als essenziellen Wert heraus, dies im Sinne von Sichtbarkeit und Kennzeichnung. Es soll für die Rezipient:innen eindeutig erkennbar sein, wenn eine Stimme KI-erzeugt wurde. Eine fehlende Kennzeichnung wird als potenziell täuschend eingestuft, weil es das Vertrauen in ein Medium oder den Journalismus insgesamt beeinträchtigen könne, so unter anderem TN 5:

> „Wir verlieren an Glaubwürdigkeit. Also wenn etwas sehr offensichtlich nicht so produziert wurde und etwas sehr offensichtlich nicht wirklich so war, wie es dann publiziert wird, dann glaubt man ja auch alles andere nicht mehr."

Der möglicherweise drohende Jobverlust durch KI und angesichts ohnehin zunehmend prekärer Arbeitsverhältnisse von Journalist:innen (vgl. Hanitzsch/Rick 2021) wird von mehreren Personen adressiert, so auch von TN 8:

> „Als Berufstätiger oder Praktiker sehe ich es halt kritisch, weil ich mir denke: Da sind ja wieder ein paar Jobs von Nachrichtensprecherinnen und Nachrichtensprecher, die ersetzt werden."

Alle Interviews eint, dass der KI-Einsatz mit ambivalenten Einschätzungen einhergeht. Keine:r der Befragten lehnt KI kategorisch ab, dies trotz eines ausgeprägten Bewusstseins über mögliche Risiken, welche die Medienschaffenden auch persönlich treffen könnten. Einerseits stellen die Interviewten die Potenziale von KI heraus. Andererseits treibt sie die Sorge um, dass durch KI menschliche Aspekte ihrer Arbeit verloren gehen könnten – die am häufigsten geäußerten Bedenken. Es zeigt sich an dieser Stelle ein latenter Technikdeterminismus. Denn die Befragten nehmen die bislang weithin ungehemmt fortschreitende KI-Integration gewissermaßen als gesetzt hin und akzeptieren diese und darüber hinaus auch das Narrativ, dass diese Entwicklung unaufhaltbare wäre. Gleichzeitig legen sie Zuversicht und Vertrauen eben nicht in KI, sondern in den Menschen: in die Hoffnung, dass Menschen lernen werden, KI-bedingten Herausforderungen auf ebenso menschlicher, sozialer und gerechterweise zu begegnen.

4.4 Theoretische Grundlagen der ethischen Leitlinien

Die in GEISST entwickelten Leitlinien orientieren sich einerseits an bestehenden journalistischen Standards, andererseits an einer aristotelisch geprägten Angewandten Ethik, welche das teleologische Ziel des „guten Lebens" verfolgt. Nach Spiekermann (2019) folgt die Digitale Ethik ebenfalls dieser eudämonistischen Tradition. Für Weber (2017: 36), der journalistische Aufgaben der politischen Sphäre zurechnet, bedeutet moralisches Handeln, für die absehbaren Folgen des eigenen Tuns einzustehen. Verantwortungslosigkeit liegt vor, wenn künftig mögliche oder bereits realisierte Auswirkungen ignoriert werden oder Journalist:innen unsachlich berichten (ebd.: 63). Hier offenbart sich erneut der Wert der Wahrheit, dessen Einhaltung im Journalismus Weber ebenfalls als wesentlich erachtet. Weber folgend, sollen Verantwortungsethiker:innen in ihren Folgenbetrachtun-

gen zudem menschliche Schwächen stets weitsichtig mitbedenken. Entsprechend dürfen sie Konsequenzen ihres Handelns nicht auf andere abwälzen.

Auf aktuelle Medienkontexte bezogen, könnte das beispielsweise bedeuten, dass einem Entscheidenden, der/die aus Kostengründen eine/r Sprecher:in durch einen KI-Avatar ersetzen möchte, sich der Konsequenzen bewusst sein muss. Wirtschaftliche Argumente allein reichen etwa nicht aus, um den Jobverlust des/der Kolleg:in zu rechtfertigen. Entsprechend sollte sie legitime Gründe für ihr Vorgehen präsentieren können. Zum Beispiel könnte der KI-Einsatz moralisch vertretbar sein, wenn er eine Medienproduktion überhaupt erst ermöglichen würde.

Daran zeigt sich: Technikethik bewegt sich stets im Spannungsfeld individueller und institutioneller Verantwortung (vgl. MacCormac 1993). Beide Seiten, Medienschaffende wie Medienorganisationen, müssen sich gemeinsam um einen verantwortungsvollen KI-Einsatz bemühen. Damit muss die KI-Verwendung in den Medien das Produkt eines (redaktionellen) Abwägungsprozesses sein, in den alle Bezugsgruppen eingebunden werden. Aufgrund der gesellschaftlichen Verantwortung von journalistischen Medien müssen sich reflektierte Praktiken damit nicht zwangsläufig an unmittelbaren ökonomischen Vorteilen messen lassen, wenngleich das eine das andere nicht ausschließt: Der größte Nutzen von moralischem Handeln liegt laut Weber vielmehr in langfristigen Erwägungen. So kann moralisches Handeln das Mediensystem stabilisieren und das Vertrauen in Medienvertreter:innen erhöhen.

In GEISST wird diese teleologische Perspektive aufgegriffen. Deren Maximen werden gleichsam ausdrücklich auf das Ziel eines guten Journalismus bezogen. Verantwortung zeigt sich in der sozial-verträglichen Kalkulation der Auswirkungen des KI-Einsatzes unter Berücksichtigung der funktionalen Bedeutung des Journalismus für eine stabile Demokratie.

Nun steht die empirische Folgenanalyse von KI in Medien aufgrund des frühen Entwicklungsstatus von selbstlernenden Systemen noch am Anfang; insbesondere Langzeitstudien fehlen (Székely et al. 2025). Gleichwohl zeigen erste Forschungen negative Effekte auf das Verstehen und Erinnern von Nachrichten, wenn diese durch synthetische Stimmen vermittelt wurden (vgl. Gong 2023). Bekannt ist allgemein, dass (audio-)visuelle Inhalte ein enormes persuasives Potenzial aufweisen: Zum einen, weil sie eine Augenzeugenschaft und damit eine dokumentarisch-objektive Wahrnehmung suggerieren (vgl. Schicha 2021: 175f.); zum anderen, weil menschliche Stimmen als authentisch wahrgenommen werden können (vgl. de Ruiter 2019: 1322; Singh 2020: 69f.). KI-generierte Medien stehen

damit im Verdacht, den epistemologischen Wert von Fotos, Videos und Audios zu bedrohen. Weil synthetische Inhalte schon jetzt kaum noch vom Original unterscheidbar sind (vgl. Frank et al. 2024), entsteht durch KI ein enormes Desinformations- und Polarisierungspotenzial. Da ein fehlendes Bewusstsein allgemein gesprochen die Anfälligkeitswahrscheinlichkeit für Polarisierung erhöht (vgl. Hagen et al. 2018: 18), sind letztlich Aufklärung, Bildung und KI-relevante Kompetenzen die Schlüsselfaktoren für einen kritischen Umgang mit Medieninhalten.

Die GEISST-Leitlinien zahlen auf diese Qualifizierungsziele ein. Weil sie Medienpraktiker:innen nach Bedarfe und Anliegen fragen und diese um KI-spezifische Einschätzungen bitten, erfüllen sie auch verantwortungsethische Maximen. Durch ihren empirischen Gehalt tragen sie ferner dazu bei, die KI-spezifische Risikoabschätzung voranzutreiben.

4.5 Ethische Leitlinien in GEISST

Im Folgenden sollen die Leitlinien, die im GEISST-Projekt erarbeitet wurden, in komprimierter Form präsentiert werden.[8] Basis derselben sind zum einen die genannten theoretischen Überlegungen (4.4.). Zum anderen stellen sie die Ableitungen der Ergebnisse der Interviewstudie dar (4.3). Da sich Normen entlang des *value-sensitive-design*-Ansatzes auf Werte beziehen lassen müssen, wurde zunächst eine Werte-Topografie angefertigt (Abbildung 1), die aus den Befragungsergebnisse extrahiert wurde.

- *Ziel- und Kontextbindung:* Künstliche Stimmen sollen nur zu journalistischen Zwecken (Information, Aufklärung etc.) eingesetzt werden, um die Integrität der Medienorganisation zu bewahren. Der Einsatz soll im jeweiligen Kontext geprüft und begründet werden. Wirtschaftliche Vorteile allein rechtfertigen einen verantwortungsbewussten Einsatz nicht.
- *Kennzeichnungspflicht:* Jede KI-Stimme soll als solche transparent erkennbar sein. Die technische Herkunft der Stimme soll offengelegt und eindeutig gekennzeichnet werden. Rezipient:innen sollen dergestalt weiterhin selbstbestimmte Entscheidungen über ihre Mediennutzung treffen können.
- *Schutz vor Täuschung:* KI-Stimmen sollen nicht zur emotionalen oder politischen Beeinflussung und nie mit manipulativer Absicht eingesetzt werden. Der KI-Einsatz darf die Verpflichtung zur wahrheitsgemäßen

8 Vollständige Fassung siehe bei Grimm/Kuhnert/Schlegel 2025.

und unvoreingenommenen Informationsvermittlung nicht untergraben. Bei sensiblen Themen soll auf KI-Stimmen verzichtet werden.

- *Rechenschaftspflicht:* Die Verantwortung für eine KI-Nutzung soll bei den Redaktionen verbleiben. Entscheidungen zum KI-Einsatz müssen dokumentiert und nachträglich einsehbar sein. Medienhäuser sollen transparente Beschwerdemechanismen etablieren, damit das Publikum mögliche Täuschungserfahrungen melden kann.
- *Schutz menschlicher Sprecher:innen:* Das Nachbilden einer Stimme soll nur mit ausdrücklicher, informierter und jederzeit widerrufbarer Zustimmung der betroffenen Person erlaubt sein. Eine KI-Stimme darf nicht über den ursprünglich vereinbarten Kontext hinaus verwendet oder verändert werden. Stimmen sind durch Persönlichkeitsrechte geschützt, die von Redaktionen zu respektierten sind – auch dann, wenn die Stimme technisch neu erzeugt wurde, aber einer realen Person ähnelt.
- *Ethische Gestaltung affektiver Sprechmodulation:* Tonfall, Tempo und Klangfarbe von KI-Stimmen sollen dem unabhängigen Informationsauftrag entsprechen. Eine emotionale Aufladung, etwa um Vertrauen oder Aufmerksamkeit zu erzeugen, ist nicht zulässig. Im Nachrichtenkontext gilt eine neutrale und sachliche Tonalität als Standard.
- *Schulung und Weiterentwicklung:* Journalist:innen sollen die Wirkung synthetischer Stimmen kritisch reflektieren und verantwortungsvoll einsetzen. Medienorganisationen sollten dafür Qualifizierungsangebote schaffen und sich an Beforschung von KI-Stimmen beteiligen. Auf diese Weise befähigen sie Mitarbeitende zur verantwortungsbewussten Teilhabe an der KI-Entwicklung.

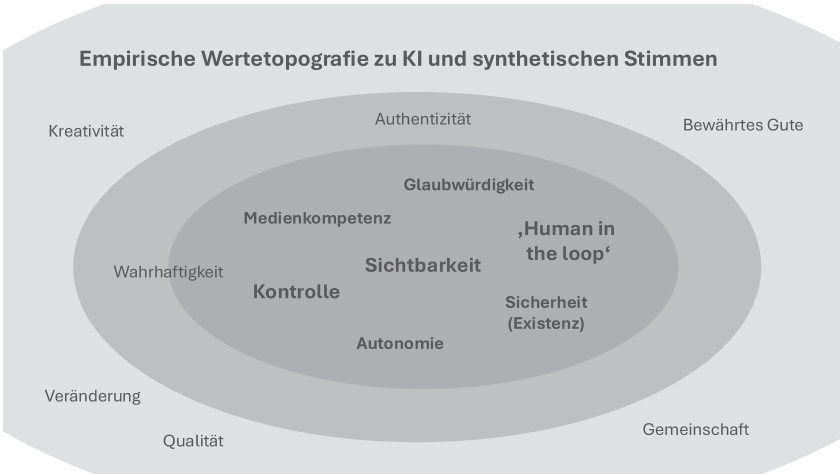

Abbildung 1: Empirische Wertetopografie zu KI und synthetischen Stimmen. Die Positionierung der dargestellten Werte reflektiert ihre relative Bedeutung in den Expert:innenaussagen, wobei eine geringere Distanz zum Zentrum mit einer höheren Nennungshäufigkeit einhergeht (eigene Darstellung).

5. Resümee

Die im Rahmen der Projekte IKID und GEISST gewonnenen Erkenntnisse verdeutlichen, dass der Einzug von KI in den Medienbereich sowohl Chancen als auch erhebliche Herausforderungen mit sich bringt. Insgesamt verdeutlichen beide Projekte, dass ein rein technischer oder ökonomischer Blick auf die KI-Verwendung in den Medien zu kurz greift. Eine verantwortungsvolle Einführung von KI erfordert daher eine enge Verzahnung von ethischer Reflexion, interdisziplinärer Bearbeitung und der Bewahrung journalistischer Kernwerte wie besonders Vertrauen, Wahrheit und Transparenz. Die Zukunft des Journalismus im Zeitalter der KI wird maßgeblich davon abhängen, ob es gelingt, die Potenziale der Technologie nutzbar zu machen, ohne die demokratisch relevanten Funktionen der Medien zu gefährden, die ein menschengemachter Journalismus etablierte und bislang absichert. Dies erfordert nicht nur die Einhaltung von ethischen Leitlinien, sondern auch die konsequente Ausbildung von Medienprofis, die in der Lage sind, die Technologie als Werkzeug unter menschlicher Kontrolle zu gestalten und zu steuern – und damit auch eine ethisch orientierte KI-Kompetenz.

Literatur

Albrecht, Steffen (2023): ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen (= TAB-Hintergrundpapier Nr. 26), Berlin.

Arnold, Norbert et al. (2020): „Wenn die KI unser Assistent bleiben kann, dann können wir viel draus ziehen" – Künstliche Intelligenz in Einstellungen und Nutzung bei unterschiedlichen Milieus in Deutschland, Konrad-Adenauer-Stiftung (online unter: https://www.kas.de/documents/252038/7995358/Künstliche+Intelligenz+in+Einstellungen+und+Nutzung+bei+unterschiedlichen+Milieus+in+Deutschland.pdf – letzter Zugriff: 15.9.2025).

Baker, Sherry / Martinson, David L. (2001): The TARES Test: Five Principles for Ethical Persuasion, in: Journal of Mass Media Ethics 16 (2&3/2001), S. 148–175.

Bandtel, Matthias / Gläser, Christine (2021): Potenziale digitaler Lehre, in: Johanna Ebeling / Henning Koch / Alexander Roth-Grigori (Hg.), Kompetenzerwerb im kritischen Umgang mit Daten, Berlin, S. 51–62.

Bentele, Günter (2021): Der Wahrheits- und Wahrhaftigkeitsanspruch in einer Ethik der Öffentlichen Kommunikation, in: Christian Schicha / Ingrid Stapf / Saskia Sell (Hg.), Medien und Wahrheit. Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", Baden-Baden, S. 59–77.

Braßler, Mirjam (2020): Praxishandbuch Interdisziplinäres Lehren und Lernen. 50 Methoden für die Hochschullehre, Weinheim.

De Ruiter, Adrienne (2021): The Distinct Wrong of Deepfakes, in: Philosophy & Technology 34, S. 1311–1332.

Deutscher Ethikrat (2023): Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz (= Stellungnahme Deutscher Ethikrat, 20. März 2023) (online unter: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf – letzter Zugriff: 15.6.2025).

Deutscher Journalistenverband (DJV 2025): Positionspapier bezüglich des Einsatzes Künstlicher Intelligenz im Journalismus (online unter: https://www.djv.de/medienpolitik/kuenstliche-intelligenz/ – letzter Zugriff 9.8.2025).

Deutscher Presserat (2025): Publizistische Grundsätze (Pressekodex), 19. März 2025 (online unter: https://www.presserat.de/pressekodex.html – letzter Zugriff: 2.7.2025).

Diakopoulos, Nicholas / Johnson, Deborah (2021): Anticipating and addressing the ethical implications of deepfakes in the context of elections, in: New Media & Society 23 (7/2021), S. 2072–2098.

Frank, Joel et al. (2024): A Representative Study on Human Detection of Artificially Generated Media Across Country, in: IEEE Symposium on Security and Privacy (SP), San Francisco, S. 55–73.

Friedman, Batya / Hendry, David G. (2019): Value Sensitive Design. Shaping Technology with Moral Imagination, Massachusetts.

Gong, Chen (2023): AI voices reduce cognitive activity? A psychophysiological study of the media effect of AI and human newscasts Chinese journalism, in: Frontiers in Psychology, 23. November 2023 (online unter: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1243078/full- letzter Zugriff: 9.8.2025).

Grimm, Petra (*2025a*): Ethik der Interdisziplinarität in der KI- und MIT-Forschung: eine Frage der Haltung?, in: Petra Grimm / Oliver Zöllner (Hg.): Ethik der Digitalisierung in Gesundheitswesen und Pflege. Analyse und ein Tool zur integrierten Forschung, Stuttgart, S. 109–125.

Grimm, Petra (*2025b*): Ethics by Design: Potenziale, Umsetzung und Grenzen, in: Dirk Lanzerath / Aurélie Halsband (Hg.), Ethics by Design. Grundlagen, Umsetzung und Grenzen ethischer Technikgestaltung. Ethik in den Biowissenschaften, in: Sachstandsberichte des DZRE, Bd. 28, Baden-Baden, S. 41–76.

Grimm, Petra (*2020*): Die Ent-wirklichung. Zum Vertrauen in Zeiten der Infodemie, in: Klaus Koziol (Hg.), Entwirklichung der Wirklichkeit. Von der Suche nach neuen Sicherheiten, München, S. 55–83.

Grimm, Petra / Kuhnert, Susanne / Schlegel, Marcel (2025): Ethische Leitlinien für den Einsatz von synthetischen und geklonten Stimmen im Nachrichtenjournalismus, zweite, überarbeitete Fassung, Stuttgart.

Habermas, Jürgen (2022): Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Berlin.

Hanitzsch, Thomas / Rick, Jana (2021): Prekarisierung im Journalismus, 1. Ergebnisbericht, März 2021 (online unter: https://www.ifkw.uni-muenchen.de›lernbereiche›hanitzsch›projekte›prekarisierung.pdf – letzter Zugriff: 29.3.2025).

Jakobs, Ilka et al. (2022): Medienvertrauen im internationalen Vergleich – Befunde aus Deutschland, Spanien, Schweden und den USA, in: UFITA – Archiv für Medienrecht und Medienwissenschaft 86 (2/2022), S. 374–401.

Jarren, Otfried (2019): Fundamentale Institutionalisierung: Social Media als neue globale Kommunikationsinfrastruktur. Der Beitrag der Kommunikationswissenschaft zur Analyse medialer Institutionalisierungsprozesse, in: Publizistik 64, S. 163–179.

Krämer, Sybille (2021): Der Verlust des Vertrauens. Medienphilosophische Perspektiven auf Wahrheit und Zeugenschaft in digitalen Zeiten, in: Christian Schicha / Ingrid Stapf / Saskia Sell (Hg.), Medien und Wahrheit. Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", Baden-Baden, S. 25–42.

Kleemann, Aldo (2023): Deepfakes – Wenn wir unseren Augen und Ohren nicht mehr trauen können. Medienmanipulationen im Konflikt. Herausforderungen und Bewältigungsstrategien, in: SWPAktuell 43, Berlin.

Kuhnert, Susanne / Grimm, Petra (2020): Die Zusammenarbeit von Industrie, Ethik und Wissenschaft im Forschungsverbund. Kommunikation – Integration – Innovation, in: Bruno Gransche / Arne Manzeschke (Hg.), Das geteilte Ganze. Horizonte Integrierter Forschung für künftige Mensch-Technik-Verhältnisse, Wiesbaden, S. 241–261.

Lenbet, Aylin (2024): Zur Aktualität des Kompetenzbegriffs und zur Bedeutung der Kompetenzentwicklung für das Coaching, in: Organisationsberatung – Supervision – Coaching 1, S. 221–232.

MacCormac, Earl R. (1993): Das Dilemma der Ingenieurethik, in: Hans Lenk / Günter Ropohl (Hg.), Technik und Ethik, Stuttgart, S. 222–244.

Manders-Huits, Noëmi (2011): What values in design? The challenge of incorporating moral values into design, in: Science and Engineering Ethics 17(2), S. 271–287.

Menzel, Christoph / Winkler, Christian (2018): Zur Diskussion der Effekte Künstlicher Intelligenz in der wirtschaftswissenschaftlichen Literatur (= Diskussionspapier Nr. 8. des Bundesministeriums für Wirtschaft und Energie), Oktober 2018 (online unter: https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Downloads/Diskussionspapiere/20190205-diskussionspapier-effekte-kuenstlicher-intelligenz-in-der-wirtschaftswissenschaftlichen-literatur.pdf?__blob=publicationFile&v=6 – letzter Zugriff: 26.5.2025).

Müller, Michael / Grimm, Petra (2016): Narrative Medienforschung. Einführung in Methodik und Anwendung, Konstanz, München.

Pancratz, Nils / Fandrich, Anatolij / Diethelm, Ira (2023): Didaktische Strukturierung von Unterrichtsmaterialien zum Thema „Künstliche Intelligenz", in: Kai Bliesmer / Michael Komorek (Hg.), Didaktische Rekonstruktion – fachdidaktischer Ansatz für aktuelle Bildungsaufgaben, Oldenburg, S. 84–96.

Pawelec, Maria / Biess, Cora (*2021*): Deepfakes. Technikfolgen und Regulierungsfragen aus ethischer und sozialwissenschaftlicher Perspektive, Baden-Baden.

Prinzing, Marlis (2024): Journalismus, in: Petra Grimm / Kai Erik Trost / Oliver Zöllner (Hg.), Digitale Ethik, Baden-Baden, S. 517–527.

Quiring, Oliver et al. (2024): Zurück zum Niveau vor der Pandemie – Konsolidierung von Vertrauen und Misstrauen. Mainzer Langzeitstudie Medienvertrauen 2023, in: Media Perspektiven 9, S. 1–14.

Rath, Matthias (2013): Medienethik – zur Normativität in der Kommunikationswissenschaft, in: Matthias Karmasin / Matthias Rath / Barbara Thomaß (Hg.), Normativität in der Kommunikationswissenschaft. Wiesbaden, S. 289–299.

Reporter ohne Grenzen (2023): Paris Charter on AI and Journalism, in: RSF (online unter: https://rsf.org/en/paris-charter-ai-and-journalism – letzter Zugriff: 19.1.2026).

Ridsdale, Chantel / Rothwell, James / Smit, et al. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, Halifax.

Schicha, Christian (2021): Bearbeitete Bilder – Techniken und Bewertungen visueller Veränderungen am Beispiel politischer Motive, in: Christian Schicha / Ingrid Stapf / Saskia Sell (Hg.), Medien und Wahrheit. Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", Baden-Baden, S. 173–203.

Schlegel, Marcel et al. (2024): Einführung in das interdisziplinäre Lehrkonzept von IKID: Ziele und Programmatik einer integrierten KI-Lehre (= Whitepaper-Serie zum Forschungsprojekt IKID: Interdisziplinäres KI-Exploratorium), (online unter: ai.hdm-stuttgart.de/research/ikid/whitepaper-serie – letzter Zugriff: 5.2.2026).

Schüller, Katharina (2019): Ein Framework für Data Literacy, in: AStA – Wirtschafts- und Sozialstatistisches Archiv 13 (3/2019), S. 297–317.

Singh, Rita (2020): The Role of Human Voice in the Communication of Digital Disinformation, in: Maya Mirchandani (Hg.), Tackling Insurgent Ideologies in a Pandemic World. ORF / Global Policy Journal, Neu-Delhi, S. 69–74.

Spiekermann, Sarah (2019): Digitale Ethik. Ein Wertesystem für das 21. Jahrhundert, München.

Stapf, Ingrid (2021): „Fake News" als eine (mögliche) Frage der Wahrheit? Medienethische Perspektiven auf Wahrheit im Kontext der Digitalisierung. In: Christian Schicha / Ingrid Stapf / Saskia Sell (Hg.), Medien und Wahrheit. Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", Baden-Baden, S. 97–119.

Stapf, Ingrid (2024): Desinformation, in: Petra Grimm / Kai Erik Trost / Oliver Zöllner (Hg.), Digitale Ethik, Baden-Baden, S. 315–326.

Székely, Éva / Miniota, Jura / Hejná, Mísa (2025): Will AI shape the way we speak? The emerging sociolinguistic influence of synthetic voices, in: Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, Association for Computational Linguistics, Mai 2025 (online unter: https://aclanthology.org/2025.iwsds-1/ – letzter Zugriff: 19.1.2026).

Tschopp, Marisa / Ruef, Marc / Monett, Dagmar (2022): Vertrauen Sie KI? Einblicke in das Thema Künstliche Intelligenz und warum Vertrauen eine Schlüsselrolle im Umgang mit neuen Technologien spielt, in: Miriam Landes / Eberhard Steiner / Tatjana Utz (Hg.), Kreativität und Innovation in Organisationen. Impulse aus Innovationsforschung, Management, Kunst und Psychologie, Wiesbaden, S. 319–346.

Weber, Max (2017): Politik als Beruf, Stuttgart, Ditzingen.

World Economic Forum (2025): Future of Jobs Report 2025: Insight Report, (online unter: https://www.weforum.org/publications/the-future-of-jobs-report-2025/ – letzter Zugriff: 19.1.2026).

Kommunikationsfreiheiten und Öffentlichkeiten als Basis verständigungsorientierter Streitkulturen in der Demokratie

Christian Schicha

Zusammenfassung

Der Beitrag setzt sich aus einer normativen Perspektive mit den verfassungsrechtlich verankerten Kommunikationsfreiheiten auseinander, die eine Voraussetzung für funktionierende Teilöffentlichkeiten bieten. Nur wenn der freie und offene Austausch von Meinungen und Argumenten möglich ist, kann sich eine konstruktive Streitkultur herausbilden, in der sachliche und zielführende Lösungen herausgearbeitet werden. Nur auf diesem Wege können demokratische Prozesse in Gang gesetzt werden, die nicht auf Macht und Ausbeutung basieren, sondern einer fairen Gemeinwohlorientierung folgen.

1. Kommunikationsfreiheiten

In dem Beitrag werden grundlegende Voraussetzungen skizziert, die erforderlich sind, damit konstruktive Kontroversen in Demokratien vom Typ der Bundesrepublik vollzogen werden können. Dazu sind deliberative Öffentlichkeiten erforderlich, die sich an den Maximen des kommunikativen Handelns orientieren. Sie vollziehen sich im persönlichen Gespräch ebenso wie über analoge und digitale Medienkanäle. Es wird die Auffassung vertreten, dass eine konstruktive Streitkultur nur auf der Basis einer freiheitlich-demokratischen Grundordnung entstehen kann, die sich an Gerechtigkeitsprinzipien orientiert, wobei sich Betroffene oder ihre Vertreter:innen ohne äußere Zwänge an öffentlichen Debatten beteiligen sollten.

Als normative Grundlage der Kommunikationsethik fungiert das Grundrecht auf freie Meinungsäußerung. Es ist im bundesdeutschen Grundgesetz, der UN-Menschenrechtscharta und in der Europäischen Menschenrechtskonvention verankert. Dazu gehören das Recht der freien Informationsbeschaffung ohne Zugangsbeschränkungen und das Recht, offen zu kommunizieren. Es werden Ideale an öffentliche Kommunikationsprozesse

formuliert, die Gerechtigkeit und Gleichheitsgrundsätze postulieren. Während die Ethik als Reflexions- und Steuerungsinstanz die Aufgabe hat, entsprechende Normen zu formulieren, geht es im Recht darum, die Missachtung derartiger Normen zu sanktionieren.

Während im traditionellen Rundfunk die Freiheit als gemeinschaftliche Aufgabe angesehen wurde, geht es in der Internetkommunikation um eine Freiheit im individualisierten Verständnis jeder einzelnen Person. Heesen (2016) zufolge kann die Freiheit als der zentrale Kern der Kommunikations- und Medienethik klassifiziert werden, in der Prozesse eines normativ angemessenen Meinungsaustausches überhaupt möglich sind. Ohne eine individuelle Freiheit ist dies nicht möglich. Darüber hinaus muss eine Gesellschaft demokratisch so aufgebaut sein, dass eine freie Rede ohne Reglementierung überhaupt möglich ist.

Die Kommunikationsfreiheit fungiert als Grundrecht in demokratischen Gesellschaften und bietet die Möglichkeit der freien Meinungsäußerung. Sie soll den offenen Austausch von Informationen und Argumenten sicherstellen. Derartige Freiheitsrechte sind auch mit einer Verantwortung verbunden. Wer sich äußert, sollte das Recht nicht zum Nachteil oder Schaden anderer missbrauchen. Insofern sind Desinformationen und Hassrede nicht durch das Recht der freien Meinungsäußerung gedeckt. Schmähungen und Hetze verstoßen gegen die in Artikel 1 des Grundgesetzes geschützte Menschenwürde.

Durch die grundgesetzlich geschützte Meinungs- und Kommunikationsfreiheit resultiert aber kein Bekenntniszwang. Niemand ist verpflichtet, sich zu artikulieren. Wer aber nicht schweigt, sondern sich öffentlich äußert, muss sich der Überprüfung seiner Aussagen stellen und Gegenargumenten begegnen. Aus der freien Wahl des Handelns oder Nichthandelns resultiert eine Verantwortung. Wer bewusst falsche Aussagen macht, kann dafür belangt werden. Dazu gehört unter anderem die strafbewährte Holocaustleugnung.

Für die juristische Bewertung ist relevant, ob eine falsche Tatsachenbehauptung als Irrtum klassifiziert werden kann, ob es sich um ein ungeprüft übernommenes Gerücht handelt oder eine bewusst verbreitete Falschmeldung, um anderen zu schaden. Vorsicht ist auch bei der Weitergabe von privaten Inhalten in sozialen Netzwerken angebracht. Wenn Menschen bereit sind, im Internet Informationen über sich weiterzugeben und diese zur moralischen Selbstdarstellung für ein positives Reputationsmanagement nutzen, ermöglichen sie es den Onlineplattformen, daraus Kapital zu schlagen, indem sie die Daten erfassen und auswerten (vgl. Hübl 2024).

Obwohl die daraus resultierten Gefahren vielfach bekannt sind, kann in diesem Zusammenhang von einem *privacy paradox* gesprochen werden, da viele Menschen bereit sind, eine Vielzahl von privaten bis hin zu intimen Daten und Informationen weiterzugeben. Paradox ist dieser Sachverhalt deshalb, weil Menschen dies tun, obwohl sie genau wissen, dass ihre Einträge für kommerzielle Zwecke genutzt und gegebenfalls auch missbraucht werden können. Aufgrund dieses Verhaltens kann neben einer personalisierten Onlinewerbung (*targeting*) eine Kontrolle und Überwachung resultieren. Darüber hinaus kann die eigene Darstellung im Netz zum Beispiel durch die Verwendung von Symbolen in Form einer Regenbogenfahne oder einer Deutschlandfahne im Social-Media-Profil einer Person Rückschlüsse auf die politische Position der Absender:innen liefern (vgl. Hübl 2024). Eine notwendige Bedingung, derartige Meinungsäußerungen offen und ohne Androhung von Sanktionen artikulieren zu können, besteht darin, dass eine demokratisch verfasste Öffentlichkeit den Raum bietet, dies gefahrlos zu tun.

2. Öffentlichkeiten

Die Öffentlichkeit steht für eine Selbstbeschreibung demokratischer Gesellschaften. In ihr findet die Begegnung zwischen Bürger:innen statt. Sie können dort miteinander diskutieren und sich eine Meinung bilden. Dabei dienen politische Diskussionen in Demokratien vom Typ der Bundesrepublik Deutschland der Koordinierung von relevanten Angelegenheiten des allgemeinen Interesses. Wo gesellschaftliche Kontroversen und Streitpunkte aufkommen, bildet sich Öffentlichkeit.

Da Kommunikation als Austauschprozess und Verständigungsmittel über Bedeutungen definiert werden kann, ist die Öffentlichkeit das Forum, die Plattform oder der Resonanzboden, in dem öffentliche Kommunikation stattfindet. Nur wenn Rahmenbedingungen für eine funktionierende Öffentlichkeit ohne äußeren Druck vorhanden sind, können Argumente und Begründungen ohne Hindernisse ausgetauscht werden. Die Herstellung von Öffentlichkeit setzt Offenheit und Transparenz voraus. Sie bildet damit den Gegenpol zur Privatheit und zum Geheimnis. „Öffentlich nennen wir Veranstaltungen, wenn sie im Gegensatz zu geschlossenen Gesellschaften, allen zugänglich sind – so wie wir von öffentlichen Plätzen sprechen oder von öffentlichen Häusern“ (Habermas 1990: 217). Öffentlichkeit fungiert als eine Sphäre kommunikativen Handelns unter der Beteiligung mehrerer

Akteur:innen. Sie kann als Bereich klassifiziert werden, in dem sich Meinungen mit Interessen verdichten können. Durch öffentliche Austauschprozesse können sich Initiativen formieren, die sich z.B. als soziale Bewegungen zusammenschließen können. Meinungen und Interessen werden im öffentlichen Raum auf diesem Wege gebündelt und organisiert.

Öffentlichkeit ist nach Habermas ein normativ gehaltvoller Begriff. Ihm zufolge werden in demokratischen Gesellschaften kollektive Probleme dort definiert und nach Relevanz und Lösbarkeit geordnet. Öffentlichkeit wird als der zentrale Raum klassifiziert, in dem die Menschen „Werte, Themen, Beiträge und Argumente" (Habermas 1992: 624) austauschen. Dort kann über divergierende Auffassungen gestritten werden, um schlussendlich ein gemeinsames Einverständnis zu erzielen. Insgesamt sollen die Interessen aller Bürger:innen eines Gemeinwesens in die öffentliche Willens- und Meinungsbildung einbezogen oder zumindest advokatorisch – etwa in Hinblick auf zukünftige Generationen oder Unmündige – durch Interessenvertreter:innen zur Geltung gebracht werden (vgl. Birnbacher 1988). Dabei wird das Prinzip der Gerechtigkeit verfolgt, das aus einer kommunikationsethischen Perspektive eine mehrfache Bedeutung umfasst. Die Partizipation soll gewährleisten, dass alle Betroffenen sowohl das Recht als auch die Möglichkeit besitzen sollten, an Diskursen teilzunehmen. Nur so kann eine dadurch gewährleistete Emanzipation dazu beitragen, dass kommunikative und gesellschaftliche Verhältnisse erreicht werden können, in denen Gerechtigkeit angestrebt wird. Die Advokation macht es möglich, dass alle Vertreter:innen der Betroffenen an Debatten teilnehmen, um dem Anspruch auf partizipative Kommunikation gerecht zu werden (vgl. Thomaß et al. 2024). Das Ziel liegt schlussendlich darin, dass gerechte Diskurse geführt werden, die sich an den Leitlinien der Partizipation aller Betroffenen oder ihrer Vertreter:innen orientieren.

Im Forum der Öffentlichkeit sollen Meinungen und Argumente artikuliert werden, Standpunkte ausgetauscht und durch Diskussion und gegenseitige Überzeugung sowie durch Abwägung von Pro und Kontra zu einer Konsens-, Kompromiss- oder Mehrheitsentscheidung gelangen. Dort steht der kommunikative Austausch im Mittelpunkt des Interesses. Diese Vorstellung fordert, dass Akteur:innen ihre Interessen in einem Argumentationsverfahren mit Begründungen abstützen. Die Öffentlichkeit manifestiert sich als Kreis mündiger und aufgeklärter Bürger:innen einer Gesellschaft, in der die aggregierten Meinungen in rationalen und freien Diskussionen zusammengetragen werden. Eine funktionierende Öffentlichkeit wird in der Vorstellung mit ethischen „Orientierungen an Gewaltfreiheit, Zivilität,

Zugangsfreiheit und Rationalität der Kommunikationspraktiken" (Höhne 2019: 9) in Verbindung gebracht. Ein aus dieser normativen Perspektive wünschenswertes Verhalten sollte dabei nicht strategisch, sondern kommunikativ ausgerichtet sein.

3. Kommunikatives Handeln

Der Begriff des kommunikativen Handelns setzt auf die Rationalitätspotenziale sprachlicher Verständigung, die aber nur initiiert werden können, wenn alle Betroffenen oder ihre Vertreter:innen am Diskurs partizipieren können: „Eine Öffentlichkeit steht und fällt mit dem Prinzip des allgemeinen Zugangs. Eine Öffentlichkeit, von der angebbare Gruppen eo ipso ausgeschlossen wären, ist nicht nur unvollständig, sie ist vielmehr gar keine Öffentlichkeit" (Habermas 1990: 156).

Sie fungiert in diesem Verständnis sowohl als Verfahrens- und Ordnungsprinzip bei Entscheidungsprozessen des politisch-administrativen Systems wie auch als zeitliches und räumliches Zugänglichkeitsprinzip. Öffentlichkeit gilt für einen offenen Kreis von Akteur:innen als prinzipielle Option der Partizipation an spezifischen Diskussionsprozessen, aus denen sich die öffentliche Meinung herausbilden kann. Die Betroffenen sollen in diesem idealtypischen Verständnis einen Einblick in den Stand der Meinungs- und Willensbildung erhalten, um mit Hilfe der dadurch erworbenen Kenntnisse ihre Interessen adäquat wahrnehmen zu können.

Habermas (1990) zufolge bewirkt das Prinzip der Öffentlichkeit eine kritische Publizität, in der herrschaftsemanzipierte, verständigungsorientierte und kontinuierlich demokratische Prozesse der auf ein Gemeinwohl gerichteten Meinungsbildung verfolgt werden. Er verlangt, dass politische Entscheidungen generell an einen permanenten Prozess öffentlicher demokratischer Meinungsbildung zurückgebunden sein müssen, und vertritt ein normatives Leitbild, in dem die Öffentlichkeit durch rational abwägende Kommunikation die Partizipation verstärken und zu einer besser legitimierten und qualitativ verbesserten Entscheidungsfindung beitragen kann. Öffentliche Diskurse behandeln praktische Fragen des kollektiven Zusammenlebens, aber auch normative Ansprüche des Ausgleichs von Interessen. Von einer Rationalität der öffentlichen Meinung ist auszugehen, sofern sie das Ergebnis freier, für alle zugängliche und vor allem diskursiver Beratungen ist. Habermas entwickelt ein normatives Idealmodell von Öffentlichkeit als kommunikativen Bereich, in dem alle Bürger:innen mit begründeten

Argumenten öffentliche Belange diskutieren sollen. Das Ergebnis dieser vernünftigen Meinungsbildung fungiert in diesem Idealmodell als Grundlage politischer Entscheidungen. Habermas setzt auf die „Produktivkraft Kommunikation" (Habermas 1990: 36). Für ihn ist die politische Öffentlichkeit ein Inbegriff von Kommunikationsbedingungen, aus denen die öffentliche Meinungs- und Willensbildung eines Publikums resultieren soll. Öffentlichkeit avanciert somit zu einer Grundkategorie einer normativ angelegten Demokratietheorie. Die politische Öffentlichkeit bezieht sich auf politische Entscheidungen und dient der Artikulation von Ansprüchen, der Definition von Themen und Streitfragen, bei denen ein öffentliches Interesse vorausgesetzt wird, sowie der Findung allgemein akzeptierter Problemlösungen. Öffentlichkeit in dieser diskursiven Variante soll nicht nur die vorhandene Pluralität der Partikularinteressen spiegeln, sondern wird mit dem Ziel reflektiert, kommunikative Verständigungen zu erreichen, die das politische System beeinflussen.

Habermas (1990) differenziert zwischen autonomer und vermachteter Öffentlichkeit. Die autonome Öffentlichkeit wird durch das Muster kommunikativer Verständigung im Sinne eines Diskurses geprägt, während die vermachtete Öffentlichkeit primär strategische Interessen verfolgt.

> „Eine ‚vermachtete Öffentlichkeit' lässt den fairen Austausch von Argumenten unmöglich werden. Die ökonomische Macht von Verlagen, reichen Einzelpersonen oder Lobbygruppen kann durch gezielte Kampagnen Debatten lenken. Die politische Macht kann durch das Strafrecht oder die Verwaltung unliebsame Meinungen unterdrücken. In all diesen Fällen wird der Kampf um Deutungsmacht unfair ausgekämpft. Keine Öffentlichkeit ist machtfrei, aber Öffentlichkeiten können mehr oder weniger vermachtet sein" (Heidenreich 2024: 7).

Ein aktuelles Beispiel für eine Form der vermachteten Öffentlichkeit ist US-Milliardär und Unternehmer Elon Musk, der Donald Trump während seines Wahlkampfes 2024 finanziell unterstützt hat und nach dessen Wahlsieg auch als dessen Berater agierte. Er versuchte auch Einfluss auf den deutschen Wahlkampf zur Bundestagswahl 2025 zu nehmen, indem er Kanzler Olaf Scholz und Bundespräsident Frank-Walter Steinmeier beleidigte und ein Gespräch mit AfD-Chefin Alice Weidel über seine Plattform X führte. Dort wurden zahlreiche Falschaussagen von beiden Diskutant:innen über die Migration und den Nationalsozialismus artikuliert. Dies hat der Faktencheck der ARD-Tagesschau nachgewiesen (vgl. Reveland/Siggelkow 2025). Aus den unterschiedlichen Formen, in denen sich Öffentlichkeiten heraus-

bilden können, können öffentliche Meinungsbildungsprozesse in Form von Anschlussdiskursen entstehen, sofern die gesellschaftlichen Kommunikations-, Informations- und Partizipationsverhältnisse diesen Schritt zulassen.

Neben der raumzeitlichen Bestimmung als Ort des Diskurses fungiert Öffentlichkeit als Prozess. Sie wird als Diskussions- und Aushandlungsprozess manifestiert und ist niemals abgeschlossen. In diesem Verständnis ist sie offen für neue Einflüsse und Akteur:innen.

Insgesamt kann von einer gesamtgesellschaftlichen Öffentlichkeit in modernen demokratischen Gesellschaften nicht mehr ausgegangen werden. Sie wird vielmehr durch eine Vielzahl von Gruppen- und Spezialöffentlichkeiten ersetzt, die sich über die unterschiedlichsten Kanäle artikulieren. Es kann also die eine Öffentlichkeit in einer komplexen und ausdifferenzierten Gesellschaft nicht geben. Sie bildet sich vielmehr auf verschiedenen Ebenen und in verschiedenen medialen Formaten heraus. Als ethisches Kriterium fungiert die Chancengleichheit zur freien Meinungsäußerung. Publizität avanciert selbst zur Norm. Denn gerade unter den Bedingungen demokratischer Herrschaftslegitimation avanciert Öffentlichkeit zum ethischen Gebot. Gerhards (1997) hat die zentralen Anforderungen an öffentliche Kommunikation wie folgt zusammengefasst: Hinsichtlich der Verteilung von Sprecher:innenrollen in der öffentlichen Arena sollen möglichst alle Positionen und Interessenlagen widergespiegelt werden. Jede Stimme in der Arena hat die gleiche Berechtigung und darf nicht von anderen am Sprechen gehindert werden. Nicht das Argument oder die Begründung der Position ist relevant, sondern der Zugang zum Diskurs. Die zentrale normative Anforderung an Öffentlichkeit ist ihre prinzipielle Zugangsoffenheit. Erforderlich ist eine argumentativ ausgerichtete Form einer normativ präferierten Öffentlichkeit, die Habermas (1992) den Bürger:innen selbst zurechnet.

4. Deliberation

Eine daraus resultierende Öffentlichkeit im Sinne eines deliberativen Verständigungsmodells umfasst auch zivilgesellschaftliche Gruppen, die an die Interessen und Erfahrungen der Menschen anknüpfen. Für Habermas stellt die Deliberation einen bedeutenden Bestandteil demokratischer Öffentlichkeit dar. Nach dieser Konzeption werden Positionen und Themen mit Argumenten begründet, die durch den geregelten Austausch von Informationen zwischen Parteien und Gruppen entsprechende Vorschläge einbringen.

Kontroverse Diskurse, speziell im Internet, orientieren sich aber auch an anderen Formen der Debatte.

> „Deliberation wird möglicherweise nicht mit dem aggressiven politischen Streit gleichgesetzt, sondern als rationale und sachliche Auseinandersetzung um Sachprobleme begriffen. Der Sache nach hat sich die Deliberation seit dem Zeitalter der Aufklärung fest in das heutige Bild liberaler Demokratien eingeschrieben. Demokratie kann grundlegend als Verständigung und damit als Kommunikationsprozess interpretiert werden. Die Präsenz von Hass, Streit und Konflikten sind im Zeitalter von Social Media aber ein Dauerzustand und für Bürger:innen zwangsläufig allgegenwärtig" (Roberts/Filipovic 2022: 133).

Das deliberative Öffentlichkeitsmodell setzt normative Standards, die sich auf Form und Inhalt der in der Öffentlichkeit kommunizierten Stellungnahmen beziehen. Öffentlichkeit ist hier nicht nur der Repräsentanz der Meinungen und Stellungnahmen, sondern auch der reflektierten kommunikativen Vermittlung zwischen diesen Positionen verpflichtet. Durch eine angemessene Form der kommunikativen Interaktion können einem deliberativen Verständnis zufolge Interessens- oder Wertkonflikte in der Öffentlichkeit zu einem Konsens, mindestens aber zu einem rationalen Kompromiss gebracht werden. In diesem Modell öffentlicher Meinungsbildung wird davon ausgegangen, dass Öffentlichkeit nicht nur die vorhandene Pluralität von Partikularinteressen widerspiegelt, sondern dass darüber hinaus argumentative Diskurse mit dem Ziel verbunden werden, kommunikative Verständigungsprozesse voranzutreiben. Der Diskurs soll in diesem Kontext als eine offengelegte politische Auseinandersetzung über relevante Themen, die das Gemeinwohl tangieren, verstanden werden.

Es sollte aber nicht übersehen werden, dass Öffentlichkeiten sich immer auch in einem Machtgefüge mit strukturellen Zwängen befinden. Gegenöffentlichkeiten der Zivilgesellschaft formieren sich in einem Raum, in dem unterschiedliche Interessen strategisch durchgesetzt werden.

Die Klimaaktivistin Luisa Neubauer bringt diesen Aspekt am Beispiel der fossilen Energiegewinnung auf den Punkt. Sie kritisiert Entscheidungen, die sich gegen nachhaltige Energieprojekte aussprechen und am Abbau der Kohle festhalten, obwohl dies nicht nachhaltig ist.

> „Es sind Blitzlichter einer Welt, und einer Gesellschaft, bei der man meinen könnte, sie habe den Verstand verloren, so etwas lässt sich nicht argumentativ oder rational, ökonomisch oder parteipolitisch erklären,

> hier reicht es auch nicht mehr, auf fossile Desinformation oder fossiles Marketing zu verweisen. Die Erklärung dafür liefert nur eins: Macht. Genaugenommen: Fossile Macht" (Neubauer 2023: 25).

Um dieser Macht entgegenzutreten und Öffentlichkeit für die eigenen Anliegen zu generieren, artikulieren sich Bewegungen wie *Fridays for Future* bei Protesten auf der Straße bis hin zu Internetkampagnen, um gegen den menschengemachten Klimawandel zu protestieren. Diese Aktionsformen werden auch in der Medienberichterstattung aufgegriffen. Dabei werden grundlegende Fragen einer Klimaethik diskutiert, die die Verantwortung gegenüber zukünftigen Generationen reflektieren (vgl. Müller-Salo 2020).

5. Medien

Öffentlichkeitstheorien sind übergreifend und verweisen nicht auf konkrete Medienkontexte mit ihren spezifischen Erscheinungsformen und strukturellen Zwängen. Gleichwohl werden konkrete Defizite im Rahmen der Medienberichterstattung benannt, die dazu führen können, dass sich keine kritische Öffentlichkeit herausbildet. Dazu gehören neben der Zensur mangelnde Recherche und eine bisweilen systematische Vernachlässigung von gesellschaftlich relevanten Themen und Meldungen.

Die Herstellung von Öffentlichkeit über gesellschaftlich relevante Sachverhalte in Demokratien wie der Bundesrepublik Deutschland gehört zu den zentralen Aufgaben von Medien, die öffentlichkeitskonstituierend agieren. Die Öffentlichkeit wird als Kommunikationssystem interpretiert, in dem Informationen und Meinungen artikuliert und ausgetauscht werden. Zentral ist dabei der offene Zugang zu Informationen ohne Blockaden. „Eine demokratisch legitimierte Öffentlichkeitssphäre erfordert einen Zugang, der thematisch offen ist, sowie von einer Gleichheit der Beteiligten und einem prinzipiell nicht abgeschlossenen Publikum ausgeht" (Winter 2010: 89).

Öffentlichkeit im Verständnis einer Kontroll- und Kritikfunktion über die Medien dient der Transparenz gesellschaftlich relevanter Entscheidungen und Entwicklungen, informiert über die Ziele von Interessensgemeinschaften und ist grundgesetzlich durch die Meinungs-, Rede-, Versammlungs- und Pressefreiheit geschützt. Insofern ist der Begriff der Öffentlichkeit normativ angelegt. Es geht um die Frage, wie eine Öffentlichkeit ausgerichtet sein sollte, um den Kriterien einer funktionierenden Demokratie zu entsprechen.

Durch Massenmedien sind plurale Öffentlichkeiten entstanden, die sich aus unterschiedlichen Zugängen (Print, Rundfunk, Internet) sowie öffentlich-rechtlichen und privat-kommerziellen Organisationsformen und Trägern zusammensetzen. Ihre Inhalte werden durch spezifische Medienstrategien aufgrund einer Orientierung an Auswahlkriterien der Neuigkeit, Verkürzung, Vereinfachung, Personalisierung und Unterhaltungszentrierung im Rahmen der konkreten Programmausgestaltung geprägt, um Interesse und Aufmerksamkeit beim Publikum zu erzeugen. Sie informieren über Entwicklungen, die über den individuellen Erfahrungshorizont hinausgehen und bilden ein frei zugängliches Podium, das Wissen verfügbar macht und einordnet. Verständigung, Urteilsvermögen, Sachkenntnis und Integrationsfähigkeit sollen nach diesem idealtypischen Verständnis eines professionellen Journalismus durch die Berichterstattung über die Kanäle bedient werden.

Massenmedien verfügen einerseits über einen sozial integrierenden und festigenden Charakter, sofern alle informiert werden, andererseits kommt ihnen eine innovative Funktion zu, indem sie über Ereignisse, Neuigkeiten und Tendenzen zu Veränderungen berichten und damit Wertewandlungsprozesse dokumentieren. Es findet zumindest über die klassischen Massenmedien keine symmetrische Dialogorientierung statt, bei der die rationale Prüfung von Geltungsansprüchen im Mittelpunkt steht. Die uneingeschränkte Information und Chancengleichheit ist ebenso wenig vorhanden wie die Möglichkeit zur Interaktion, von Leser:innenbriefen einmal abgesehen. Es handelt sich primär um Prozesse der einseitigen Informationsaufnahme, die jedoch Anschlussdiskurse zulassen.

Von der Medienberichterstattung in der Demokratie wird insgesamt erwartet, eine freie und umfassende Meinungsbildung der Öffentlichkeit möglich zu machen. Dabei ist über gesellschaftlich relevante Sachverhalte so zu berichten, dass das Publikum einen fundierten Überblick über Ereignisse erhält, die glaubwürdig eingeordnet werden. Die Nachrichten, Kommentare und Meldungen sollten eine Meinungsvielfalt abbilden und sich am ethischen Leitwert der Wahrheitsfindung orientieren. Medien können durch die Qualität der Berichterstattung dazu beitragen, die gesellschaftliche Integration zu fördern. Das Publikum soll dadurch in die Lage versetzt werden, sich aufgrund der angebotenen Informationen ein eigenes Bild über Ereignisse machen zu können.

6. Digitale Zugänge

Habermas hat sich in seiner Habilitationsschrift *Strukturwandel der Öffentlichkeit* im Jahr 1962 mit Formen einer idealtypischen bürgerlichen Öffentlichkeit auf Basis der Aufklärung und der Entwicklung der modernen bürgerlichen Gesellschaft mit Bezug auf das Privateigentum und die Ökonomie beschäftigt. Dabei nimmt er Bezug auf die Funktionslogik spezifischer Medien u.a. in Form der literarischen Kritik, den Konflikt um die öffentliche Meinung und das massenmediale Entertainment. Nachdem von ihm zunächst eine schädliche Wirkung von Massenmedien auf die Öffentlichkeit aufgrund einer unreflektierten und unkritischen Berichterstattung konstatiert worden ist, wird im Vorwort der Neuauflage von 1990 das kritische Potenzial der Rezipient:innen betont, Medieninhalte eigenständig einzuordnen und angemessen zu bewerten. Aktuell geht Habermas auf den digitalen Strukturwandel der Öffentlichkeit als qualitative Veränderung ein. Er verweist auf digitale Plattformen, die den Nutzer:innen die Möglichkeit geben, sich unabhängig von linearen Medienangeboten zu informieren und zu unterhalten.

> „Sie verändern auf radikale Weise das bisher in der Öffentlichkeit vorherrschende Kommunikationsmuster. Denn sie ermächtigen alle potenziellen Nutzer prinzipiell zu selbstständigen und gleichberechtigten Autoren. Die ‚neuen' unterscheiden sich von den traditionellen Medien dadurch, dass sich digitale Unternehmen diese Technologie zunutze machen, um den potenziellen Nutzern die unbegrenzten digitalen Vernetzungsmöglichkeiten wie leere Schrifttafeln für eigene kommunikative Inhalte anzubieten" (Habermas 2022: 44).

Die angebotenen Inhalte der Plattformen werden von den Rezipient:innen ausgewählt und zu einer selbst bestimmten Zeit konsumiert. Dies gilt auch für lineare Programme, die über Mediatheken abgerufen werden können. Sie bieten

> „[…] eine vielseitig vernetzungsoffene kommunikative Verbindung für den spontanen Austausch möglicher Inhalte zwischen potenziell vielen Nutzern. Diese unterscheiden sich nicht schon aufgrund des Mediums in ihren Rollen voneinander; sie begegnen sich vielmehr als prinzipiell gleiche und selbst verantwortliche Teilnehmer am kommunikativen Austausch von spontan gewählten Themen" (Habermas 2022: 45).

Insofern entsteht eine größere Handlungsautonomie auf Seiten der Rezipient:innen, die selbstbestimmt entscheiden können, wann sie an welchem Ort über welche digitalen Kanäle Inhalte konsumieren oder interaktiv mit anderen teilen möchten. Damit erweitern sich die Kommunikationsmöglichkeiten, die alternative Optionen der Partizipation bereitstellen. Somit bilden sich neue Teil- und Gegenöffentlichkeiten, die öffentliche Debatten neu organisieren und strukturieren. Es entstehen neue Formen der Vernetzung und Artikulation von Interessen, die auch marginalisierten Gruppen die Möglichkeit bieten, sich an Diskursen publikumswirksam zu beteiligen. Dies gilt aber auch für weitere Bereiche.

> „Der Begriff der automatisierten Öffentlichkeiten bringt zum Ausdruck, dass technische bzw. automatisierte Akteure wie z. B. Sprachassistenten, Softbots oder Suchmaschinenalgorithmen eine wachsende Rolle für die mediale Kommunikation spielen. Sie beeinflussen die demokratische Meinungsbildung durch Selektionsprozesse (Gatekeeping), aber auch durch aktive Manipulationen der Meinungsbildung in den sozialen Medien mithilfe von Social Bots" (Heesen 2024: 250).

Rechtspopulistische, antifeministische und rassistische Bewegungen gewinnen ebenfalls an Deutungsmacht. Das Spektrum der sozialen Situationen, Kontexte und Praktiken, in denen unterschiedliche Öffentlichkeiten verortet werden, hat sich erweitert. (vgl. Jung/Kempf 2023). Auch ohne klassische Instanzen wie Verlage, Redaktionen und Medieninstitutionen ist es nun möglich, eigene Beiträge zu verfassen und aktiv in öffentliche Debatten einzugreifen. Durch die Nutzung u.a. von X, Facebook und Instagram ist es möglich, sich zu artikulieren und zu reagieren. Aussagen über die Qualität der digitalen Austauschprozesse lassen sich aber erst nach Prüfung der Inhalte treffen. Habermas (2022: 45) vertritt hier eine kritische Position, indem er anmerkt: „Dieses große emanzipatorische Versprechen wird heute zumindest partiell von den wüsten Geräuschen in fragmentierten, in sich selbst kreisenden Echokammern übertönt." Eine professionelle Auswahl und diskursive Prüfung der Qualität von Medieninhalten findet nicht flächendeckend statt und obliegt den Mediennutzerinnen selbst. Diese „Plattformisierung der Öffentlichkeit" (Habermas 2022: 56) verdrängt zunehmend klassische Medienformate. Klinger konstatiert:

> „Die Funktionsweise von Öffentlichkeiten, die Dynamik von Informationsflüssen und Diskursen hat sich fundamental gewandelt. Die einstigen Gatekeeper – Redaktionen, Verlage, Multiplikatoren – verlieren massiv

> an Bedeutung. Zugleich nimmt der Einfluss von Technologien auf Meinungsbildungsprozesse, Informiertheit und Diskurse enorm zu. Soziale Medien, Suchmaschinen und Online-Portale sind keineswegs unkuratierte Umgebungen, sondern maßgeblich von Algorithmen geprägt, die auswählen, welche Inhalte welche Nutzerinnen und Nutzer sehen oder auch nicht sehen" (Klinger 2020: 50f.).

Es gibt also eine veränderte Kommunikationspraxis, die durch eine abnehmende Qualität des politischen Diskurses und einer Polarisierung der Gesellschaft geprägt ist. Heesen beschreibt die Problematik wie folgt:

> „Während das Ziel eines normativ ausgelegten Öffentlichkeitsbegriffs gesellschaftliche Verständigung und Integration durch Kommunikation sind, ist das Ziel der Plattformbetreiber die Optimierung ihrer Geschäftsinteressen durch die Bereitstellung von Kommunikationsmitteln. Dieser Zielkonflikt spiegelt sich zum Beispiel wider in dem Bestreben der Plattformbetreiber, ihre Nutzerinnen und Nutzer durch möglichst viele Anreize in ihren Diensten zu halten und zur Kommunikation anzuregen. Im Vordergrund stehen also nicht Qualität und Relevanz der Kommunikation, sondern die Generierung von Kommunikation als solcher" (Heesen 2024: 239).

Dabei spielen die so genannten Sozialen Medien eine zentrale Rolle. Habermas (2022: 52) konstatiert einen „dramatischen Bedeutungsverlust der Printmedien", ein „sinkendes Anspruchsniveau" öffentlicher Diskurse und sieht eine Gefahr der Demokratie durch Soziale Netzwerke. Er sieht durch die Angebote sozialer Medien – ganz in der Tradition der Kritischen Medientheorie – einen Trend zur Entpolitisierung, da sich politische Programme auch an Unterhaltungs- und Konsumangeboten orientierten. Insofern ist es zentral, Diskurse über die verschiedenen Medien so zu organisieren, dass ein öffentlicher Austausch über relevante Sachverhalte gewährleistet ist.

> „Es ist deshalb keine politische Richtungsentscheidung, sondern ein verfassungsrechtliches Gebot, eine Medienstruktur aufrechtzuerhalten, die den inklusiven Charakter der Öffentlichkeit und einen deliberativen Charakter der öffentlichen Meinungs- und Willensbildung ermöglicht" (Habermas 2022: 67).

Dabei sind grundlegende Rahmenbedingungen und Regeln erforderlich, die zum Teil missachtet werden. Es geht um konstruktive Argumentations-

verfahren, die Angemessenheitsbedingungen für diskursive Verfahren im analogen und digitalen Raum aufzeigen.

7. Eine konstruktive Streitkultur als Basis für die Demokratie

> „Das liberale Demokratieverständnis stellt […] hohe Anforderungen an die öffentliche Kommunikation. Inklusivität, Responsivität, Zivilität sowie die Bedingung der argumentativen öffentlichen Begründung stellen die normativen Grundannahmen auch der medienvermittelten öffentlichen Kommunikation dar" (Roberts/Filipovic 2022: 137).

Lessenich (2019: 7) vertritt die Auffassung, dass die Demokratie „möglicherweise der Hochwertbegriff der westlichen Moderne schlechthin ist". In dieser Staatsform gelten Grundwerte wie Gleichheit und Freiheit als Menschenrechte. Dabei ist die Meinungsfreiheit eine zentrale Voraussetzung.

> „Die Meinungsfreiheit ist eines der grundlegenden Menschenrechte in modernen Demokratien. Einerseits ermöglicht sie die individuelle Persönlichkeitsentfaltung und Meinungsbildung. Andererseits ist sie zentral für das Funktionieren demokratischer Prozesse, weil sie gewährt, divergierende Ansichten öffentlich zum Ausdruck zu bringen, sich gegenüber der Politik zu artikulieren und in freien, diskursiven Aushandlungsprozessen zu konsensfähigen Lösungen politischer Fragen zu kommen" (Rothut et al. 2023: 50).

Das Grundrecht auf Meinungsfreiheit gibt den Bürger:innen nicht nur die Möglichkeit, sich an öffentlichen Diskussionen zu beteiligen, sondern es fördert die persönliche und öffentliche Meinungsbildung, sofern eine aktive Teilnahme an Diskussionen zu gesellschaftlich relevanten Themen in einem geschützten Rahmen möglich ist, die einen gewaltfreien kommunikativen Ausgleich von Interessen ermöglicht. Dafür muss die Demokratie verteidigt und geschützt werden. In einer Demokratie zu leben bedeutet, dass Menschen in die Lage versetzt werden, selbst über ihr eigenes Leben zu verfügen. Das setzt Rahmenbedingungen voraus, in denen Solidarität und Empathie die Grundlage jedes gerechten Gemeinwesens bilden (vgl. Süß/Torp 2021). Hierbei sind faire Verhältnisse erforderlich, die durch freie Wahlen und konstruktive Debatten zum Ausdruck kommen. Die Orientierung am Allgemeinwohl und der Abbau von Ungleichheiten sind wichtig für die demokratische Staatsform. Die Verfolgung von Prinzipien der Menschenwürde ist ebenso unverzichtbar wie die Gewaltenteilung durch eine

unabhängige Justiz und die Kontrolle der Regierung durch das Parlament. Für eine funktionierende Mediendemokratie sind Kommunikations- und Medienfreiheiten zentral, die ebenfalls eine Kontrolle und Kritik gegenüber den Mächtigen durch eine kritische Berichterstattung ausüben sollten (vgl. Görlach 2021).

> „Journalismus beeinflusst die Voraussetzung von Demokratie – die politische Kultur; die Sachlichkeit der Debatte und Qualität der Meinungsbildung; die entsprechende Lernbereitschaft der Bürgerinnen und Bürger; ihren mehr oder minder ausgeprägten Gemeinsinn; und den Zusammenhalt des Gemeinwesens – kurzum die Kraft der Demokratie" (de Weck 2024: 12).

Die Demokratie kann als besondere Form der Selbstbestimmung klassifiziert werden, die als normative Grundprinzipien dem Anspruch von Freiheit und Gleichheit aller Menschen genügt. Nur unter diesen Rahmenbedingungen, die auch den Schutz der Privatsphäre umfasst, kann eine offene Diskurskultur vollzogen werden, die auf der Basis vernünftiger Argumente mit guten Gründen diskursiv vollzogen wird. Einschüchterungen und Ausgrenzungen sind hingegen kontraproduktiv. (vgl. Nida-Rümelin 2023). Gleichwohl ist konstruktiver Streit nach den Regeln der Fairness ein Wesensmerkmal demokratischer Debatten.

> „Streiten und die freien Meinungsäußerungen sind der Sauerstoff der Demokratie. Demokratie eröffnet Diskursräume, schützt öffentliche Räume, in denen die unterschiedlichsten Perspektiven, Meinungen, Argumente angstfrei ausgetauscht werden können. Demokratie bedeutet Streiten. Opposition. Widerspruch. Im Gegensatz zur Diktatur, in der jegliche Kritik gegenüber den Mächtigen brutal unterdrückt wird – wie in Russland, China, Belarus, immer noch in großen Teilen der Welt – ist es in der Demokratie möglich, dass selbst diejenigen, die diese Demokratie zerstören wollen, ihre Meinung öffentlich kundtun können" (Friedmann 2021: 26).

Insofern sind die Rahmenbedingungen in Demokratien vom Typ der Bundesrepublik in Bezug auf die Meinungsäußerungsfreiheit deutlich besser. Daraus folgt, dass auch konträre Positionen ohne eine sachliche Fundierung ertragen werden müssen.

> „Tatsächlich darf man in offenen, freien Gesellschaften sehr viel sagen, was im Umkehrschluss bedeutet, dass man auch sehr viel aushalten

> muss. Meinungsfreiheit schließt die Meinung des anderen ein, wie dumm und beschränkt, wie überheblich oder engstirnig, wie verrückt oder gar wahnsinnig sie einem auch erscheinen mag" (Roßbach 2018: 12).

Unterschiedliche Interessen und Positionen sollten politisch zur Geltung kommen und wechselseitig ausgehandelt werden, um die Meinungsbildung transparent zu machen und Entscheidungsprozesse zu generieren. Dass dies aufwändig, anstrengend und teilweise langsam vonstattengeht, ist zwar mühsam, aber auf längere Sicht durchaus zielführend und demokratisch legitimiert.

Durch die rasante Verbreitung von Meldungen im Netz kommt es zu einer Menge an Informationsangeboten, die eben nicht mehr vernünftig verarbeitet werden können. Durch die permanente Überforderung funktioniert die sozialmediale Öffentlichkeit weniger nach rationalen, sondern primär nach emotionalen Prinzipien.

> „Wenn man jeden Tag mit hunderten Informationsbröckchen konfrontiert wird, zu denen man sich irgendwie glaubt verhalten zu müssen, dann ist fast die einzige Möglichkeit, dem Bauchgefühl zu folgen. Das Bauchgefühl kommt anhand von wissenschaftlich noch nicht vollständig geklärten Anhaltspunkten in sagenhaften 50 Millisekunden zu einem Urteil. Like, Dislike, Wut, Mitleid, Freude, Zusammengehörigkeit, Hass, Neid – das sind alles emotionale Dropse, die der Bauch beziehungsweise das Gehirn nach weniger als einer Sekunde bereits zu Ende gelutscht hat“ (Lobo 2016: 23).

Insofern ist ein reflektierter und zeitlich begrenzter Umgang mit den bisweilen problematischen Inhalten im Netz erforderlich, um eine substanzielle Meinungs- und Willensbildung zu bewerkstelligen. Dabei ist die Kommunikationsfreiheit, zu der die Wahl der Inhalte und technische Nutzung der Plattformen gehört, substanziell für die analoge und digitale Souveränität der Bürger:innen.

8. Fazit und Ausblick

Es mag naiv anmuten, wenn Kommunikationsfreiheiten in diesem Aufsatz angepriesen werden, die in der Praxis vielfach nicht existieren. Dennoch sind regulative Ideen auf der Idealebene aus einer normativen Perspektive ein wichtiges Fundament, um Ziele zu formulieren, die faktisch häufig

nicht umgesetzt werden. Sie dienen der Selbstvergewisserung einer demokratischen Gesellschaft, die auf einer konstruktiven Form des Meinungsaustausches in Form von Argumenten basiert. Dabei sollten möglich viele Betroffene unmittelbar oder advokatorisch mit in Entscheidungsprozesse mit eingebunden werden.

Grundsätzlich ist es wichtig, dass Macht legimiert ist und Machtmissbrauch sanktioniert wird. Um dies zu bewerkstelligen, ist eine Gewaltenteilung unverzichtbar. Die in Demokratien vom Typ der Bundesrepublik Deutschland gewählte Aufteilung in die Legislative (gesetzgebende Gewalt), Exekutive (ausführende Gewalt) und Judikative (rechtssprechende Gewalt) bieten die Möglichkeit der wechselseitigen Kontrolle, um Machtmissbrauch zu verhindern und den Schutz der Freiheitsrechte von Bürger:innen zu ermöglichen. Auch die als vierte Gewalt klassifizierte Medienberichterstattung besitzt nach wie vor wichtige Funktion, um Mächtige zu kontrollieren und Missstände, Machtmissbrauch und weitere Verfehlungen transparent zu machen. Die über die digitalen Medien verbreiteten Debatten können ebenfalls einen wichtigen Beitrag zur Ausübung von Kommunikationsfreiheiten leisten, sofern argumentative Mindeststandards eingehalten werden. Hass und Hetze verstoßen allerdings dagegen und sind zu sanktionieren.

Zudem sollten nicht nur die Produzierenden für die eingestellten Inhalte verantwortlich gemacht werden, sondern auch die Plattformanbieter*innen, die damit ihr Geld verdienen. Insofern sind Medienregulierungen erforderlich, die unabhängig von staatlichen Einflüssen, diese Aufgabe übernehmen. Eine konstruktive Streitkultur auf der Basis eines deliberativen Verständigungsmodells ist nach wie vor ein wichtiger Garant für die eine demokratisch legitimierte Öffentlichkeit, in der kontroverse Debatten in konstruktiver Form vonstattengehen sollten. Solidarität, Empathie, Respekt und wechselseitige Anerkennung als Grundhaltungen sind dabei notwendige Bedingungen, um diesem Anspruch gerecht zu werden (vgl. Schicha 2025).

Literatur

Birnbacher, Dieter (1988): Verantwortung für zukünftige Generationen, Stuttgart.

De Weck, Roger (2024): Das Prinzip Trotzdem: Warum wir den Journalismus vor den Medien retten müssen, Berlin.

Friedmann, Michel (2021): Streiten? Unbedingt! Ein persönliches Plädoyer, Berlin.

Gerhards, Jürgen (1997): Diskursive versus liberale Öffentlichkeit. Eine empirische Auseinandersetzung mit Jürgen Habermas, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 49, S. 1–34.

Görlach, Alexander (2021): Demokratie, Stuttgart.

Habermas, Jürgen (1990): Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft, 2.Aufl., Frankfurt am Main.

Habermas, Jürgen (1992): Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats, Frankfurt am Main.

Habermas, Jürgen (2022): Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Frankfurt am Main.

Heesen, Jessica (2024): Ethik in der öffentlichen digitalen Kommunikation. In: Barbara Thomaß et al. (Hg.), Ethik der öffentlichen Kommunikation. Eine kommunikationsethische Einführung, Wiesbaden, S. 227–261.

*Heesen, Jessic*a (2016): Freiheit, in: dies. (Hg.), Handbuch Medien- und Informationsethik, Stuttgart, S. 52–58.

Heidenreich, Felix (2024): Demokratie – ein umstrittener Begriff, in: Informationen zur politischen Bildung 361 (4/2024), S. 6–11.

Höhne, Florian (2019): „Öffentlichkeit" als Imagination und Ensembles sozialer Praktiken. Zur Relevanz einer Schlüsselkategorie öffentlicher Theologie in digitalen Kontexten, in: Ethik und Gesellschaft. Ökonomische Zeitschrift für Sozialethik (1/2019). https://doi.org/10.18156/eug-1-2019-art-1

Hübl, Philipp (2024): Moralspektakel. Wie die richtige Haltung zum Statussymbol wurde und warum das die Welt nicht besser macht, München.

Jung, Simone / Kempf, Victor (Hg.) (2023): Entgrenzte Öffentlichkeit. Debattenkulturen im politischen und medialen Wandel, Bielefeld.

Klinger, Ulrike (2020): Diskurskiller Digitalisierung? Warum das Internet nicht an allem Schuld, aber trotzdem ein Problem ist, in: Stephan Russ-Mohl (Hg.), Streitlust und Streitkunst. Diskurs als Essenz der Demokratie, Köln, S. 48–65.

Lessenich, Stephan (2019): Grenzen der Demokratie. Teilhabe als Verteilungsproblem, Stuttgart.

Lobo, Sascha (2016): Das Ende der Gesellschaft. Von den Folgen der Vernetzung, Tübingen.

Müller-Salo, Johannes (2020): Klima, Sprache und Moral. Eine philosophische Kritik, Stuttgart.

Negt, Oskar / Kluge, Alexander (2001): Entfremdete Öffentlichkeit. Öffentlichkeit und Erfahrung, in: Günter Helmes / Werner Köster (Hg.), Texte zur Medientheorie, Stuttgart, S. 285–288.

Neubauer, Luisa (2023): Sagen, was ist. Die Klimakrise im Diskurs, Tübingen.

Nida-Rümelin, Julian (2023): „Cancel Culture" – Das Ende der Aufklärung? Ein Plädoyer für eigenständiges Denken, München.

Reveland, Carla / Siggelkow, Pascal (2025): Gespräch von Musk und Weidel. Falschaussagen von Migration bis Nationalsozialismus, in: Tagesschau, 04. Februar 2025 (online unter: https://www.tagesschau.de/faktenfinder/musk-weidel-102.html – letzter Zugriff: 13.11.2025).

Roberts, Cindy Ricarda / Filipovic, Alexander (2022): Deliberation als Streitkultur? (Un-) Möglichkeiten der Deliberationstheorie in digitalen Zeiten, in: Christian Gürtler/ Marlis Prinzing / Thomas Zeilinger (Hg.), Streitkulturen. Medienethische Perspektiven auf gesellschaftliche Diskurse, Baden-Baden, S. 133–168.

Roßbach, Nikola (2018): Achtung Zensur! Über Meinungsfreiheit und ihre Grenzen, Berlin.

Rothut, Sophia et al. (2023): Meinungsfreiheit in Gefahr? Wie politische Einstellungen und individuelle Erfahrungen die Wahrnehmung der Meinungsfreiheit in Deutschland prägen, in: Studies in Communication and Media 12 (1/2023), S. 48–86.

Schicha, Christian (2025): Kommunikationsethik. Grundlagen – Debatten -Lösungsansätze, München.

Seeliger, Martin / Sevignani, Sebastian (2021): Zum Verhältnis von Öffentlichkeit und Demokratie. Ein neuer Strukturwandel? in: Martin Seeliger / Sebastian Sevignani (Hg.), Ein neuer Strukturwandel der Öffentlichkeit? Leviathan Sonderband 37, Baden-Baden, S. 9- 42.

Süß, Dietmar / Torp, Cornelius (2021): Solidarität. Vom 19. Jahrhundert bis zur Corona-Krise, Bonn.

Thomaß, Barbara et al. (2024): Ethik der öffentlichen Kommunikation. Eine kommunikationsethische Einführung, Wiesbaden.

Winter, Rainer (2010): Widerstand im Netz. Zur Herausbildung einer transnationalen Öffentlichkeit durch netzbasierte Kommunikation, Bielefeld.

Geltungsansprüche „pluralistisch-transzendentaler Öffentlichkeit“ als Grenzziehungen gegen holistische „qualitative Öffentlichkeit“. Überlegungen zur Integration der Ansätze von Habermas und Manheim im Rekurs auf Seyla Benhabib

Stefanie Averbeck-Lietz

Zusammenfassung

Der vorliegende Aufsatz begründet den Wert Pluralismus aus einer kommunikationsethischen Perspektive. Ausgangspunkt ist der interaktive Universalismus von Seyla Benhabib, der auf einer kritischen Lesart von Jürgen Habermas aufbaut. Diese wird um eine historische Dimension erweitert mit Blick auf Ernst Manheims frühe Überlegungen zu holistischen versus pluralistischen Öffentlichkeitsvorstellungen. Der Artikel ist theorieorientiert, gleichwohl lassen sich Anwendungsbezüge zu Fragen unter anderem von Populismus in seiner expressiven Reklamation holistischer, non-pluraler Perspektiven ebenso herstellen wie solche zur Entwicklung von Kommunikationskompetenz im Sinne der Anerkennung von Pluralität und Diversität. Dies führt auch zu der Frage, wie Gesellschaften sich der Problematik gegenseitiger Anerkennung als Gleiche und Gleichwertige stellen können – konfrontiert mit neuen, nicht-humanen Aktanten im öffentlichen und privaten Raum: der sogenannten Künstlichen Intelligenz, beziehungsweise Large Language Models.

1. Einleitung

Der vorliegende Aufsatz[1] möchte „Pluralismus als Wert“ (Mandry 2012: 229) kommunikations- und öffentlichkeitstheoretisch begründen. Chris-

1 Dieser Aufsatz baut auf der Antrittsvorlesung der Verfasserin am 11.9.2024 an der Universität Greifswald sowie ihres Vortrags anlässlich der gemeinsamen Tagung des Netzwerks Medienethik, der DGPuK-Fachgruppe Kommunikations- und Medienethik sowie der Akademie für politische Bildung in Tutzing im Februar 2025 auf.

toph Mandry[2] spezifiziert diesen als Meta-Wert des Zusammenlebens in (westlichen) Gesellschaften. Es handele sich um eine „demokratisch-politische Werthaltung oder auch Tugend", relevant für den gesellschaftlichen Zusammenhalt (ebd.: 235, vgl. auch Filipović 2021: 290). Nur scheinbar paradox zu einer pluralistischen Position bezieht sich der vorliegende Artikel auch auf *universalistische* Positionen, die jedem Menschen die gleichen Rechte zuweisen und Menschen nicht als Mittel, sondern als Zweck, als an sich wertvoll begreifen und daraus die Menschenwürde ableiten (vgl. Thomaß 2025; Fenner 2025: 80–125). Erst in *Anerkennung* unserer Verschiedenenheit können wir uns als gleichwertige Personen betrachten, als gleich wertvoll in Unterschiedlichkeit. Dabei muss gegenseitige Anerkennung Fragen der Akzeptanz oder Nicht-Akzeptanz spezifischer Verhaltensweisen, Sitten und Gebräuche etwa im interkulturellen Kontext nicht ausschließen (vgl. Thomaß 2025). Diskrepanzen auf der Ebene der Moral schließen sich gerade nicht aus, wenn der gemeinsame Bezug eine ethische Reflektion ist: sich gegenseitig in Verschiedenheit anerkennen *zu wollen*. Dies läßt sich mit Ernst Manheim, aber auch mit Jürgen Habermas argumentieren, wie im Folgenden gezeigt werden soll.

Vom Pluralismus *als* Wert und in diesem Sinne verstanden als Zielnorm ist der Begriff des *Wertepluralismus*, der empirischen Beschreibungen gesellschaftlicher Realität dienen kann, zu unterscheiden. Wertepluralismus als *empirisches* Phänomen wird in diesem Aufsatz nicht behandelt, gleichwohl spielt er als Bezugsrahmen eine Rolle: Einen Pluralismus *der* Werte, als eine „Einheit in Verschiedenheit" (Mandry 2012: 232), kann eine Gesellschaft als Differenzphänomen nur dann realisieren – so die Argumentation in diesem Aufsatz –, wenn sie die universelle Menschenwürde als Voraussetzung akzeptiert und auf dieser Basis Religionsfreiheit, sexuelle Selbstbestimmung oder andere Formen der Diversität, folglich *individuelle Rechtsansprüche als universelle Forderung*, einschließlich solcher der Kommunikationsfreiheit postuliert (vgl. Saxer 1996; Thomaß 2025: 58–59, 62; Beck 2021): „Anerkennung des Pluralismus hat als normativen Kerngehalt die Anerkennung des Anderen als gleichermaßen würdige Person [...] in ihrer konkreten Andersheit" (Mandry 2012: 235).

Es geht mithin in diesem Aufsatz weniger um die empirische Beschreibung von Wertepluralismus, sondern um die theoretische Begründung, *warum* kommunikationsethische Positionierungen *Pluralismus selbst als*

2 Mandry (2012: 230), von dem ich den Terminus „Pluralismus als Wert" entlehne, baut eine Argumentation aus „theologisch-ethischer" Sicht auf.

Wert einfordern können und sollten – in der konkreten Anwendung gerade gegenüber anti-pluralistischen Positionen zumal im politischen Diskursraum.

Der Soziologe Ernst Manheim (1900–2002) hat vor mehr als 90 Jahren anti-plurale, personell und/oder inhaltlich autoritär strukturierte Öffentlichkeiten mit Blick auf den aufstrebenden Nationalsozialismus beschrieben. Er nannte sie „qualitative Öffentlichkeiten" (Manheim 1975 [1933]: 60–63), gemeint sind Öffentlichkeiten, die bestimmte „Qualitäten" markieren (was durch Tradition, Macht oder beides umgesetzt werden kann), sowohl um Menschen und ihre Werte, Einstellungen und Meinungen zu inkludieren, als auch um solche gegebenenfalls gezielt zu exkludieren.

Der im Titel dieses Aufsatzes genannte Begriff „pluralistisch-transzendentale Öffentlichkeit" ist eine Zusammenführung von zwei weiteren Öffentlichkeitstypen, die Manheim konzeptualisierte, den der „pluralistischen Öffentlichkeit", die für die vielfältigen Meinungen in Demokratien steht und dabei strategische, auch antagonistische Kommunikation einschließt, und den der „transzendentalen" Öffentlichkeit, die Manheim folgend in Anlehnung an Kant verständigungs- und konsensuell orientierte Kommunikation beschreibt (Manheim 1975 [1933]: 49–60). Auf den Transzendenzbegriffs bei Habermas (1991) – der sich nicht auf Manheim bezieht, der dessen Buch aber nachweisbar kannte[3] (Habermas 1996 [1962]: 95–96, 371),[4] gehe ich im Folgenden ebenfalls ein. Dieser Transzendenzbegriff wird wichtig, um das typisch Humane der Kommunikationsethik zu präzisieren. Mit dieser Notwendigkeit konfrontieren uns neuerdings KI-Anwendungen, indem sie eben nur scheinbar *mit* uns „kommunizieren" (vgl. Beck 2024). Dass gerade KI mit ihren Anschlüssen an bestehende Verzerrungen und Stereotypisierungen, wie sie in online verfügbaren Inhalten ohnehin be-

3 Nicht eingehen kann ich als Nicht-Philosophin auf die Bezüge und Kritik beider Autoren, Manheim und Habermas an Kant, insbesondere dessen Transzendenzbegriff. Dies müsste im Vergleich zwischen Habermas und Manheim erst erarbeitet werden.

4 In der 1996er Auflage des „Strukturwandels der Öffentlichkeit" wird Manheim von Habermas falsch belegt: Manheims Buch wird seitens Habermas mit einem Publikationsdatum „1923" versehen, es ist jedoch 1933 erschienen, frühere Ausgaben von Manheims Monografie sind nicht nachweisbar. Manheim schrieb sein Buch, das als Habilitationsschrift geplant war ab 1932, wie er mir gegenüber in einem Interview 1995 gesagt hat.

stehen, Pluralität und Diversität eher entgegensteht und ethnozentrische Orientierungen aufweist, ist ein weiteres Problem.[5]

Das epistemologische Problem, dass universalistische Prinzipien (letztlich auch der universelle Anspruch diskursive Verfahren anzuwenden, wie wir ihn bei Habermas finden) *idealtypisch* sind, das heißt auf Regeln zielen, die (noch) nicht umgesetzt sind und reales Handeln als Abstraktum zwar anleiten, dieses Handeln und die entsprechende Kommunikation aber gleichwohl aufgrund von *Macht- und Ungleichheitskontexten* defizitär bleiben (vgl. Thomaß 2025: 57, 60), reproduziert auch der vorliegende Aufsatz. Allerdings weist gerade Manheim, der die konkrete Situativität öffentlicher Kommunikation begrifflich und empirisch fassbar machen wollte (vgl. Reitz 2005), kommunikationswissenschaftlichem Denken eine Perspektive, die einschließt, dass Akteur:innen universalistische Prinzipien gegebenenfalls nur behaupten oder vortäuschen, tatsächlich aber Partikularität – auch und gerade mittels Sprache–vernichten *wollen*. Diesen Gedankengang beinhaltet Manheims Konzept der „qualitativen Öffentlichkeit", die Pluralität per se unter Beobachtung stellt und potenziell bedroht und bekämpft (vgl. Averbeck-Lietz 2015: 101–149).

Ähnlich Manheim vor ihm bewegt sich Seyla Benhabib (2016: 55–56) explizit von der Abstraktion weg, wenn sie schreibt, dass ein kantianisch geprägter „Begründungsuniversalismus" (in dem sie Habermas verortet)[6] in der situativ gelagerten Realität untrennbar verbunden ist mit moralisch *divers und different* orientierten Individuen, also der Anerkennung der „Anderen" als vollständige, gleichwertige, aber gegebenenfalls in vielfacher Hinsicht Unterschiedliche, die in dieser Alterität zu respektieren sind (in diesem Sinne liest Redshaw 2020 Benhabib). Dies ist relevant, gerade um Universalismus für heutige Gesellschaften transkulturell oder auch kosmopolitisch[7] (vgl. Benhabib 2016) zu konzeptualisieren.

5 Die Problematik kann hier nur angesprochen, aber nicht ausgeführt werden kann. Vgl. weiterführend unter anderem Fenner 2025, Heesen 2024.

6 Allerdings vereinfacht diese Zuweisung Habermas auch, denn Habermas legt dar, dass sprachlich vorgebrachten „Gründen" ein universelles Prinzip zu eigen sei, da Gründe in allen Sprachen formulierbar sind und über Sprache für alle Kulturen zugänglich, seien sie „noch so verschieden" (Habermas 2024: 178). Habermas ist somit für unterschiedliche lebensweltliche Kontexte und Impulse offen.

7 Vgl. auch zu neueren Ansätzen „kosmopolitischer Kommunikationswissenschaft" Richter/Radue/Horz-Ishak et al. 2025.

An diese Überlegungen anschließend, geht die nachfolgende Argumentation in einem analytischen Dreischritt vor:

1. Zunächst lege ich dar, warum ein Universalismus der Würde (Benhabib 2016) holistischen, anti-pluralen Vorstellungen von „Volk“ wie sie populistische politische Strömungen scheinbar universalistisch reklamieren, etwas entgegenzusetzen hat. Mit Benhabibs Konzeption des „interaktiven Universalismus“ als kommunikativem „Aushandeln von Einheit und Vielfalt“ (Benhabib 2016: 68–69) sind universalistische *und* partikularistische Positionen adressierbar. Sie stehen sich dann gerade nicht diametral gegenüber (zur Opposition Universalismus/Partikularismus Thomaß 2025: 56);
2. versuche ich, den interaktiven Universalismus öffentlichkeits- und kommunikationstheoretisch sowie -ethisch zu vertiefen durch die Integration der Konzepte von Habermas und Manheim im Rekurs auf Benhabib, insbesondere, um illegitime, deviante Kommunikation und zugehörige Konzepte holistisch definierter „qualitativer“ Öffentlichkeit besser identifizierbar zu machen;
3. möchte ich auf dieser Basis die Rolle einer diskursiv verstandenen Kommunikationsethik als Antwort auf sowie als Prävention gegen holistische Vorstellungen vom „Volk“ oder der Einforderung einer vermeintlich ‚einheitlichen‘ (völkischen) öffentlichen Meinung betonen, wie Ernst Manheim sie für den aufstrebenden Nationalsozialismus Anfang der 1930er Jahre beobachtete.

Dieser Dreischritt ist analytisch gemeint und bezieht sich nicht auf eine Reihenfolge innerhalb der Struktur dieses Artikels, die folgenden Abschnitte greifen ihn auf verschiedene Weise auf und verschränken die Ansätze der Bezugsautoren Benhabib, Habermas und Manheim miteinander.

2. Universalismus der Würde und gegenseitige Anerkennung als Verschiedene

Universalismus ist gerade kein holistischer oder homogenisierender Begriff, denn er bezieht sich mit Benhabib (2016) auf die *gemeinsame Würde der Unterschiedlichen*. Dies gilt, obwohl „Universalismus“ eine hoch belastete Geschichte hat und als ‚Deckmantel‘ für Unterdrückung herhalten musste und muss (Thomaß 2025: 58–64).

Ich orientiere mich an dem Begriff des „moralischen Universalismus" Benhabibs (2016: 55–56), der von der Gleichwürdigkeit aller Menschen ausgeht, und wenn aller, *potenziell auch aller unterschiedlichen Individuen*:

> „Ich würde dieses Verständnis des Universalismus als das Prinzip definieren, wonach alle menschlichen Wesen, unabhängig von [,]Rasse[‘], Geschlecht, sexueller Orientierung, körperlichen oder geistigen Fähigkeiten sowie ethnischer, kultureller und religiöser Prägung das gleiche Recht auf moralischen Respekt haben" (Benhabib 2016: 56).

Dieser moralische Universalismus der gegenseitigen Anerkennung als Person ist zugleich ein „interactive universalism" (Benhabib 2007: 19, dazu Bracci 2002; Abbott 2022, 2024: 195–222). Die „difference" konkreter Anderer besteht immer in Relation zur "sameness" generalisierter Anderer (Bracci 2002: 476, 479; Redshaw 2020). Empirisch gesehen sind Individuen beides, konkrete und generalisierte Andere (vgl. Mead 1975 [1934], dazu Krallmann/Ziemann 2001: 2001–229). Moralität und ethische Reflexion entwickeln sich in der Spannung zwischen konkreter Person und ihrem sozialen Umfeld sowie generalisierter, auch stereotyp wahrgenommener oder vorgestellter Andersheit: „Moral judgement requires the enlarged thinking that is done in the context of other viewpoints" (Bracci 2002: 477).

Dass diese *viewpoints* gegebenenfalls nicht dialogisch oder diskursiv, sondern *faktisch durch Macht und die Erreichung von Hegemonie durchgesetzt werden,* bleibt unbenommen. Gleichwohl leitet strategische Macht(-kommunikation) kommuni-kationsethische Perspektivierungen nicht an, sondern konterkariert ihre *Geltung* (vgl. Habermas 1988 [1981]; Habermas 1998).

Benhabib setzt ihren Begriff „interactive universalism" in diese Spannung: Wir weisen uns – jedenfalls potenziell – als allgemeine Andere gegenseitig Würde zu, *weil* wir alle kommunizierende Menschen sind, aber das, *worum* es uns geht, wird kontextbezogen und konkret ausgehandelt. Das kulminiert – ob nun diskursiv oder strategisch – darin, dass wir moralische Urteile *zwangsläufig* im Kontext anderer, abweichender Sichtweisen fällen *müssen*.[8] Moralischer Universalismus entwickelt sich bei Benhabib *in* der Interaktion. Sich dabei *nur* als Verschiedene zu verstehen, würde lebensweltlich (vgl. Schütz/Luckmann 2003) und auch ethisch keine Sinn-

8 Vgl. auch die Beiträge in Niesen/Herborth 2007, die Habermas Theorie des Kommunikativen Handelns mit Blick auf politisch hochumkämpfte Machtfelder, die wie in der Außenpolitik von strategischer Kommunikation geprägt sind, diskutieren.

haftigkeit erzeugen, weil eine gemeinsame Basis nicht gedacht und somit auch nicht zu einem in der Alltagswelt relevanten Bezugspunkt werden könnte. Es wäre dann auch nicht unmoralisch andere, zum Beispiel Minderheiten auszugrenzen. Benhabib (2016) begründet aus einer kommunikationsethischen Position der kommunizierenden Unterschiedlichen mit gleicher Würde, dass Ausgrenzung ethisch gerade nicht begründbar ist und moralisch daher falsch. Sich allerdings als „Gleiche" ohne konkrete Andersheiten zu begreifen, bliebe abstrakt auf den nur „generalisierten Anderen" bezogen und kann Individuen und ihren verschiedenen *Identitäten* kaum Rechnung tragen (vgl. auch Redshaw 2020: 1). Personen sind als allgemeine Andere folglich gleichwertig und zugleich verschieden als konkrete/r Andere mit unterschiedlichen moralischen und anderen Präferenzen sowie ihren spezifisch situativ gelagerten Erfahrungen.

Anthropologisch lässt sich die Denkfigur der interaktiv *und* kooperativ handelnden Individuen aus Michael Tomasellos (paläo-)anthropologischen Schriften ableiten (Tomasello 2006, 2010), auf die Habermas (2024: 91) sich nach eigenen Worten „heute stützt", auch wenn dem eine kritische Debatte zwischen ihm und Tomasello zum Ursprung der Sprache vorausging (Vgl. Habermas 2009a). Interessanterweise beziehen sich alle drei, Benhabib (2016: 57), Tomasello (2006: 94) und Habermas (1988 [1981]: Bd. 2, 9–23) auf Meads Interaktionismus. Ausgehend von Mead (1975 [1934]), bei der generalisierten und konkreten Anderen keine normativen Konzepte sind, sondern empirische (auch epistemische), die Identitätsentwicklung im sozialen Umgang miteinander beschreiben, wird gerade bei Benhabib die *moralische Anerkennung des Anderen in seiner/ihrer Konkretheit* zur bestimmenden Größe, dies hat Sarah Redshaw herausgearbeitet: „In contrast to the ‚generalized other', Benhabib describes the ‚concrete other' as based on *acknowledgment* of difference" (Redshaw 2020: 1).

Die/der konkrete Andere ist nach Benhabib jede/r in seiner oder ihrer Identität, mit unterschiedlichen Erfahrungen und gegebenenfalls differenten moralischen Standpunkten zu spezifischen Themen und Fragestellungen (vgl. Benhabib 2016: 61–64). Jede:r konkrete Andere kann jede:n andere:n konkrete:n Andere:n anerkennen und genau dies kann als universelles Konzept verstanden werden, in dem zugleich Geschichtlichkeit, Identitätsbewusstsein und affektiv-emotionale Komponenten vermittelt werden.

Ernst Manheim sollte sich erst nach 1945, in seiner Phase als US-amerikanischer Soziologe und Anthropologe mit Meads „Mind, Self and Society"

befassen, es ist unwahrscheinlich, dass er die Schriften Meads in seinem akademischen Leben in Deutschland bis 1934 kennengelernt hat.[9]

Warum sind Denker:innen wie Habermas oder Benhabib sich sicher, dass Kommunikation als Modus der gegenseitigen Anerkennung möglich ist?[10] Die Begründung liegt darin, dass sie die grundlegende Fähigkeit zur Kommunikation nicht nur in verbaler Form, auch paraverbal sowie mittels Gestik und Mimik als menschliche Universalie verstehen, die dazu dient, Handlungen und Beziehungen zu leben und zu strukturieren.[11] Die anthropologische Fähigkeit zur Kommunikation liegt dabei weit vor der sozialisierten Fähigkeit zum Diskurs. Tomasello hat am Max Planck Institut für Paläoanthropologie in Leipzig nachgewiesen, was bis dahin die Sozialwissenschaftler:innen meist von Mead, also aus der sehr viel älteren Theoriebildung, entnahmen: *Kooperativ orientierte Interaktion* (und sei es nach Tomasello in der einfachen Variante vorsprachlich über Zeigegesten) ist die Voraussetzung auf einem höheren Niveau zu kommunizieren und legt nahe, dass wir qua Geburt zur Rollenübernahme fähige, gesellige Wesen sind (vgl. Tomasello 2006, 2009, 2010). Als interaktionsfähige Wesen versuchen wir, das hatte Mead herausgearbeitet, Situationen und Rollen anderer Menschen permanent einzuordnen und zu verstehen und darauf wiederum kommunikativ zu antworten und Antwort auch zu erwarten (vgl. Krallmann/Ziemann 2001: 201–229). Benhabib bezieht sich genau in diesem Sinne auf Mead: Der allgemeine Andere ist jeder von uns in seinem Menschsein. Das ist auch eine moralische Feststellung, denn sie sagt aus: Wir sind uns alle ähnlich. Der hier zugrunde liegende Universalismus der Würde ist nicht hintergehbar und greift weiter als ein nur prozeduraler Universalismus, der die Diskursethik als rein formale Ethik universalisiert

9 Dies geht aus dem Aktenkonvolut mit Manheims Vorlesungsskripten, Bibliografien und Referatethemen für Studierende in den von Manheim so benannten Feldern „History of Sociology" und „Symbolic Interactionism" hervor. Vgl. Archiv der Universität Kansas City, Missouri Signatur 31, 2.8.-2.9, eigene in Kansas City verfertigte Abschriften der Autorin 1997. Zu dieser Zeit war der Bestand noch unsortiert, siehe heute: Ernst Manheim papers, MS 373, Kenneth Spencer Research Library, University of Kansas (online).

10 Axel Honneth und gegebenenfalls weitere Denker:innen wären hier im Falle weiterer Forschung hinzuzunehmen.

11 Auch zur Kommunikation zwischen Tieren liegt Forschung vor, die interaktive Kommunikation nicht mehr nur als Fähigkeit von homo sapiens einordnet. Darauf kann hier nicht weiter eingegangen werden (vgl. weiterführend Rutz/Bronstein et al. 2023). Die Fähigkeit zur Metakommunikation, nämlich im Sinne von Habermas über Kommunikation zu kommunizieren (und sie zu erforschen), scheint ausschließlich Menschen zuzukommen.

(vgl. Brosda 2010). Dies auch deshalb, da Benhabib nicht nur auf Diskurse, sondern darüber hinaus auf das Netzwerk von Relationen und Narrationen, in das Menschen eingebunden sind und das sie produzieren, eingeht (vgl. Abbott 2024: 201). Dies ist meines Erachtens durchaus kompatibel mit Habermas' an Schütz anschließender Konzeption der „Lebenswelt“ und ihrer Verankerung „moralischer Gefühle“, einschließlich solcher für die gegenseitige „Verpflichtung“ (vgl. Habermas 2009b: 36).

Denkt man dies für die Alltagskommunikation pluraler Gesellschaften weiter, so können Menschen folglich *in gegenseitiger Anerkennung ihrer konkreten Verschiedenheit* und ihres allgemein Menschlichen einander Ähnlich-Seins kontrovers kommunizieren, streiten und sich empören und stellen dabei Öffentlichkeit als ein plurales, vielstimmiges Netzwerk von Stellungnahmen und Meinungen her (vgl. Habermas 1998: 436; weiterführend Wessler 2018). Legitimität entsteht so grundlegend durch gegenseitige Achtung (vgl. Rühl/Saxer 1981 aus systemtheoretischer Perspektive und sich dabei ebenfalls auf Mead beziehend). Aus verständigungsorientierter Perspektive (Habermas 1988 [1981]) sprechen sich Menschen in ihrer lebensweltlichen Kommunikation und Interaktion gegenseitig *Geltung* (Habermas 1988 [1981], Bd. 1: 410–411, 439) zu, auch dann, wenn sie zumal in der öffentlichen Kommunikation nicht in der Lage sind, sie zu erfüllen: „Zustimmung zu Themen und Beiträgen bildet sich erst als Resultat einer mehr oder weniger erschöpfenden Kontroverse, in der Vorschläge, Informationen und Gründe mehr oder weniger rational verarbeitet werden“ (Habermas 1998: 438).

Diese Rationalität entfaltet sich als „kommunikative Rationalität“ (ebd.: 28), unterscheidbar von strategischer Zweckrationalität oder „kognitiver instrumenteller Rationalität“ (ebd. sowie S. 446), eben über die wechselseitige Anerkennung von Geltungsansprüchen:

3. Die Geltungsansprüche nach Habermas

Habermas konzeptualisiert für das „verständigungsorientierte Handeln“ und einzig für dieses, nicht für das am Geltungsanspruch „Wirksamkeit“ orientierte „strategische Handeln“ *Wahrheits-, Richtigkeits- und Wahrhaftigkeitsansprüche*. Hinzu kommen Verständlichkeitsansprüche (vgl. Habermas 1988 [1981], Bd. 1: 45, 439). Wir alle leben in „drei Welten“ (Habermas 1988 [1981], Bd.1: 115–137, in Auseinandersetzung unter anderem mit Popper

ebd.: 148–149), auf die sich diese Ansprüche beziehen, gleichzeitig: In der „objektiven Welt" versuchen wir Fakten nahe zu kommen (zum Beispiel durch Wissenschaft, generell durch Belege und Begründungen), in der „sozialen Welt" versuchen wir soziale Beziehungen zu leben und Normen des Respekts einzuhalten, dabei möglichst wahrhaftig zu sein und erwarten das auch von anderen. Außerdem wollen wir verstehen können, was andere sagen. Von Bedürfnissen geleitete Gefühle der „subjektiven Welt" begleiten diese Prozesse (ebd.: 138, 149–149). Zentral ist meines Erachtens und dies kann Überlegungen zur gegenseitigen Anerkennung erweitern: Wir haben als kommunizierende Menschen *Anspruch auf Geltung* unterhalb der juristischen Ebene, auf einer moralischen. Ich lehne mich hier an Hannah Arendts Diktum (aufgegriffen von Benhabib 2016: 64, 75) an, das besagt Menschen haben „das Recht, Rechte zu haben". Benhabib leitet daraus ein gegenseitiges moralisches, kein nur juristisches Anerkennungsprinzip ab (ebd.: 62–63): Rechte haben Menschen nach Benhabib, *weil* diese auf moralischen Ansprüchen *aufbauen* (ebd.: 105), nicht weil sie formaljuristisch gesetzt sind. Aus einer Perspektive der Menschenwürde baut Recht auf Moralität auf, nicht umgekehrt Moralität erst auf Recht.[12] Somit haben Menschen Anspruch darauf, ihre Geltungsansprüche (im Sinne von Habermas) in die Kommunikation einzubringen; grundsätzlich alle jederzeit und in ihrer Verschiedenheit, auch wenn dies faktisch nicht erfüllt ist und durch Macht, Abwertung und Nicht-Gelten-Lassen überlagert sein kann. Spätestens bei verletzten Geltungsansprüchen – damit verletzter Würde und es reicht, dass dies subjektiv so gefühlt wird – kommen *Emotionen* ins Spiel, auch Empörung. Wir werden angeschrien und fühlen uns jenseits des Inhalts verletzt. Anschreien mag wahrhaftig sein und sich für die wütende Person passend anfühlen, aber ist sicher in der sozialen Welt nicht richtig. Auch dann nicht, wenn die Fakten stimmen. Die Tatsache stimmt, die Abwertung der Person ist *nicht* richtig. Allein dass man bemerkt, dass Geltungsansprüche *nicht* erfüllt sind, hat Auswirkungen auf unser Miteinander. Auf der Basis der Geltungsansprüche kann Kommunikation als misslungen kritisiert oder gekennzeichnet werden, dass Menschen exkludiert werden. Habermas argumentiert bewusst kontrafaktisch. Seine Kommunikationsethik lässt sich deshalb machtkritisch verwenden (vgl. Peters 2007: 67; Averbeck-Lietz 2015: 143).

Habermas geht – das wird oft übersehen – über kognitive Rationalität hinaus. Menschen leben innerhalb von „Gefühls- und Vorstellungswelten"

12 Benhabib (2016: 105) argumentiert hier gegen Martha Nussbaum und Amartya Sen.

(Habermas 2024: 94). Die „Fähigkeit zum moralischen Urteilen" als Möglichkeit des Erkennens von Unrecht (gegen andere Menschen) vollzieht sich über Empathie und übersteigt ein formales Procedere. Daran lässt der späte Habermas keinen Zweifel, aber auch nicht daran, dass es *sinnvoll* ist in formal organisierte Diskurse einzutreten (vgl. Habermas 2024: 87).

Emotionen ordnet Habermas den „praktischen Diskursen" über moralische Fragen zu, sie haben rechtfertigenden Charakter und stoßen gesellschaftliche Debatten an (vgl. Wessler 2018: 148–151; Habermas 1988 [1981], Bd. 1: 45). Allerdings kann und soll in der weiteren diskursiven Bearbeitung gegebenenfalls auch wieder von Emotionen Abstand genommen werden: Im „theoretischen Diskurs" um Fragen der Wahrheit, in dem rationale Argumente verwendet werden sollen (Habermas 1988 [1981], Bd. 1: 45), sind Menschen aufgefordert von ihren subjektiven Emotionen zurückzutreten (von Wut, Enttäuschung, auch von Freude). Die emotionale Distanzierung, auch von sich selbst, ist nötig, um die – möglicherweise gegenläufigen – Emotionen und Meinungen anderer Menschen anerkennen zu können, auch und gerade, wenn man sie nicht teilt. Das mag nicht immer vollständig gelingen, ist aber relevant, um überhaupt zu wechselseitigen Argumentationen, die Pluralität zulassen, vorzudringen. In der Konsequenz nicht sehr weit davon entfernt heißt es bei Benhabib: Empathie im Sinne von „Anteilnahme" und „Mitgefühl" solle sich nicht nur auf die eigene Gruppe beziehen. Erst dann können gegenseitige Anerkennungsansprüche erfüllt werden (vgl. Benhabib 1995: 196). Das ist rationalisierbar, eben als Respekt oder Achtung gegenüber anders Meinenden, Denkenden, Fühlenden oder solchen, die differente Sinnbezüge aufgrund anderer Erfahrungen herstellen. Ein Schritt darüber hinaus wäre der Versuch, gemeinsam doch zu einem Dialog (wenn schon nicht zu einem Diskurs) zu kommen, gegebenenfalls nur in Teilen: Kompromiss statt Übereinstimmung. Auch die Publizistin und Habermas-Schülerin Carolin Emcke erläutert Gefahren für den gesellschaftlichen Zusammenhalt durch gefühlsgeleitete Empathie, sobald diese holistisch auf die *Eigengruppe* bezogen wird, weil sie dann Fremdgruppen gegebenenfalls nicht mehr als gleichwertige Partikularitäten, nämlich als „Teil eines universalen Wir" anerkennt. Emcke bezieht sich auf gewalttätige und gewalttolerierende „Wutbürger", die 2015 in Clausnitz einen Bus mit Geflüchteten massiv bedrängt und blockiert sowie die Insass:innen, darunter Kinder, bedroht und verängstigt haben (Emcke 2019: 54).

Die Geltungsansprüche ermöglichen uns nicht nur situative Einordnungen und empirische Analysen, wann sie wie von wem erhoben werden

(vgl. Venema/Averbeck-Lietz 2017), sie betreffen auch die „Reflexions- und die Steuerungsfunktion" der Kommunikationsethik (Debatin 2002: 262). Wir geben auf der Basis von Reflektion Kommunikationsprozessen steuernd eine dialogische Richtung. Jenseits der Geltungsansprüche werden die Meinungen und Einstellungen zu Themen unterschiedlich bleiben, sie „konkurrieren" (Habermas 2022: 25), aber die Geltungsansprüche sind eine Grundlage, Kommunikation *trotzdem* vollziehen zu können. Das macht ihre reflexive, reagierende und zugleich ihre präventive Funktion aus.

In Fällen von Hate Speech, Dangerous Speech, aggressiver völkischer Parolen und tätlicher Angriffe greift potenziell das Recht. Das ändert an der kommunikationsethischen Aufgabe von Reflektion und Steuerung nichts, denn sie liegt auf einer anderen Ebene als das Recht. Wenn man mit Benhabib moralischen Universalismus, kulminierend in Würde, im gemeinsamen Menschsein, das per se kommunikativ entwickelt wird, annimmt, bleibt nichts anderes übrig, als neben dem Einsatz von Rechtsmitteln langfristig mit dem Ziel des Einbeziehens möglichst vieler Beteiligter weiter zu kommunizieren. Zugleich sind Kommunikationsstrategien zwecks Zersetzung von Geltungsansprüchen zu entlarven und zu analysieren.

Wie können wir es erkennen, wenn öffentlichkeitsrelevante Geltungsansprüche bedroht werden? Die Geltungsansprüche und deren Einforderung und Erhalt sind grundlegend für die gegenseitige Anerkennung in Verschiedenheit. Allerdings müssen ergänzend die strukturellen Bedingungen der Öffentlichkeit betrachtet werden, unter denen das geschehen kann, möglich ist und bleibt, gerade um Öffentlichkeit als „Raum der Gründe" (Habermas 2024: 108) schützen zu können.

Um diesen Denkschritt gehen zu können, greife ich auf Manheims Öffentlichkeitskonzeption zurück:

a. Die Interdependenz verständigungsorientierter und strategischer Formen von Kommunikation auslotet und
b. Die Grenzen zwischen Pluralismus als Normalfall der Demokratie und „radikaler" Polarisierung als die Demokratie gegebenenfalls destabilisierend erkundet.

4. Pluralistische, transzendentale, qualitative Öffentlichkeiten und ihre Kommunikationsformen nach Manheim

Ernst Manheim führte 1933 einen dreifachen Öffentlichkeitsbegriff ein. Ihm folgend kommt es auf drei idealtypische, gegebenenfalls realtypisch mit-

einander verwobene Ausprägungen von Öffentlichkeit an: „pluralistische Öffentlichkeit" (Gruppeninteressen, die teils konfliktär sind), „transzendentale Öffentlichkeit" (verständigungsorientierte, rationale Argumente zielend auf Konsens) und „qualitative Öffentlichkeit" (thematische und normative Schließungen qua Tradition oder qua Macht (ausführlich Averbeck-Lietz 2015: 125–131; Beetz 2005: 156–157). Die qualitative Öffentlichkeit erzeugt Kommunikationstabus und sanktioniert sie, hier herrscht die holistische Repräsentation, nicht die Diskussion. Auch das Motiv finden wir bei Habermas in dem, was er über vermachtete und repräsentative Öffentlichkeit schreibt (unter anderem Habermas 1996 [1962]: 58–67).

Manheim folgend tarieren die Öffentlichkeitsformen sich gegenseitig aus, kippen aber in nicht mehr legitimierbare anti-demokratische und/oder totalitäre Tendenzen, wenn qualitative Elemente strukturbildend werden und plurale und transzendente, also diskursive Anteile, mittels Macht und gegebenenfalls Gewalt delegitimieren. Dann ist pluralistische Öffentlichkeit gefährdet oder wird vernichtet (vgl. Manheim 1939; 1964; 1976 [1933]).

Die „Feinde" der „offenen Gesellschaft" (Popper 1945) sind dann die, die ihre „qualitativen" Setzungen durchdrücken wollen und können – gegen den Diskurs und die in ihm vermittelten Geltungsansprüche, schlimmstenfalls auch gegen die (transzendente) *Idee eines Diskurses*. Manheim beobachtete dies im aufstrebenden Nationalsozialismus (vgl. Manheim 1939; Averbeck/Manheim 1999). In Manheims Werk finden wir meines Erachtens eine der relevantesten Fragen, die sich mit Blick auf Öffentlichkeit stellt: Wann wird Polemik strukturell? Wann kippen Antagonismen in Ideologien, die die ersten Schritte zu totalitären Formen der Kommunikation sein können und Öffentlichkeit als Raum der Gründe zerstören? Bei Manheim liegt der Mechanismus nicht nur in der Sprache, sondern auch in der Kommunikations*situation*, einschließlich der systemischen Frage nach dem Verhältnis von Achtung/Missachtung und Macht, schließlich Gewalt. Die Öffentlichkeitsvorstellung der NS-Zeitungswissenschaft spiegelte Manheims Gedanken verzerrt ins Negative wider: In einer sogenannten „Volksgemeinschaft" seien unterschiedliche Meinungen und Ansprüche von Minderheiten nicht mehr bedeutsam (vgl. Averbeck 1999: 134–142 in Auswertung von Schriften der NS-Zeitungswissenschaft). Der „Volkskörper" habe – holistisch – das gleiche zu meinen und zu wollen. Meinungspluralität wurde von NS-Wissenschaftlern wie Karl Kurth, Direkter des Wiener Instituts für Zeitungswissenschaft oder Hans Amandus Münster, Direktor des Leipziger Instituts für Zeitungswesen, für obsolet erklärt und aus ihren professoralen, mit dem NS-System paktierenden Machpositionen heraus bekämpft

(vgl. ebd.). Bei Manheim ist der Begriff des *Wollens* anders bestimmt: Es geht um das Wollen oder den Willen, in Diskurse einzutreten (siehe unten), nicht um bestimmte Inhalte des Wollens, die vorab gesetzt werden. Die Grenze eines kommunikativen Pluralismus liegt folglich da, wo Menschen versuchen, das Plurale willentlich zu zerstören, auch mittels gezielter kommunikativer Polarisierung (dazu aus heutiger Perspektive Holtz-Bacha 2019).

Manheim war Schüler von Ferdinand Tönnies, in seiner Konzeption verbirgt sich auch dessen Konzept des *Kürwillens,* der reflektierten Entscheidung zu einem bestimmten Handeln (vgl. Averbeck-Lietz 2015: 65–69). Manheims Konzept von Öffentlichkeit ist eines, in dem die „Träger der öffentlichen Meinung" (so der Untertitel seines Buches zur „Soziologie der Öffentlichkeit" von 1933), ob Publizist:innen, Poliker:innen oder auch die Massenpresse, sich freiwillig entscheiden, an der öffentlichen Diskussion teilzunehmen und diese dabei kommunikativ absichern und legitimieren: „In der öffentlichen Diskussion erweitert sich der Raum der Gruppenidentifikation zum transzendentalen Gesamtraum der Öffentlichkeit. Man diskutiert aufgrund verschiedener Meinungen, aber gleichen Wollens." (Manheim 1979 [1933]: 53)

Konsens ist dabei in einem ersten Schritt nicht inhaltlich gemeint, sondern liegt darin, überhaupt in eine verständigungsorientierte kommunikative Beziehung einzutreten, beziehungsweise *eintreten zu wollen:*[13] „Der Ausdruck ‚Konsensus' bezeichnet eine verbale, willensmäßige Form der Identifikation, in deren Zeichen Subjekt und Adressat in die kommunikative Beziehung […] eintreten" (Manheim 1979 [1933]: 31).

Erst dann, wenn der Idee eines konsensorientierten Diskurses willentlich zugestimmt werden kann, wird Pluralität möglich: „[…] es sollen alle Ausgangsmeinungen, alle Standpunkte in sie [die sinnvolle Unterredung] einbezogen werden" (Manheim 1979 [1933]: 53).

Diese Ausgangsmeinungen können höchst unterschiedlich sein. Was diejenigen *eint*, die sie vertreten, ist, dass sie sie in einen Diskurs einbringen.

13 Das menschliche Wollen hat auch den späten Habermas beschäftigt. Im Interview mit Müller-Doohm und Yos geht es auch darum, warum man Diskurse wollen sollte (Habermas 2924: 148), warum sie sinnvoll für ein soziales Miteinander und eine „Beteiligungsperspektive" sind (ebd.: 12).

Auch Habermas (2022: 25) spricht demokratischer Öffentlichkeit ihren widerstreitenden, „agonalen Charakter"[14] gerade nicht ab:

> „Nur über [...] die Ermutigung zum reziproken Neinsagen entfaltet sich das epistemische Potential der widerstreitenden Meinungen im Diskurs, denn dieser ist auf die Selbstkorrektur von Teilnehmern angelegt, die ohne gegenseitige Kritik nicht voneinander lernen könnten. [...] Vor diesem konsentierten Hintergrund besteht der gesamte demokratische Prozess aus einer Flut von Dissensen" (Habermas 2022: 25).

Das heißt auch: Der Diskurs geht nicht völlig in verständigungsorientierter Kommunikation auf (vgl. Günther 2009: 305).

Manheim folgend kann die potentielle Polarisierung zu der pluralistische Öffentlichkeiten mit ihrem „polemischen" (Manheim 1976 [1933]: 57) Kommuni-kationsstil neigen, nur durch den transzendentalen Typus, also Konsensorientierung abgemildert werden, ansonsten kommt es zum Zerfall der Öffentlichkeit, dem „radikalen Grenzfall" (siehe unten), gut beobachtbar aus Manheims Perspektive am Ende der Weimarer Republik mit ihrer politischen Publizistik, den Aufmärschen und Straßenkämpfen (vgl. Averbeck 2005; Koenen/Sax 2025). „Pluralistisch wird das Gefüge dieser Öffentlichkeit grundsätzlich dadurch, dass in ihr im radikalen Grenzfall keine affirmativen Willensgehalte vorhanden sind [...]. Die Öffentlichkeit [...] beruht [...] auf diesen polaren Scheidungen" (Manheim 1979 [1933]: 54).

Als US-Soziologe transferierte Manheim, der 1934 aus Deutschland fliehen musste (zu seiner Biografie Welzig 1997) seine Typenbildung (vgl. Manheim 1964) ins Englische, der pluralistische Typ wird zum „liberalen". Der qualitative heißt nun „authoritarian" oder „illiberal", der zu „imperativer" Kommunikation neige. Der dominante Kommunikationstyp ist nach Manheim polemisch im pluralistischen Umfeld, imperativ im qualitativen Umfeld und es muss je gefragt werden, wer damit eigentlich von welchen Kommunikatoren wie erreicht und adressiert werden soll; was also *strategisch* bewirkt werden soll (vgl. Manheim 1976 [1933]: 56]. Einzig die transzendentale Form – die sich mit anderen mischen und verbinden kann – ist nicht strategisch, sondern verständigungsorientiert.

14 Agonale Öffentlichkeitstheorien wie die von Chantal Mouffe bleiben in diesem Aufsatz unberücksichtigt, eine solche Berücksichtigung würde die gegenseitige Rezeption von Mouffe und Habermas sowie deren Abgrenzung voneinander einzuschließen haben.

Anders als Habermas schließt Manheim den empirischen Normalfall strategische Kommunikation elementar in sein Analyseraster ein. Schon in den 1920er Jahren wollte er Soziologie als „Wirklichkeitswissenschaft" betreiben.[15] Im Londoner Exil erkannte er im Rahmen seiner zweiten Dissertationsschrift, die er bei seinem Cousin Karl Mannheim und Bronislaw Malinowski verfasste, noch etwas anderes, dass nämlich Gesellschaften Non-Konformität – also nicht nur Pluralität, sondern auch Abweichung – belohnen müssen, sollen sie demokratisch stabil sein: „There is, however, in every society in different degrees at different points scope for public dissension. Legitimate and coordinated non-conformity in public is a stabilizing force just as submission to social norms" (Manheim 1937: 7). Die "qualitative Öffentlichkeit", die Manheim 1933 beobachtete, ließ keine Nonkonformität mehr zu.

Habermas konzeptualisiert wie schon Manheim den Transzendenzbegriff innerweltlich, also nachmetaphysisch (Vgl. Müller-Doohm 2020). Er spricht von der „Transzendenz von Innen" oder „transzendierenden Geltungsansprüchen" (Habermas 1991: 155; 2024: 36, 113–117). Sie kulminieren im „Universalitätsanspruch der Wahrheit" und stellen sich in „lebensweltlich verfasster kommunikativer Alltagspraxis" her (ebd.: 146), einschließlich „Symbolen und Bildern, Indexen und Ausdrucksgesten" (ebd. Habermas im Anschluss an Peirce sowie Habermas 2024: 116–117).

Transzendenz vollzieht sich durch Kommunikation als Möglichkeit und Notwendigkeit gegenseitiger Verständigung. Manheim und Habermas sind hier kompatibel, auch wenn die Theorie des kommunikativen Handelns, gerade im Anschluss von Habermas an Peirce und die Sozial- und Sprachpragmatik, viel weitreichender für eine allgemeine Kommunikationstheorie ist als Manheims Öffentlichkeitstheorie. Indes hat Habermas die lebensweltlich verfasste Alltagspraxis von „Symbolen, Bildern, Indexen und Ausdrucksgesten" (siehe oben) nicht selbst erforscht, Manheim ebenso wenig. Vieles bleibt sowohl bei Manheim als auch bei Habermas offen. Wichtig aber ist: Über welche Zeichenform auch immer, Menschen sind potenziell in der Lage sich *in gegenseitiger Anerkennung zu verständigen*, trotz partiku-

15 Manheim referierte auf einen weiteren Lehrer: Hans Freyer (der ein Freund wurde bis zum Tode Freyers, trotz dessen zeitweiliger Huldigung des NS-Wissenschaftssystem). Manheim reklamierte in einem Interview mit der Autorin 1995 die Idee in Anlehnung an Max Weber „konkrete" Soziologe als „Wirklichkeitswissenschaft" zu betreiben für sich „und Freyer nahm's an" (Averbeck/Manheim 1999). Manheim hatte zuvor bei Theodor Litt und Freyer als zweitem Gutachter mit einer Dissertation zur „Logik des konkreten Begriffs" (Manheim 1930) promoviert.

larer Interessen und Machtgefälle. Die Auseinandersetzung mit Machtgefällen muss wiederum kommunikativ bearbeitet werden.

5. Schlussfolgerung

Ulrich Saxer (1996: 155) hat vor 30 Jahren festgestellt, es handele sich mit Blick auf „Letztwerte" wie Menschenwürde und die daraus ableitbare Meinungsäußerungsfreiheit in der deutschen und anderen westlichen Gesellschaften um weitgehend „konsentierte" Werte. Daran dürfen Zweifel angebracht werden, umso mehr heute angesichts der Afd mit hohen Wahlerfolgen in Deutschland (und vergleichbarer Parteien anderswo), die genau dahingehend, in ihrer Haltung zur Menschenwürde, derzeit der Verfassungsschutz beobachtet.

Ein normatives Ziel, das sich aus diesem Aufsatz ableiten lässt, ist die gegenseitige Achtung (Universalität der Würde) bei Unterschiedlichkeit der Standpunkte und Gruppenidentifikationen (Partikularität, Pluralität) in einer Gesellschaft. Wie kann dies vermittelt werden? Aus einer Habermasianischen Position müsste dieses Postulat immer wieder in gesellschaftliche Debatten eingehen und auf verschiedenen, auch institutionellen Ebenen thematisiert und besprochen werden. „Die Anerkennung von Pluralismus ist anstrengend" formuliert Mandry (2012: 236). Aus kommunikationswissenschaftlicher Perspektive können wir zu dieser Anstrengung beitragen mit Blick auf die langfristige Entwicklung kommunikativer Kompetenzen nicht nur auf der Individual-, sondern auch auf der Mesoebene spezifischer Organisationen und ihrer Institutionen, wie den Schulen, den Universitäten und Stätten der Erwachsenen- und der politischen Bildung.[16] Schlussfolgernd aus dem vorliegenden Aufsatz möchte ich verschiedene Orientierungen einer solchen Bildung benennen:

- Sensibilisierung für die gegenseitige Achtung (Universalität der Würde) bei Unterschiedlichkeit der Standpunkte und Gruppenidentifikationen (Partikularität, Pluralität)
- Sensibilisierung für die Verletzung und Verletzbarkeit von Geltungsansprüchen (eigener und anderer)
- Sensibilisierung für die strukturellen u. systemischen Bedingungen in (medienvermittelten) Öffentlichkeiten

16 Manheim war als Doktorand in Leipzig und später als Professor in Kansas City auch in der außerakademischen Erwachsenenbildung tätig.

- Sensibilisierung für das Pluralitätspostulat von Öffentlichkeiten, an dem heute die Nutzer:innen als Produser:innen selbst teilhaben.
- Kommunikationsethische Bildung und Vermittlung von Wissen aus der Kommunikationswissenschaft bezüglich interpersonaler Kommunikation, massenmedialer, digitaler Kommunikation, plattformisierter Kommunikation und KI
- Sensibilisierung für den Zusammenhang zwischen der Vermittlung von Kommunikations- und Medienkompetenz sowie politischer Bildung.

Diese Sensibilisierungen sind relevant, weil das Anerkennen von Geltungsansprüchen in Interessenskonflikten, die typisch sind für Demokratien, nicht per se erfolgt. Der transzendentale Aspekt verständigungsorientierter Kommunikation, die grundsätzliche Bereitschaft zur Konsensbildung, stellt sich nicht von allein her, sie muss institutionell abgesichert werden (vgl. Habermas 1998), im Sinne Manheims *gewollt* werden auch über Regelbildungen in Organisationen wie zum Beispiel Parlamenten.

Ein weiterer Aspekt, der die Verständigungsfähigkeit von Menschen und damit auch die akademische Kommunikationsethik herausfordert, ist die Kommunikation mittels und mit KI. Es gilt den grundsätzlichen *epistemologischen Unterschied* zwischen Humankommunikation und ihrer Möglichkeit zu verständigungsorientierter Kommunikation auf der Basis von Geltungsansprüchen und von KI und Large Language Models (LLM) als selbstlernenden neuronalen Netzen voneinander abzugrenzen. LLM kommunizieren nicht verständigungsorientiert oder basierend auf einer „reziprok eingenommenen Zweipersoneneinstellung" (Habermas 2024: 120–121 in Anlehnung an Schütz). Sie haben keinen lebensweltlichen Alltag, sie sind keine menschlichen Interaktionswesen. Auch wenn sich KI-Anwendungen noch so ‚echt' anhören und die „Kommunikation" mit ihnen sprachlich erfolgt: Roboter oder KI sind im Sinne Bruno Latours „Aktanten" (Beck 2024: 532). Sie sprechen nicht, sie können Intersubjektivität allenfalls über Sprachmodelle *imitieren*, aber die „Transzendenz von Innen" im Sinne von Habermas erfüllen sie nicht. Denn diese meint mehr als die Sprache als ‚Tool', sie meint die umfassende, körperliche, zeit-räumliche Bezogenheit der Menschen aufeinander. Nur Menschen vollziehen Sinngebung durch wechselseitige *Selbst*kundgabe und Meta-Kommunikation im Sinne des *Sprechens über Sprechen* und der gegenseitigen Erwartens-Erwartungen, die wiederum reflektiert werden können und in neue Kommunikationen eingehen. Roboter sind hier nicht eingeschlossen (vgl. auch Beck 2023, 2024). Sie imitieren und erfinden, ‚halluzinieren' gar, aber sie *leben* nicht.

Sie können funktional in Sinne einer Assistenz mit uns die Rollen tauschen oder uns als Assistenz unterstützen, sie werden gleichwohl nicht zu Menschen, die körperlich in Zeit und Raum verankert sind und ein Bewusstsein ihres eigenen, geistigen und körperlichen Todes haben. Menschen ihrerseits können die Rolle gerade nicht mit ihren KI-Assistenten oder anderen algorithmisch programmierten Systemen tauschen, ihre langsameren, erratischeren, da auch emotional, durch Erinnerungen und Zukunftserwartungen getriebenen kognitiven Prozesse lassen das nicht zu. Aufgaben, die wir einer KI geben, wie Texterstellung oder Recherchen *in kürzester Zeit*, können Menschen nicht erfüllen. Zeitlichkeit und Räumlichkeit als zentrale Komponenten von Kommunikationsprozessen (vgl. Merten 1977) unterscheiden die Prozesse der KI von Humankommunikation.

Habermas hat sich in der „Theorie des Kommunikativen Handelns" 1981 zwar nicht mit künstlicher Intelligenz befasst, aber mit „operativ erzeugten Gebilden", die zwar auf rationalen Entscheidungsmechanismen beruhen, aber keine „selbstgenügsamen Existenzen" sind, die *aus sich selbst heraus* Geltungsansprüche in Anschlag bringen könnten:

> „Operativ erzeugte Gebilde können, für sich betrachtet als mehr oder weniger korrekt, regelkonform oder wohlgeformt beurteilt werden; sie sind aber nicht wie Handlungen einer Kritik unter Gesichtspunkten der Wahrheit, Wirksamkeit, Richtigkeit oder Wahrhaftigkeit zugänglich; denn sie gewinnen nur als Infrastruktur anderer Handlungen einen Bezug zur Welt" (Habermas 1988 [1981], Bd. 1: 147).

Geltungsansprüche werden *von Menschen reziprok erhoben*, die sie intersubjektiv anerkennen können, aber (jedenfalls bisher) nicht von LLM. Diese sind allenfalls strategisch an „Wirksamkeit", sprich Effizienz orientiert und damit an einem normativ ‚leeren' Geltungsanspruch (vgl. Habermas 1988 [1981], Bd. 1: 439). Wirksamkeit als solche kann keine Legitimität kommunikativer Rationalität beanspruchen.

Literatur

Abbott, Owen (2022): Interactive Universalism, the Concrete Other and Discourse Ethics: A Sociological Dialogue with Seyla Benhabib's Theories of Morality, in: European Journal of Social Theory 26 (3/2022), S. 335–353.

Abbott, Owen (2024): Social Theorists of Morality. Essays on Moral Agency, London.

Averbeck, Stefanie (1999): Kommunikation als Prozess. Soziologische Perspektiven in der Zeitungswissenschaft 1927–1935, Münster, London.

Averbeck, Stefanie / Manheim, Ernst (1999): Gespräche mit Ernst Manheim (geb. 1900), jüdischer Emigrant aus Deutschland und amerikanischer Soziologe. Aufgezeichnet im August 1995 in Martha's Vineyard/Massachussetts, in: Jahrbuch für Soziologiegeschichte 1995, S. 53–86.

Averbeck, Stefanie (2005): Ernst Manheims „Träger der öffentlichen Meinung": Eine Theorie der Öffentlichkeit 30 Jahre vor Jürgen Habermas, in: Frank Baron / David N. Smith / Charles Reitz (Hg.), Authority, Culture and Communication. The Sociology of Ernest Manheim, Heidelberg, S. 43–70.

Averbeck-Lietz, Stefanie (2015): Soziologie der Kommunikation. Die Mediatisierung der Gesellschaft und die Theoriebildung der Klassiker, München.

Beck, Klaus (2021): Kommunikationsfreiheit, Wiesbaden.

Beck, Klaus (2023): Digitale Kommunikation, in: Christian Neuhäuser / Marie-Luise Raters / Ralf Stoecker (Hg.), Handbuch Angewandte Ethik. 2. Aufl, Wiesbaden, S. 903–908.

Beck, Klaus (2024): Kommunikation, in: Petra Grimm / Kai Erik Trost / Oliver Zöllner (Hg.), Digitale Ethik, Baden-Baden, S. 529–539

Beetz, Michael (2005): Die Rationalität der Öffentlichkeit, Konstanz.

Benhabib, Seyla (1995): Selbst im Kontext. Gender Studies, Frankfurt am Main

Benhabib, Seyla (2007): Another Universalism: On the Unity and Diversity of Human Rights, in: Proceedings and Addresses of the American Philosophical Association 81 (2/2007), S. 7–32.

Benhabib, Seyla (2016): Kosmopolitismus ohne Illusionen. Menschenrechte in unruhigen Zeiten, Frankfurt am Main.

Bracci, Sharon L. (2002): Seyla Benhabib's Interactive Universalism: Fragile Hope for Radically Democratic Conversational Model, in: Qualitative Inquiry 8 (4/2002), S. 463–488.

Brosda, Carsten (2010): Diskursethik, in: Christian Schicha / Carsten Brosda (Hg.), Handbuch Medienethik, Wiesbaden, S. 83–106.

Debatin, Bernard (2002): Zwischen theoretischer Begründung und praktischer Anwendung: Medienethik auf dem Weg zur kommunikationswissenschaftlichen Teilsdisziplin, in: Publizistik 47 (3/2002), S. 259–264.

Emcke, Carolin (2019): Gegen den Hass, Frankfurt am Main.

Fenner, Dagmar (2025): Digitale Ethik, Tübingen.

Filipović, Alexander (2021): Ethik der Pluralität. Impulse für eine Medienethik pluraler Gesellschaften, in: Communicatio Socalis 54 (3/2021), S. 288–297.

Günther, Klaus (2009): Diskurs, in: Hauke Brunhorst / Regina Kreide / Christina Lafont (Hg.), Habermas Handbuch, Stuttgart, S. 302–306.

Habermas, Jürgen (1988): Theorie des Kommunikativen Handelns, 2. Bd., Frankfurt am Main [zuerst 1981].

Habermas, Jürgen (1991): Exkurs: Transzendenz von innen, Transzendenz ins Diesseits, in: ders., Texte und Kontexte, Frankfurt am Main, S. 127–156.

Habermas, Jürgen (1998): Faktizität und Geltung, Frankfurt am Main.

Habermas, Jürgen (1996): Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft, Frankfurt am Main [zuerst 1962].

Habermas, Jürgen (2009a): Es beginnt mit dem Zeigefinger, in: Die Zeit, 22 Dezember 2009 (online unter: https://www.zeit.de/2009/51/Habermas-Tomasello – letzter Zugriff: 30.7.2025).

Habermas, Jürgen (2009b): Diskursethik, in: ders., Philosophische Texte, Bd. 3, Frankfurt am Main, S. 31–140.

Habermas, Jürgen (2022): Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Frankfurt am Main.

Habermas, Jürgen (2024): „Es musste etwas anders werden“. Gespräche mit Stephan Müller-Doohm und Roman Yos, Frankfurt am Main.

Heesen, Jessica (2024): Ethik in der öffentlichen digitalen Kommunikation, in: Barbara Thomaß / Günter Bentele / Nils C. Borchers (Hg.), Ethik der öffentlichen Kommunikation. Eine kommunikationswissenschaftliche Einführung, Wiesbaden, S. 227–261.

Holtz-Bacha, Christina (2021): Performing Populism. Communication Strategies for Polarization, Provocation and Fearmongering, in: Reinhard Heinisch / Christina Holtz-Bacha / Oscar Mazzoleni (Hg.), Political Populism. A Handbook of Concepts, Questions and Strategies for Research, Baden-Baden, S. 429–440.

Koenen, Erik / Sax, Simon (Hg.) (2025): Prodemokratische Propaganda, Pressekultur und politische Kommunikation in der Weimarer Republik, München.

Krallmann, Dieter / Ziemann, Andreas (2001): Grundkurs Kommunikationswissenschaft, München.

Mandry, Christoph (2012): Pluralismus als „Wert“ – Chancen und Hindernisse aus theologisch-ethischer Sicht, in: Konrad Hilpert (Hg.), Theologische Ethik im Pluralismus. Studien zur theologischen Ethik, Fribourg, S. 229–238.

Manheim, Ernest (1937): The Psychology of Social Conformity. Typoscript London, in: Archiv für die Geschichte der Soziologie in Österreich (AGSÖ), Signatur 31/2 (online unter: https://agso.uni-graz.at/product/31-2-010-manheim-ernest-the-psychology-of-social-conformity-1937-1938/ – letzter Zugriff: 30.7.2025).

Manheim, Ernest (1939): The role of Small Groups in the Formation of Public Opinion, in: Frank Baron / David N. Smith / Charles Reitz (Hg.), Authority, Culture and Communication. The Sociology of Ernest Manheim, Heidelberg, S. 175–180.

Manheim, Ernest (1964): The communicator and his audience: Liberals and traditionalists in Eighteenth Century Germany, in: Werner J. Cahnmann / Alvin Boskoff (Hg.), Sociology and History. Theory and Research, London, S. 505–513.

Manheim, Ernest (1979): Aufklärung und öffentliche Meinung [Die Träger der öffentlichen Meinung]. Studien zur Soziologie der Öffentlichkeit. Stuttgart [zuerst 1933].

Mead, George H. (1975): Geist, Identität und Gesellschaft [Mind, Self and Society. From the Standpoint of a Social Behaviorist], Frankfurt am Main [zuerst 1934].

Merten, Klaus (1977): Kommunikation. Eine Begriffs- und Prozessanalyse, Opladen.

Müller-Doohm, Stefan (2020): Habermas hat die begrifflichen Fundamente für eine Theorie der Moderne entwickelt [Stephan Müller-Doohm im Interview mit Harro Zimmermann], 27. Februar 2020 (online unter: https://www.hoheluft-magazin.de/2020/02/habermas-hat-die-begrifflichen-fundamente-fuer-eine-theorie-der-moderne-entwickelt/ – letzter Zugriff: 27.7.2025).

Niesen, Peter / Herborth, Benjamin (Hg.) (2007): Anarchie der kommunikativen Freiheit. Jürgen Habermas und die Theorie der internationalen Politik, Frankfurt am Main.

Peters, Bernhard (2007): Der Sinn von Öffentlichkeit, Frankfurt am Main.

Popper, Karl (1945): The Open Society and its Enemies, London.

Reitz, Charles (2005): The Call to Concrete Thinking. Ernest Manheims „Zur Logik des Konkreten Begriffs", in: Frank Baron / David N. Smith / Charles Reitz (Hg.), Authority, Culture and Communication. The Sociology of Ernest Manheim, Heidelberg, S. 27–43.

Redshaw, Sahrah (2020): Generalized Other, in: George Ritzer / Chris Rojek (Hg.), The Blackwell Encyclopedia of Sociology, New York.

Rühl, Manfred / Saxer, Ulrich (1981): 25 Jahre Deutscher Presserat. Ein Anlaß für Überlegungen zu einer kommunikationswissenschaftlich fundierten Ethik des Journalismus und der Massenkommunikation, in: Publizistik 26 (4/1981), S. 471–507.

Saxer, Ulrich (1996): Ethik der Kommunikation, in: Gerhard Wittkämper / Anke Kohl (Hg.), Kommunikationspolitik. Einführung in die medienbezogene Politik, Darmstadt, S. 146–168.

Schütz, Alfred / Luckmann, Thomas (2003): Strukturen der Lebenswelt, Stuttgart.

Thomaß, Barbara (2025): Global Media and Communication Ethics. The Tension between Universalism and Cosmopolitism, in: Carola Richter et al. (Hg.), Cosmopolitan Communication Studies. Toward Deep Internationalisation, Bielefeld, S. 53–86.

Tomasello, Michael (2006): Die kulturelle Entwicklung des menschlichen Denkens, Frankfurt am Main.

Tomasello, Michael (2009): Die Ursprünge der menschlichen Kommunikation, Frankfurt am Main.

Tomasello, Michael (2010): Warum wir kooperieren, Frankfurt am Main.

Rutz, Christian et al. (2023): Using machine learning to decode animal communication, in: Science 381, S. 152–155.

Welzig, Elisabeth (1997): Die Bewältigung der Mitte. Ernst Manheim, Soziologe und Anthropologe, Wien.

Wessler, Hartmut (2018): Habermas and the media, Cambridge.

Venema, Rebecca / Averbeck-Lietz, Stefanie (2017): Moralizing and Deliberating in Financial Blogging. Moral Debates in Blog Communication During the Financial Crisis 2008, in: Andreas Hepp / Andreas Breiter / Uwe Hasebrink (Hg.), Communicative Figurations, London, S. 241–265.

Öffentlich-rechtliche Medien als Gestalter konstruktiver Debattenräume: der Public Spaces Incubator

Henning Eichler

Zusammenfassung

Als Mitgestalter von Öffentlichkeit stehen die öffentlich-rechtlichen Medien (ÖRM) vor der Aufgabe, auch digitale Debattenräume zu entwickeln und zu gestalten, die ihrem Funktionsauftrag gerecht werden. Abseits der kommerziellen Logiken sozialer Netzwerke sollen selbst gestaltete Kommunikationsangebote Meinungsvielfalt, konstruktive Debatten und Teilhabe ermöglichen. Vor dem Hintergrund der steigenden Abhängigkeit von global agierenden Plattformen hat das internationalen Forschungs- und Entwicklungsprojekt Public Spaces Incubator (PSI) zum Ziel, eigenständige, gemeinwohlorientierte Kommunikationsräume für den Austausch zwischen Nutzer:innen und Redaktionen der ÖRM aufzubauen. Die hier vorgestellten Ergebnisse einer qualitativen Fallstudie beim beteiligten Zweiten Deutschen Fernsehen (ZDF) zeigen, dass die Entwicklung solcher Räume neben technischen und organisationalen auch ethische Herausforderungen mit sich bringt. Aus den Erwartungen von Redakteur:innen und Community-Manager:innen an die Kommunikation im PSI lassen sich Normen und Werte ableiten, die Auswirkungen für ÖRM als Gestalter öffentlicher Kommunikation haben. Der Beitrag skizziert, wie Innovationen orientiert an operationalisierbaren Werten eine erneuerte Legitimationsgrundlage für ÖRM jenseits des Plattformkapitalismus werden können. Systematischer Bestandteil des Innovationsmanagements der ÖRM sollten ethische Prozesse sein, auch, um erwünschte Makroeffekte evaluieren zu können. Dazu wird das Modell der Prinzipienethik vorgestellt, mit dem Werte und Normen auf Kohärenz geprüft und in Innovationsprozesse integriert werden können.

1. Öffentlich-rechtliche Debattenräume zwischen Plattformökonomie und Auftrag

Öffentlich-rechtliche Medien (ÖRM) sind entscheidende Akteure in der Organisation und Gestaltung öffentlicher Kommunikation. Sie sollen dazu beitragen, einen rational übergreifenden Diskurs aufrechtzuerhalten, „der die Voraussetzung für eine auf umfassende und vielfältige Information gestützte Willensbildung und damit für eine funktionsfähige Demokratie ist" (Dörr 2017: 46). Ein in der Debatte um Legitimation und Zukunft der ÖRM vielfach erwarteter und geforderter *public value* entsteht „auch durch die kommunikativen Beziehungen, die die Institution mit der Zivilgesellschaft unterhält" (Serong 2017: 27).

Debattenräume zu gestalten und zu unterhalten ist Teil des gesetzlichen Auftrags (Ministerpräsidentenkonferenz 2020) und hat zudem eine ethische Dimension. Für Alexander Filipović stehen die Funktionsaufträge, wie sie in den Medienstaatverträgen ausformuliert sind, „mehr oder weniger explizit in Verbindung mit sozialethischen und medienethischen Grundsätzen". Er weist auf damit zusammenhängende Leitwerte wie Meinungsbildung, Gemeinwohl, Integration und Vielfalt hin (Filipović 2019: 100).

Für das Herstellen von Öffentlichkeit haben soziale Netzwerke eine zentrale Rolle eingenommen, vor allem bei der Verbreitung journalistischer Inhalte und für jüngere Nutzer:innen (vgl. Newman et al. 2023; ARD/ZDF Forschungskommission 2023; Nielsen/Fletcher 2020). Drittplattformen sind daher auch in Bezug auf Publikumsbeziehungen fester Bestandteil der Digitalstrategien der öffentlich-rechtlichen Medien (vgl. Eichler 2022). Die Empfehlungsalgorithmen der Plattformen priorisieren jedoch polarisierende, emotionale, zugespitzte oder extremistische Inhalte und schaden so einer konstruktiven Debattenkultur und ausgewogenen Meinungsbildung (vgl. Lewandowsky et al. 2020). In der Kommunikation in sozialen Netzwerken sind Redakteur:innen und Community Manager:innen mit destruktiven Beiträgen und der Bildung von Gruppen Gleichgesinnter konfrontiert (vgl. deCinelli et al. 2021). Folgen sind unter anderem eingeschränkte Informiertheit, kursierende Falschinformationen und Polarisierungstendenzen (vgl. Lewandowsky et al. 2020; Lorenz-Spreen et al. 2022).

Zudem erschwert die intransparente Änderung algorithmischer Funktionen eine zuverlässige Distribution journalistischer Angebote und den Austausch zwischen Redaktion und Publikum. Beispielsweise hat das Unternehmen Meta im Februar 2024 die Einstellungen für seine Netzwerke

Facebook, Threads und Instagram so verändert, dass den Nutzer:innen politische Inhalte von Accounts, denen sie nicht folgen, nur angezeigt werden, wenn diese Einstellungen manuell deaktiviert werden (vgl. Meta 2024).[1]

Auch die Plattformisierung (vgl. Eisenegger 2021; van Dijck/Poell/Waal 2018) journalistischer Inhalte und Arbeitsweisen wirkt auf die Interaktionen zwischen Redaktionen und Publikum zurück. Durch den Einfluss algorithmischer Systeme wird „the public interest-principle […] systematically diminished and marginalized" (Napoli 2019: 159). Diese Entwicklung wird durch Management-Entscheidungen der globalen Plattformen, wie beispielsweise die Beendigung des Fact-Checkings durch Meta, noch verstärkt (vgl. Stippler et al 2025). Gleichzeitig wachsen die Erwartungen an öffentlich-rechtliche Medien, Öffentlichkeit auch in digitalen Räumen herzustellen. So fordert der Rat für die Zukunft der ÖRM, dass diese sich auch als Dialoganstalten verstehen sollten (vgl. Jäkel et al. 2024: 12).

Ein Vorstoß, werteorientierte Debattenräume zu gestalten ist der *Public Spaces Incubator* (PSI), ein internationales Forschungs- und Entwicklungsprojekt, das zum Ziel hat, offene und respektvolle Online-Kommunikation zu fördern (vgl. ZDF 2024; Sgarro/Chan 2024). Die sechs beteiligten ÖRM (ARD und ZDF für Deutschland, ABC Australia, CBC / Radio Canada, SRG / SSR aus der Schweiz, RTBF aus Belgien) wollen Kommunikationsräume aufbauen, die sich an ihren Funktionsaufträgen orientieren und mit selbst entwickelten Technologien realisiert werden. In den eigenen Mediatheken oder Apps werden verschiedene PSI-Anwendungen integriert, sodass Nutzer:innen redaktionell ausgewählte Video- oder Textbeiträge kommentieren, ihre Meinung äußern und sich mit anderen Nutzer:innen oder Redakteur:innen austauschen können (zur Funktionsweise ausgewählter Prototypen siehe https://www.youtube.com/@wearenew_public/videos).

2. Erwartungen an Journalismus-Publikum-Beziehungen und normative Aspekte

PSI soll der Abhängigkeit der ÖRM von kommerziellen Netzwerken entgegenwirken und ist eine Reaktion auf veränderte Erwartungen in den Beziehungen zwischen Journalismus und seinem Publikum. Da von ÖRM

1 Eine entsprechende Regelung wurde im Januar für den deutschen Markt kommuniziert (Meta Transparency Center (2025). In einer anderen Mitteilung kündigt Instagram-Geschäftsführer Adam Mosseri wiederum an, in den Empfehlungsalgorithmen wieder mehr politische Inhalte zuzulassen (vgl. Mosseri, 2025).

in besonderem Maße erwartet wird, gesellschaftliche Anforderungen zu erfüllen und den Austausch mit Nutzer:innen aktiv zu gestalten (vgl. Stehle et al. 2022), gilt es, die wechselseitigen Erwartungen zwischen Journalismus und seinem Publikum unter veränderten Medienroutinen weiter zu untersuchen.

Die Forschung zu *Audience Engagement* (für einen Überblick vgl. Gajardo/Costera Meijer 2023) zeigt: Nutzer:innen digitaler Debattenräume erwarten, dass sie Meinungen in öffentlichen Dialogen mit Anderen und Journalist:innen äußern können, dass sie Journalist:innen Feedback geben und Fragen zu deren Arbeit stellen können (vgl. Detel et al. 2023; Uth 2025). Spezifische Publikumserwartungen an ÖRM sind, dass Interaktionen von Journalist:innen moderiert werden, um einen respektvollen und konstruktiven Dialog zu gewährleisten, aufgeheizte Debatten zu versachlichen und als Vermittler:innen zwischen Bürger:innen und Politik zu fungieren (vgl. Mothes/Prinzing 2025).

Während mittlerweile also erste Erkenntnisse dazu vorliegen, welche Erwartungen das Publikum an den Austausch mit Journalist:innen richtet, besteht bezüglich der Erwartungen von Journalist:innen an ihr Publikum jedoch eine Forschungslücke. Eine erste Studie von Detel et al. (2023) zeigt, dass die Erwartungen von Journalist:innen an Interaktionen mit ihrem Publikum sich primär auf Höflichkeit und Konstruktivität in der Interaktion sowie dem Wunsch nach thematischem Input konzentrieren.

Neben erwünschten individuellen Effekten auf einer Mikroebene sind für eine erneuerte Legimitation der ÖRM nachweisbare und operationalisierbare Makrowirkungen entscheidend, argumentiert Neuberger (2024). Dazu müsse der Journalismus „seine Vermittlungsfunktion nachjustieren, indem er die Werte der liberalen Demokratie für den digitalen Kontext innovativ operationalisiert, normativ absichert und dafür Akzeptanz im Publikum gewinnt“ (ebd.: 35). Der öffentlich-rechtliche Auftrag ließe sich präzisieren, indem Bezug auf bestimmte Werte genommen und diese durch konkrete Normen verwirklicht werden. Werte sollten in einem partizipativen Public Value-Prozess als „Fixpunkte einer normativen Analyse“ (ebd.: 28) dienen. Aus dem Wertekatalog, den Neuberger (2024) für Medienbewertungen aufstellt, sind für die vorliegende Untersuchung die Werte Integration, Vielfalt und Diskursqualität besonders relevant. Das ergibt sich aus den geäußerten Erwartungen in den Leitfadeninterviews sowie aus für diese Studie zugänglich gemachten internen Dokumenten des ZDF aus der strategischen Planungsphase.

Bei der Einführung externer Innovationen (wie beispielsweise journalistischen Arbeitsweisen, die sich stark an den Logiken kommerzieller Plattformen orientieren) besteht die Gefahr eines *normative failure* (vgl. Siegelbaum/Thomas 2016). Journalistische Funktionsaufträge können dann nicht ausreichend erfüllt werden. Zielführender sind daher interne, wertegeleitete Innovationen, die in Eigenregie entwickelt und am Auftrag und an der Funktion der ÖRM orientiert umgesetzt werden. Für diesen Beitrag ist die Fragestellung leitend, welche ethischen Herausforderungen sich bei Entwicklung und Evaluation interner Innovationen abzeichnen.

PSI ist eine solche interne Innovation. Der folgende Abschnitt stellt Ergebnisse einer qualitativen Fallstudie zu den Erwartungen der Journalist:innen, in der Journalismus-Publikum-Beziehung vor.

3. Studiendesign

PSI ist eine Initiative, die 2023 von CBC / Radio-Canada, SRG / SSR (Schweiz), RTBF (Belgien) und dem ZDF in Zusammenarbeit mit New_ Public gegründet wurde. Inzwischen sind die ARD und die australische ABC dem Projekt beigetreten, das zum Ziel hat, geschützte digitale Räume für konstruktive und respektvolle Diskurse im Onlinebereich zu schaffen. In der Entwicklungsphase entstanden mehrere Prototypen, die verschiedene Formen der Interaktion zwischen Nutzer:innen untereinander sowie zwischen Nutzer:innen und Redaktionen ermöglichen. Die Prototypen sollen in ausgewählte journalistische Angebote – darunter Webseiten, Apps und Mediatheken – integriert werden.

Drei dieser Prototypen bilden den Ausgangspunkt für die vorliegende Studie. Sie wurden in einer Testumgebung von Journalist:innen beim ZDF erprobt:

- *Public Square View*: Nutzer:innen haben die Möglichkeit, während eines Livestreams über Emojis auf Inhalte zu reagieren und an Kurzumfragen teilzunehmen. Anschließend können sie sich in einem Chat über die Inhalte austauschen.
- *Representing Perspectives*: Bei dieser Funktion geben Nutzer:innen an, welche Perspektive bzw. Rolle sie beim Verfassen eines Kommentars einnehmen (möchten). Die Rollen variieren je nach Thema und werden im Vorfeld von den Redaktionen festgelegt.
- *Comments Slider*: Für Nutzer:innen, die selbst keinen Kommentar schreiben möchten, steht ein verschiebbarer Regler zur Verfügung, mit

dem sie ihre Position zu einem Thema anzeigen können. Neben binären Antworten sind auch nuancierte Haltungen möglich, die durch Begründungen ergänzt werden können.

Aufbauend auf den gesammelten Erfahrungen der Journalist:innen stehen folgende Forschungsfragen im Zentrum der Untersuchung:

F1: Welche Erwartungen an PSI bestehen in Bezug auf Debattenkultur und Meinungsbildungsprozesse?
F2: Welche Normen und Werte sind in PSI leitend und warum?
F3: Wie gestalten und verändern sich Rolle und Auftrag der ÖRM in der Herstellung und Gestaltung von Debattenräumen?
F4: Welche ethischen Herausforderungen für ÖRM als Gestalter öffentlicher Kommunikation ergeben sich aus den vorigen Fragen?

Um diese Forschungsfragen zu beantworten, wurden qualitative, semistrukturierte Leitfadeninterviews mit 16 Mitarbeiter:innen des ZDF geführt, darunter Journalist:innen, Community Manager:innen sowie Vertreter:innen aus dem strategischen Management. Die Datenerhebung erfolgte im Zeitraum vom 21. Juni bis 16. Juli 2024. Die Auswertung basiert auf einer zusammenfassenden und strukturierenden qualitativen Inhaltsanalyse nach Mayring (2015). Die Kategorien wurden zunächst deduktiv entwickelt und anschließend induktiv anhand des empirischen Materials erweitert.

Die Interviews fokussierten sich auf drei Themenbereiche:

(1) die Beziehung zwischen Journalismus und Publikum auf kommerziellen Plattformen (einschließlich deren Relevanz, bisheriger Erfahrungen und normativer Erwartungen),
(2) die bisherigen Erfahrungen der Befragten mit den PSI-Prototypen im Vergleich zu Interaktionen auf kommerziellen Plattformen und
(3) das wahrgenommene Potenzial von PSI, Abhängigkeiten von kommerziellen Plattformen zu verringern.

Dieser Beitrag legt vor diesem Hintergrund einen Fokus auf ethische Herausforderungen, die sich für öffentlich-rechtliche Medien aus der Entwicklung und Gestaltung von PSI-Debattenräumen ergeben.

Ein qualitativer Forschungsansatz in Form einer Fallstudie ist angemessen, da es sich bei PSI um ein neuartiges und bislang empirisch nicht untersuchtes Phänomen handelt (vgl. Yin 2009; Speier-Werner 2006). Methodisch begrenzt ist die Studie insofern, als die Einschätzungen der

Befragten auf Erfahrungen in Testumgebungen basieren und somit keine Rückschlüsse auf reale Anwendungsszenarien zulassen. Die Verlässlichkeit und Gültigkeit der Ergebnisse wurde durch die gemeinsame Durchführung der Datenerhebung und -auswertung durch alle Autor:innen sowie durch kontinuierliche Reflexion und Abgleiche im Codierprozess sichergestellt.

Die Namen der Interviewpartner:innen werden – mit Ausnahme von Paul-Christian Britz und Robert Amlung – anonymisiert. Bei diesen beiden Personen handelt es sich um die Verantwortlichen des Projekts beim ZDF, die nach Rücksprache in ihrer offiziellen Funktion namentlich zitiert werden dürfen.

4. Ergebnisse

In den nachfolgenden Abschnitten werden die Erwartungen der befragten Personen an das Projekt PSI sowie die daraus resultierenden ethischen Herausforderungen dargestellt. Sofern nicht ausdrücklich anders angegeben, beziehen sich die im Text dargestellten Aussagen auf alle getesteten Prototypen.

4.1 Journalismus-Publikum-Beziehung: Erwartungen an PSI

Die zentrale Erwartung der befragten Personen an PSI liegt in der Förderung eines respektvollen und konstruktiven Austausches – sowohl zwischen den Nutzenden als auch zwischen Journalist:innen und Publikum. Ein besonderes Anliegen ist dabei, einen geschützten Kommunikationsraum zu etablieren, der frei von Hate Speech und anderen inzivilen Beiträgen ist. Die Befragten verbinden mit PSI die Hoffnung, dass sich dort auch Nutzer:innen beteiligen, die sich auf kommerziellen Plattformen aus Sorge vor Hassrede oder unzureichender Moderation bislang nicht geäußert haben.

Mit PSI verbinden manche Befragte die Chance, sich von der Logik der Aufmerksamkeitsökonomie zu lösen. Besonders hervorgehoben wird die Möglichkeit, Inhalte und Interaktionen jenseits der marktgetriebenen Dynamiken von Reichweite, Geschwindigkeit und Reizüberflutung zu gestalten. Eine befragte Person formuliert dies so: „Ich glaube, das ist auch ziemlich ein Kernproblem bei den kommerziellen Plattformen. Es geht halt ausschließlich um Aufmerksamkeitsökonomie und Geschwindigkeit“ (I10).

In diesem Kontext wird auch der Wunsch nach stärker individualisierten, dialogorientierten Kommunikationsumgebungen geäußert, in denen Nutzer:innen sich sicher und begleitet fühlen. PSI könnte – so eine weitere befragte Person – dazu beitragen, weniger „Massenabfertigung" zu erleben, sondern vielmehr „eine hohe individuelle Begleitung, also ein sehr hohes Maß an Safe Space" zu realisieren (I12).

Der Austausch soll entschleunigt werden und Raum schaffen für inhaltliche Tiefe: „Dass irgendwie der Fokus weggeht von dem: Was macht am meisten Traffic, sondern was bezieht sich am ehesten auf den Beitrag? Was ist inhaltlich oder auch kommunikativ wertvoll […]?" (I10). PSI wird emnach als Gegenentwurf zur Logik kommerzieller Plattformen verstanden, der nicht auf maximale Sichtbarkeit, sondern auf sinnvolle, respektvolle Interaktion abzielt.

Wiederholt wurde außerdem die potenziell entlastende Wirkung im Community Management genannt. Die Befragten verweisen auf den hohen zeitlichen und emotionalen Aufwand, der mit der Moderation von Kommentaren auf kommerziellen Plattformen verbunden ist. Sollten PSI-Prototypen künftig mit KI-basierten Funktionen ausgestattet werden, könnten sie hier einen Beitrag leisten, indem sie Vorfilterungen vornehmen, problematische Inhalte identifizieren und gruppieren sowie potenziell passende Diskussionspartner:innen zusammenführen. Eine befragte Person formuliert dies wie folgt: Die PSI-Tools könnten dabei helfen, Nachrichten „gruppenweise vielleicht zu beantworten, Dinge zu beantworten oder Diskussionsgruppen besser zu sortieren und Gleichgesinnte, Gleich-Interessierte zusammenzuführen" (I4). Diese technische Unterstützung wird als Vorteil gegenüber bestehenden Plattformmechanismen wahrgenommen, da sie nicht nur die Arbeitsbelastung von Moderator:innen reduzieren, sondern auch die Qualität der Diskurse verbessern könnte.

Gleichwohl wurden in der Erhebung auch kritische Perspektiven formuliert. Eine wiederkehrende Sorge bezieht sich auf die Konkurrenzfähigkeit von PSI gegenüber etablierten kommerziellen Netzwerken. Das routinierte Mediennutzungsverhalten vieler Menschen – insbesondere die Gewohnheit, sich über wenige, große Plattformen zu informieren und zu beteiligen – könnte ein wesentliches Hindernis für die Nutzer:innenbindung an PSI darstellen. Darüber hinaus wird das Risiko fragmentierter Teilöffentlichkeiten angesprochen, die sich in PSI entwickeln könnten: „Die Gefahr ist natürlich auch, dass man sich die Ultra-Blasen aufbaut. […] ich habe Sorge vor Parallelgesellschaft" (I11).

Dem entgegen ist ein grundlegendes Ziel von PSI, geschützte digitale Debattenräume zu schaffen, die vielfältige und inklusive Formen der Publikumsbeteiligung ermöglichen. Eine Erwartung ist, Nutzende anzusprechen, die sich bislang nur selten oder gar nicht an Online-Diskursen beteiligt haben. Dies spiegelt sich auch in den handlungsleitenden Werten und Normen wider.

4.2 Handlungsleitende Werte und Normen bei PSI

Die Interviews verdeutlichen, dass gelingende Kommunikation für die Befragten vor allem durch ein Gefühl von Sicherheit, Offenheit und gegenseitigem Respekt charakterisiert ist. Ein:e Interviewpartner:in beschreibt dies wie folgt:

> „Wann ist eine gute Kommunikation? […] Also wenn ich das Gefühl habe, ich kann mich äußern, ohne dass ich dafür an den Pranger gestellt werde. Ich kann meine Gefühle äußern. Ich kann auch mal Kritik äußern und gleichzeitig haben alle so eine Ebene, dass sie dabei respektvoll bleiben, dass man jemandem zuhört" (I7).

Diese Aussage unterstreicht, dass Leitwerte wie Respekt, Toleranz und Meinungsfreiheit als Voraussetzungen für eine funktionierende Debattenkultur betrachtet werden.

Darüber hinaus ist Transparenz ein zentraler Aspekt des PSI: Die Nutzerinnen sollen nachvollziehen können, welche Daten erhoben werden und zu welchen Zwecken sie verwendet werden. Ziel der Prototypen ist es laut einem Befragten, durch Transparenz und Zugänglichkeit Diversität und Vertrauen zu fördern und sich damit von kommerziellen Plattformen zu unterscheiden. Dies spiegelt sich laut den Interviewpartner:innen auch darin wider, dass PSI als Open-Source-Software entwickelt und veröffentlicht wird. Somit können auch andere Akteur:innen darauf zurückgreifen und PSI weiterentwickeln. Transparenz wird schließlich auch im redaktionellen Kontext als bedeutsam erachtet. Mehrere Befragte erwähnen, dass Nutzer:innen dadurch frühzeitig Einblicke in redaktionelle Prozesse erhalten können und journalistische Arbeit so nachvollziehbar wird.

4.3 Erfahrungen mit kommerziellen Plattformen und Potentiale für mehr Unabhängigkeit

Die Erwartungen an die ideale Interaktion zwischen Redaktion und Publikum wie sie im vorangegangenen Abschnitt für PSI beschrieben wurden, treffen den Erfahrungswerten der Befragten zufolge auf kommerzielle Plattformen nur in begrenztem Umfang zu. Zwar wird der Austausch mit Nutzer:innen auf kommerziellen Plattformen von den Befragten insgesamt als bereichernd beschrieben – insbesondere hinsichtlich Themenvorschlägen und direktem Feedback zu Inhalten. Gleichzeitig zeigen die Interviews jedoch, dass dieser Austausch auch als problematisch erlebt wird. Kritisiert werden vor allem die eingeschränkte algorithmisch gesteuerte Ausspielung einzelner Inhalte, die fehlende Transparenz der algorithmischen Auswahlmechanismen sowie unklare Datenschutzpraktiken. Mehrere Befragte äußern die Vermutung, dass Inhalte entweder algorithmisch gedrosselt oder nur bestimmten Nutzer:innengruppen ausgespielt werden, wodurch sich Teilöffentlichkeiten bilden. In solchen Kommunikationsräumen können sich problematische Interaktionen wie Hassrede, Beleidigungen oder diskriminierende Äußerungen häufen. Insgesamt wird in den Interviews deutlich, dass eine selbstbestimmte und redaktionell kontrollierbare Interaktion mit dem Publikum unter den Bedingungen kommerzieller Plattformen sehr herausfordernd ist.

Besonders das Community Management stellt die Redaktionen vor erheblichen Hürden. Mehrere Interviewpartner:innen kritisieren eine mangelnde Benutzer:innenfreundlichkeit, etwa beim Melden oder Moderieren von problematischen Kommentaren. Gleichzeitig entsteht der Eindruck, dass die Sichtbarkeit kontroverser und negativer Kommentare algorithmisch begünstigt wird. Eine befragte Person formuliert dies so: „[…] leider ist die Kommentarspalte in Social Media teilweise wirklich hochtoxisch. Das finde ich wirklich nicht gut. Für niemanden" (I8).

Diese Funktionsweisen kommerzieller Plattformen führen nach Einschätzung mehrerer Befragter dazu, dass manche Nutzer:innen öffentliche Online-Debatten meiden oder sich in kleinere oder private Räume zurückziehen. Das wird als problematisch bewertet, weil dadurch öffentliche Debattenräume schrumpfen und Perspektiven fehlen: „Das ist jetzt meine ganz persönliche Einschätzung […], dass Leute sich tatsächlich eher von diesen Plattformen zurückziehen. Sie ziehen sich zurück in kleinere Räume oder in private Räume, weil dieses Environment so ist, wie es ist und das

wiederum sehe ich als Risiko, weil dann fällt diese öffentliche Diskussion weg […]" (Paul-Christian Britz).

Vor dem Hintergrund dieser Herausforderungen reflektieren die Befragten die Aktivitäten der ÖRM auf kommerziellen Plattformen kritisch. Einerseits wird anerkannt, dass diese Plattformen für die Erfüllung des öffentlich-rechtlichen Auftrags unverzichtbar sind – insbesondere, um Reichweite und Sichtbarkeit bei bestimmten Zielgruppen zu generieren. Andererseits bieten eigene Infrastrukturen wie PSI zumindest die Möglichkeit, sich von problematischen Funktionslogiken kommerzieller Plattformen abzugrenzen – besonders dann, wenn deren ökonomische Prinzipien den Leitwerten und dem Funktionsauftrag der ÖRM widersprechen. Gleichzeitig betonen die Befragten, dass PSI allein das Abhängigkeitsverhältnis zu kommerziellen Plattformen nicht auflösen könne: „Wir werden die großen amerikanischen Plattformen nicht los, und wir werden sie nutzen müssen für den Auftrag, weil wir anders bestimmte Reichweiten nicht generieren" (Robert Amlung).

PSI wird demnach weniger als Ersatzlösung, sondern vielmehr als ergänzende Infrastruktur verstanden, um den Austausch mit dem Publikum zu stärken. PSI könnte insbesondere jene Nutzer:innen erreichen, die sich von der Diskussionskultur auf kommerziellen Plattformen bislang nicht angesprochen fühlen. Gleichwohl werden Zweifel geäußert, inwiefern PSI gerade für jüngere Zielgruppen ein attraktives Angebot darstellen kann: „Wenn sie sowieso auf drei bis fünf Plattformen unterwegs sind, wollen sie da jetzt noch mitmachen? Es sind immer dieselben, die man anspricht […] oder zu wenige Leute" (I14).

Erschwerend kommt hinzu, dass PSI funktional in manchen Aspekten nicht mit den großen Plattformen konkurrieren kann. So ist zum Beispiel eine direkte Verlinkung zu externen Quellen bislang nicht möglich – ein Umstand, der die Attraktivität für Nutzer:innen einschränkt. Trotz dieser Limitationen äußern einige Befragte konkrete Strategien, wie sich PSI mittelfristig als Gegenmodell zu kommerziellen Plattformen etablieren lassen könnte. So könnten etwa ausgewählte Inhalte ausschließlich auf eigenen Plattformen wie der ZDF-Mediathek veröffentlicht werden, während auf kommerziellen Kanälen lediglich verknappter Content in Teaser-Funktion publiziert wird. Laut einigen Befragten ist denkbar, dass PSI punktuell Abhängigkeiten reduzieren könnte: Etwa durch eine schrittweise Verlagerung bestimmter Inhalte von YouTube-Kanälen exklusiv in die Mediathek. Auch das gezielte Deaktivieren der Kommentarfunktion auf kommerziellen Platt-

formen wird als Maßnahme diskutiert, um Debatten in die Mediatheken und damit zum PSI zu lenken.

Den Befragten ist insgesamt bewusst, dass PSI keine vollständige Abkehr von kommerziellen Plattformen bedeutet. Vielmehr könnte PSI eine Balance schaffen zwischen notwendiger Präsenz auf kommerziellen Plattformen und dem Aufbau eigener, wertegeleiteter Infrastrukturen.

5. Innovationen ethisch gestalten

Der folgende Abschnitt setzt die zuvor präsentierten Ergebnisse in einen ethischen Rahmen. Es wird diskutiert, welche Herausforderungen für ÖRM in der Entwicklung interner Innovationen aus ethischer Perspektive entstehen.

Kommerzielle soziale Netzwerke funktionieren nicht auf Basis gemeinwohlorientierter Normen. Unsere Studie zeigt einen Strategiewechsel der beteiligten ÖRM im Umgang mit diesen Plattformen. Durch die Entwicklung eigener Kommunikationsräume sollen Interaktionen und Engagement des Publikums neugestaltet und Abhängigkeiten von globalen Digitalkonzernen verringert werden. Mit diesen *internen,* wertegeleiteten Innovationen sollen zugleich normative Anforderungen an ÖRM besser erfüllt und dysfunktionale Effekte wie *normative failure* vermieden werden.

Eine völlige Abkehr von kommerziellen Netzwerken ist unter den derzeitigen Gegebenheiten jedoch unrealistisch. Zu groß sind die Reichweiten, zu groß wären die Publikumsverluste und die vertanen Chancen auf kommunikative Beziehungen. Umso mehr wird es darauf ankommen, die Journalismus-Publikum-Beziehungen bewusst zu gestalten und aktiv zu managen. Ergebnisse unserer Studie weisen darauf hin, dass es insbesondere darauf ankommen wird, mimetische Isomorphie (vgl. Lischka 2024) – also das Ausrichten an organisationalen Strukturen oder das Nachahmen von Verhaltensweisen kommerzieller Plattformen – zu vermeiden. Stattdessen sollten Innovationen orientiert an operationalisierbaren Werten (vgl. Neuberger 2024) entwickelt, kommuniziert und evaluiert werden.

Es wird deutlich, dass solche wertebasierten Innovationen nicht ohne eine ethische Reflektion auskommen. Das Management von Medieninnovationen, insbesondere im Bereich der ÖRM, sollte sich dabei am Modell des *normative turn* (vgl. Litschka/Krainer 2022) orientieren. Dieses geht davon aus, dass sich eine Medienorganisation in einem unbegrenzten öf-

fentlichen Legitimationsprozess im Habermas'schen Sinne einer Diskursethik befindet.

Moralische Begründungen und empirisch-ethische Analysen (Filipović 2019) gehen Hand in Hand, wenn Handlungsweisen, technische Funktionen und Infrastrukturen aus dem Funktionsauftrag (mit Leitwerten wie Meinungsbildung, Gemeinwohl, Vielfalt, Integration) abgeleitet und begründet werden können. Konkrete Beispiele für selbst gesetzte Normen innerhalb eigener Infrastrukturen wie dem PSI sind z.B. der Einsatz Künstlicher Intelligenz im Community Management, das Blockieren destruktiver Nutzer:innenbeiträge oder selbst entwickelte algorithmische Empfehlungssysteme für das Kuratieren redaktioneller Inhalte.

6. Werte und Normen als Bestandteil der Legitimation

Für das zielführende Entwickeln und Gestalten interner Innovationen nach ethischen Kriterien bedeutet dies, einen Kanon von Werten als Grundlage zu haben, aus dem heraus sich Normen als konkrete Handlungsregeln ableiten lassen. Mit einer solchen systematischen Herangehensweise lässt sich Innovationsmanagement im Sinne eines *normative turn* – orientiert am Funktionsauftrag – realisieren.

Ein Modell zum Abgleich von Werten als Handlungszielen und Normen als konkreten Handlungsregeln bietet der kohärentistische Ansatz der Prinzipienethik (vgl. Beauchamp 2016). Dabei werden in einem Verfahren der Spezifizierung einzelne Werte und davon abgeleitete Handlungsregeln auf ihren Kohärenzgehalt diskutiert und überprüft. Ausgehend von einem Grundprinzip wird ein Geltungsbereich eingegrenzt, mit dem Ziel, eine Kohärenz zwischen Normen und Werten zu erreichen. Voraussetzung dafür ist ein Set von Prinzipien mittlerer Reichweite, die vorerst ohne weitere argumentative Unterstützung akzeptabel und in sich kohärent sind. Diese Prinzipien gelten nicht als absolut, erheben aber universellen Geltungsanspruch. Konkret könnten die ÖRM also ein Set von Werten formulieren und daran orientiert überprüfen, welche Normen handlungsleitend sein sollen. Wie eine Prinzipienethik für den Bereich der Medienethik und konkret für das Spannungsfeld zwischen normativen Ansprüchen an ÖRM, Plattformlogiken und erneuerter Legitimation ausgelegt werden kann, skizziert Eichler (2024).

Auch um erwünschte Makroeffekte (wie zum Beispiel eine integrierte statt einer fragmentierten Öffentlichkeit, Diskursqualität) zu erzielen und

systematisch nachweisen zu können, sind im Management der ÖRM umfassende medienethische Prozesse nötig. Diese transparent und partizipativ zu gestalten, muss Teil ihrer Vermittlungsfunktion werden.

Insgesamt wird deutlich, dass auf öffentlich-rechtliche Medien durch technische und organisationale Innovationen auch neue ethische Aufgaben zukommen. Die Herausforderung besteht darin, diese Innovationen so zu gestalten, dass individuelle Effekte auf der Mikroebene mit nachweisbaren Makroeffekten einhergehen.

Unsere Forschungsergebnisse am Beispiel des PSI weisen darauf hin, dass eine werteorientierte Gestaltung des *Audience Engagement* sowohl einen Beitrag zu einer funktionsfähigen Demokratie in einer zunehmend fragmentierten und plattformisierten Medienlandschaft leisten als auch Bestandteil einer ethisch begründeten Legitimation der ÖRM sein kann.

In einer nächsten Forschungsphase können Befragungen von Nutzer:innen im Rahmen von Tests der Prototypen durchgeführt werden. Der Fokus sollte dabei ebenfalls auf Werten liegen: Wie gut werden konstruktive Debatten ermöglicht? Wie überzeugend werden Teilhabe, Vielfalt und Partizipation realisiert? Wie stark ist die integrierende Wirkung der PSI-Angebote? Mit solchen Erhebungen aus Nutzerperspektiven ließen sich weitere Erkenntnisse zu den hier skizzierten ethischen Aspekten wertegeleiteter interner Innovationen gewinnen und der Bogen zur Evaluation erwünschter Makrowirkungen schlagen.[2]

Hinweis: Teile dieses Beitrages wurden aus einem bereits erschienenen Text für Digital Journalism (vgl. Eichler et al. under review) in überarbeiteter Form übernommen. Dort werden PSI und die getesteten Prototypen näher vorgestellt.

Literatur

ARD/ZDF Forschungskommission (2023): ARD/ZDF Onlinestudie, 30. Dezember 2023 (online unter: https://www.ard-zdf-onlinestudie.de/ – letzter Zugriff: 8.12.2025).

2 An der Forschung beteiligte: Dr. Henning Eichler ist Professor für medienadäquate Inhalteaufbereitung an der Hochschule für Technik, Wirtschaft und Kultur Leipzig. Vanessa Kokoschka, M.Sc. ist Doktorandin und wissenschaftliche Mitarbeiterin am Fachbereich Media der Hochschule Darmstadt. Dr. Bernadette Uth ist wissenschaftliche Mitarbeiterin (postdoctoral) am Journalism Studies Center am Institut für Publizistik und Kommunikationswissenschaft der Universität Wien. Hannah Lea Ötting, M.A., M.Sc. ist wissenschaftliche Mitarbeiterin am Institut für Kommunikationswissenschaft der Universität Münster.

Beauchamp, Tom L. (2016): The Principles of Biomedical Ethics as Universal Principles, in: Mohammed Ghaly (Hg.), Islamic Perspectives on the Principles of Biomedical Ethics. Muslim Religious Scholars and Biomedical Scientists in Face-To-Face Dialogue with Western Bioethicists, Doha, S. 91–119.

deCinelli, Matteo et al. (2021): The echo chamber effect on social media, in: Proceedings of the National Academy of Sciences of the United States of America 118 (9/2021). https://doi.org/10.1073/pnas.2023301118

Detel, Hanne et al. (2023): The impact of mutual interaction expectations on journalist-audience relations in digital media contexts: An exploratory study (= Paper presentation, 73rd Annual Conference of the International Communication Association (ICA), Toronto).

Dörr, Dieter (2017): Zukunftsfähiger Funktionsauftrag des öffentlich-rechtlichen Rundfunks, in: Medienwirtschaft. Zeitschrift für Medienmanagement und Medienökonomie 14 (4/2017), S. 40–48.

Eichler, Henning (2022): Journalismus in sozialen Netzwerken. ARD und ZDF im Bann der Algorithmen? (= OBS-Arbeitsheft, Band 110), Frankfurt am Main.

Eichler, Henning (2024): Mit Prinzipienethik aus der Plattformfalle. Ein Verfahren für Digitalstrategie und erneuerte Legitimation der öffentlich-rechtlichen Medien, in: Communicatio Socialis 57 (2/2024), S. 169–185.

Eichler, Henning et al. (under review): Debate spaces for everyone? How public service media aim to reshape audience engagement and interactions, in: Digital Journalism.

Eisenegger, Mark (2021): Dritter, digitaler Strukturwandel der Öffentlichkeit als Folge der Plattformisierung, in: Mark Eisenegger et al. (Hg.), Digitaler Strukturwandel der Öffentlichkeit, Wiesbaden, S. 16–39.

Filipović, Alexander (2019): Öffentlichkeitsbegriff und Gemeinwohlrelevanz des öffentlich-rechtlichen Rundfunks, in: Marianne Heimbach-Steins (Hg.), Öffentlich-rechtliche Medien, Münster, S. 87–112.

Gajardo, Constanza / Costera Meijer, Irene (2023): How to tackle the conceptual inconsistency of audience engagement? The introduction of the Dynamic Model of Audience Engagement, in: Journalism 24, S. 1959–1979.

Jäkel, Julia et al. (2024): Bericht des Rates für die zukünftige Entwicklung des öffentlich-rechtlichen Rundfunks, Januar 2024 (online unter: https://rundfunkkommission.rlp.de/fileadmin/rundfunkkommission/Dokumente/Zukunftsrat/ZR_Bericht_18.1.2024.pdf – letzter Zugriff 2.7.2024).

Lewandowsky, Stephan et al. (2020): Technology and democracy. Understanding the influence of online technologies on political behaviour and decision-making, Luxembourg.

Lischka, Juliane A. (2024): Isomorphie durch Innovation, in: Sonja Kretzschmar / Daniel Nölleke / Annika Sehl (Hg.), Innovationen im Journalismus. Theorien, Modelle, Potentiale?, Wiesbaden, S. 35–44.

Litschka, Michael / Krainer, Larisssa (2022): The Normative Turn in the Organisation of Media. Ethical Considerations for Change Management in Media Enterprises, in: Matthias Karmasin / Sandra Diehl / Isabell Koinig (Hg.), Media and Change Management. Creating a Path for New Content Formats, Business Models, Consumer Roles, and Business Responsibility, Cham, S. 331–341.

Lorenz-Spreen, Philipp et al. (2022): A systematic review of worldwide causal and correlational evidence on digital media and democracy, in: Nature Human Behaviour 7, S. 74–101.

Mayring, Philipp (2015): Qualitative Inhaltsanalyse. Grundlagen und Techniken, Weinheim.

Meta (2024): Continuing our Approach to Political Content on Instagram and Threads, 09. Februar 2024 (online unter: https://about.instagram.com/blog/announcements/continuing-our-approach-to-political-content-on-instagram-and-threads – letzter Zugriff: 19.6.2025):

Meta Transparency Center (2025): Unser Ansatz für politische Inhalte, 07. Januar 2025 (online unter: https://transparency.meta.com/de-de/features/approach-to-political-content/ – letzter Zugriff: 8.12.2025).

Mosseri, Adam (2025): Mitteilung Threads, 08. Januar 2025 (online unter: https://www.threads.com/@mosseri/post/DEk65zdTVmX – letzter Zugriff 19.6.2025).

Mothes, Cornelia / Prinzing, Marlis (2025): Was macht öffentlich-rechtlichen Journalismus wertvoll? Der Public Value öffentlich-rechtlicher Medien aus Sicht der Nutzerinnen und Nutzer, in: MediaPerspektiven 8, S. 1–33.

Napoli, Philip (2019): Social Media and the Public Interest. Media Regulation in the Disinformation Age, New York.

Neuberger, Christoph (2024): Werte als Maßstab der liberal-demokratischen Öffentlichkeit, in: Marlis Prinzing et al. (Hg.), Regulierung, Governance und Medienethik in der digitalen Gesellschaft, Wiesbaden, S. 23–43.

Newman, Nic et al. (2023): Reuters Institute Digital News Report 2023, hrsg. von Reuters Institute for the Study of Journalism, Oxford (online unter: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf – letzter Zugriff: 30.12.2023).

Nielsen, Rasmus K. / Fletcher, Richard (2020): Democratic Creative Destruction? The Effect of a Changing Media Landscape on Democracy, in: Nathaniel Persily / Joshua A. Tucker (Hg.), Social media and democracy. The state of the field, prospects for reform, New York, S. 139–162.

Serong, Julia (2017): Die Öffentlich-Rechtlichen und Public Value. Über das ungenutzte Potential des Public-Value-Begriffs, in: Communicatio Socialis 50, S. 20–34.

Sgarro, Victoria / Chan, Min L. (2024): How we're building with international broadcasters. A deep dive into product design for public conversation, in: New Public, 23. Juni 2024 (online unter: https://newpublic.substack.com/p/how-were-building-with-international – letzter Zugriff: 26.6.2024).

Siegelbaum, Sasu / Thomas, Ryan J. (2016): Putting the Work (back) into Newswork, in: Journalism Practice 10, S. 387–404.

Speier-Werner, Petra (2006): Die Einzelfallstudie, in: Harald Barrios / Christoph H. Stefes (Hg.), Einführung in die Comparative Politics, Berlin, Boston, S. 52–58.

Stehle, Helena / Uth, Bernadette / Detel, Hanne (2022): Erwartungen in der Journalismus-Publikum Beziehung in digitalen Kontexten, in: ORF (Hg.), Public Value Studie. Öffentlich-rechtliche Qualität im Diskurs, Wien, S. 197–220.

Stippler, Felix et al. (2025): Meta beendet Faktenchecks auf Facebook und Instagram, in: Handelsblatt, 08. Januar 2025 (online unter: https://www.handelsblatt.com/technik/it-internet/tech-konzern-meta-beendet-faktenchecks-auf-facebook-und-instagram/100099044.html – letzter Zugriff: 29.5.2025).

Uth, Bernadette (2025): What does the audience expect when interacting with journalists? A Q-sort study, in: Journalism Practice. https://doi.org/10.1080/17512786.2025.2551986

van Dijck, José / Poell, Thomas / Waal, Martijn d. (2018): The platform society. Public values in a connective world, New York.

Yin, Robert K. (2009): Case study research. Design and methods, Los Angeles.

ZDF (2024): Mehr konstruktiver Diskurs im Netz: ZDF und internationale Partner stellen Prototypen vor. Public Spaces Incubator für mehr Respekt im öffentlichen Austausch. Pressemitteilung, in: ZDF, 07. Mai 2024 (online unter: https://presseportal.zdf.de/pressemitteilung/mehr-konstruktiver-diskurs-im-netz-zdf-und-internationale-partner-stellen-prototypen-vor – letzter Zugriff: 26.6.2024).

Qualitätsjournalismen? Eine (Re-)Definition flexibler Kriterien der Qualität alternativer journalistischer Berichterstattungsmuster

Janis Brinkmann

Zusammenfassung

Anknüpfend an aktuelle Diskurse über Qualität und mögliche (Fehl-)Leistungen im Journalismus sowie die in diesem Band aufgeworfenen Fragen zur Deutungsmacht von Sprache argumentiert der Beitrag für stärker ausdifferenzierte Kriterien zur Bewertung journalistischer Qualität und eine Erweiterung des in Forschung und Praxis immer noch recht eng am Informationsjournalismus orientierten Begriffs des „Qualitätsjournalismus". Nach einer öffentlichkeitstheoretischen Verortung nimmt der Beitrag alternative journalistische Berichterstattungsmuster wie den investigativen, narrativen oder konstruktiven Journalismus in den Blick und plädiert für ein flexibleres Set von Qualitätskriterien, die die spezifischen Charakteristika unterschiedlicher Journalismus-Konzepte bzw. „Journalismen" in der Praxis gegenstandadäquater abbilden können. Abschließend erörtert der Beitrag die Vor- und Nachteile einer solchen Perspektive auf journalistische Qualität.

1. Ausgangspunkt: Die „Entgrenzung" journalistischer Qualität

Diskussionen über journalistische Qualität entzünden sich in Medien, aber auch in der Öffentlichkeit gegenwärtig oft an der Berichterstattung über internationale Krisen und Konflikte wie der Corona-Pandemie oder dem Krieg in der Ukraine (vgl. Maurer et al. 2021; 2022), bei denen der Vorwurf einer einseitigen, verzerrten oder unausgewogenen Thematisierung und Bewertung schnell artikuliert und teilweise auch empirisch gestützt wird (vgl. Neuberger/Hohlfeld 2024). Der mediale und öffentliche Diskurs über die (Fehl-)Leistungen von Journalismus ist vor diesem Hintergrund „stark von den Schlagworten Vertrauens- oder Glaubwürdigkeitskrise bestimmt" (Arnold 2023: 93). Zwar zeigen Langzeiterhebungen, dass deutsche

Nutzer:innen (1) im internationalen Vergleich Medien viel Vertrauen entgegenbringen (vgl. Nielsen/Fletcher 2024) und (2) das Medienvertrauen in Deutschland seit 2008 vergleichsweise stabil geblieben ist (vgl. Quiring et al. 2024). Dennoch scheint eine subjektive Wahrnehmung diese Befunde zu konterkarieren. Spektakuläre, medial wie öffentlich diskutierte Fälle journalistischen Fehlverhaltens – von den erfundenen Reportagen Claas Relotius' im Spiegel (vgl. Niggemeier 2018) bis zu den „Lügen"- und „Framing"-Vorwürfen des Influencers Rezo gegen das investigative funk-Format STRG_F (vgl. Niggemeier 2024) – tragen dazu bei, „dass die Qualität im Journalismus immer häufiger und in immer breiteren Kreisen thematisiert wird" (Arnold 2023: 93).

Erschwert wird die Debatte über journalistische Qualität – stark vereinfacht die Frage: Was ist *guter* Journalismus? (vgl. Meier 2018: 239–240; Reineck 2024: 541; Brinkmann 2024: 251) – durch eine fortschreitende, keineswegs nur technologisch induzierte „Entgrenzung" des Journalismus (vgl. Loosen 2016). In der systemtheoretischen Journalismusforschung wird ein solches Verschwimmen der Grenzen des Journalismus und seine schwindende „Unterscheidbarkeit von anderen Kommunikationsformen" (ebd.) schon länger befundet und beobachtet. Wenn aber die Grenzen und damit die Funktionen des Journalismus zu anderen Teilen des Mediensystems zunehmend unscharf werden (z. B. durch *Native Ads* oder *Affiliate Links* zur Werbung, durch Praktiken des Influencer Marketings in sozialen Netzwerken, durch Haltungsjournalismus zum Aktivismus, durch politische Content Creator:innen zum *Campaigning* usw.), dann verschwimmen auch journalistische Qualitätskriterien.

Denn: Journalismus ist heute *fluid*. Nicht nur sein jahrzehntelang massenmedial manifestierter institutionell-organisatorischer *Kontext* hat sich vor dem Hintergrund eines als komplex wie krisenhaft wahrgenommenen Medienwandels teilweise liquidiert: „Die etablierten massenmedialen Strukturen der Produktion, Verteilung und Nutzung journalistischer Inhalte lösen sich im Zuge von ökonomischen, technologischen und sozialen Veränderungen der Digitalisierung zunehmend auf" (Buschow 2018a: 516; vgl. auch Meier/Neuberger 2016: 9–10). Ebenso verliert der post-industrielle Journalismus (vgl. unter anderem Anderson/Bell/Shirky 2014; Creech/Nadler 2018) in der Transformation journalistischer Arbeitsweisen spätestens seit der Jahrtausendwende „als fest umrissener, identifizierbarer Sinn- und Handlungszusammenhang deutlich an Konturen" (Weischenberg 2001: 77). Es ist also zunehmend durchlässiger geworden, was Journalismus selbst ist, beziehungsweise sein soll und welches *Handeln* gegen-

wärtig als „journalistisch" gelten kann (vgl. Buschow 2018b: 317). Damit verbunden sind erhebliche Folgen für die journalistische Praxis und Forschung, die sich mit Definitions- und Abgrenzungsproblemen konfrontiert sehen (vgl. vertiefend auch Brinkmann 2025a: 7ff.), aber eben auch für die Frage, was als guter bzw. „Qualitätsjournalismus" gelten kann. Gleichzeitig ‚verflüssigt' sich der Journalismus auch nach innen (vgl. Deuze/Witschge 2018), differenziert sich also zunehmend in spezielle Formen und Konzepte diverser Sub-Journalismen oder „New Journalisms" aus (vgl. Papacharissi 2015; Fowler-Watt/Jukes 2020; Loosen et al. 2020). Für den Qualitätsdiskurs in der Journalismusforschung, der noch immer von der Stephan Ruß-Mohl zugeschriebenen Pudding-Metapher (vgl. 1992: 85) illustriert wird, ist es also noch schwieriger geworden, den „verflüssigten" Journalismus an die metaphorische Wand zu nageln, um dessen Qualität zu bestimmen.

Dieser Beitrag unternimmt den Versuch, Berichterstattungsmuster als „Wertungsobjekte" (Reineck 2024: 554) journalistischer Qualität in den Blick zu nehmen und flexiblere Sets von Qualitätskriterien herzuleiten, die die sich *in praxi* ausdifferenzierenden Journalismen gegenstandadäquater bewerten können.

2. Theoretischer Rahmen: Öffentlichkeitstheorien und (alternative) journalistische Berichterstattungsmuster

Die deutschsprachige journalistische Qualitätsforschung ist bis heute vor allem von einem liberalen Öffentlichkeitskonzept geprägt, das als zentrale Qualitätsmerkmale die Vielfalt, Relevanz und Professionalität journalistischer Berichterstattung anlegt (vgl. Schatz/Schulz 1992; Jandura/Friedrich 2014: 352–357). Als Referenzgröße gilt noch immer der demokratietheoretisch fundierte Informations- und Nachrichtenjournalismus (vgl. Riedl 2024), der hohe Ansprüche an Neutralität, Ausgewogenheit und Objektivität formuliert (vgl. Mothes 2014; Post 2015; Neuberger 2017). Dass diese Perspektive angesichts eines sich ausdifferenzierenden Journalismus zu begrenzt sein könnte, hat der Autor zuletzt für Formen des subjektiven Journalismus angeregt (vgl. Brinkmann 2025b). Journalistische Qualitätskriterien sind eben keine „Eigenschaften, sondern Beobachter:innenkonstrukte und damit subjektiv" (Sehl/Eder/Kretzschmar 2022: 47; zur sozialen Konstruktion journalistischer Qualität vgl. auch Reineck 2018). Unter den Versuchen, den journalistischen Qualitätsbegriff auf alternative Journalismus-Konzepte (vgl. Wyss/Keel 2010) zu erweitern (vgl. unter anderem Engesser

2013; Habers 2016; Radü 2018) ist mit Blick auf das oben formulierte Ziel einer Re-Definition etablierter Qualitätskriterien für alternative Berichterstattungsmuster beziehungsweise Journalismen der Beitrag von Jandura und Friedrich (2012) besonders fruchtbar: Die Autor:innen fokussieren auf den Boulevardjournalismus als Referenz- und besonders exponierten Kontrapunkt des Informationsjournalismus für journalistische Qualitätsdiskussionen, indem sie ein partizipatives Öffentlichkeitsmodell zugrunde legen, dessen Leistungsanforderungen an eine boulevardeske Politikvermittlung sich von denen eines liberalen öffentlichkeitstheoretischen Modells stark unterscheiden (vgl. ebd.: 409): Partizipative, inklusive, emotionale und sogar unterhaltende oder solidarische Kriterien gewinnen im Zuge dieser „öffentlichkeitstheoretischen Neubewertung" an Einfluss, die Jandura und Friedrich jedoch nur als Startpunkt für eine Weiterentwicklung des journalistischen Qualitätsdiskurses verstehen:

> „Zusätzlich sollte die Wissenschaft aber auch unterhaltende Medienangebote auf ihr Leistungsvermögen hin untersuchen, alternative Teilpublika zu erreichen und anders politisierte Teilöffentlichkeiten zu erzeugen. Das partizipative Öffentlichkeitsmodell bietet einen Ansatzpunkt, die entsprechenden politischen Bezüge in diesen Medienangeboten präziser zu modellieren, als es mit dem liberalen Öffentlichkeitsmodell bislang möglich ist" (ebd.: 415).

So wie das partizipative das liberale Öffentlichkeitsmodell nicht ersetzt, sondern lediglich ergänzt, ergänzen auch vom Informations- und Nachrichtenjournalismus abweichende Berichterstattungsmuster das Funktions- und Leistungsspektrum des Journalismus (vgl. Meier 2018: 195; 2019). In diesem Sinne können auch die besonderen *Qualitäten* alternativer Berichterstattungsmuster stärker Gegenstand der Debatten um journalistische Qualität sein (vgl. Brinkmann 2025b) – insbesondere, da die Werturteile über Qualität im Journalismus keineswegs statisch sind, sondern immer relational erfolgen müssen (vgl. Brinkmann 2024: 254): „Sie werden von einem zeitlichen, räumlichen und sozialen Standort aus gefällt" (Reineck 2024: 544). Doch was guten bzw. hochwertigen Journalismus ausmacht – im Sinne seiner Leistungsfähigkeit, also seiner zum Beispiel ökonomischen, ethischen oder publizistischen *Performanz* (vgl. Brinkmann 2024: 229ff.) – hängt nicht nur von der jeweils gewählten theoretischen Perspektive ab (zur Übersicht über eine funktional-professionelle, eine werte- und kodexorientierte, eine markt- und publikumsorientierte Perspektive

oder eine integrative Perspektive vgl. Arnold 2023: 95ff.). Es ist ebenso abhängig vom jeweiligen journalistischen Programm und der darin eingewobenen journalistischen Praktiken. Denn spezifische Themen, Kanäle, Darstellungsformen und eben auch Berichterstattungsmuster beeinflussen als Dimensionen journalistischer Programme zur Wirklichkeitskonstruktion (vgl. Meier 2018: 202) die journalistischen Praktiken wie Recherche, Storytelling oder Präsentation und damit auch die Leistungspotenziale seiner Produkte:[1] Während ein Nachrichtenredakteur zum Beispiel mit einer möglichst aktuellen Meldung Informationen sachlich vermittelt, will eine Boulevard-Journalistin vielleicht mit einer möglichst emotionalen, exklusiven Geschichte die Aufmerksamkeit der Leser:innen erzwingen – und ein Investigativ-Journalist mit Hilfe hartnäckiger Recherche einen politischen Skandal aufdecken und darüber eine umfangreiche Story schreiben.

Die journalistischen Werteobjekte können also nicht bloß vertikal nach Mediensystemen, Medienorganisationen, Medienprodukten oder Medienakteur:innen differenziert werden, sondern auch horizontal nach journalistischen Genres – so verweist Reineck (2024: 544) auf die „spezifischen Qualitäten des Lokaljournalismus, Sportjournalismus, Wissenschaftsjournalismus oder Gesundheitsjournalismus" (vgl. hierzu auch Reineck 2018: 111–112) – oder eben nach Berichterstattungsmustern. Dieser vergleichsweise kleine Bereich der Journalismusforschung differenziert und typologisiert verschiedene Journalismus-Konzepte, zum Beispiel nach den damit verbundenen Intentionen, Rollenbildern, Arten der Faktenpräsentation und Formen der Recherche (vgl. Meier 2018: 196f.; 2019). Obwohl die oben beschriebene Ausdifferenzierung des Journalismus in der Praxis dazu geführt hat, das auch theoretische Konzepte wie das der Berichterstattungsmuster kleinteiliger werden (und zunehmend Sub-Muster ausbilden; vgl. Meier 2019: 105), fokussiert dieser Beitrag neben dem Informations- und Nachrichtenjournalismus jene fünf alternativen Berichterstattungsmuster, die sich über die vergangenen Jahre in der journalistischen Praxis (und damit auch in der Forschung) stabilisiert haben und deren unterschiedliche Charakteristika sich klar erkennbar gegen die Kriterien des Informations- und Nachrichtenjournalismus konturieren (vgl. Abb. 1).

1 Zum konzeptionellen Zusammenwirken journalistischer Programme, Praktiken und Leistungspotenziale vgl. das PPP-Modell bei Brinkmann (2024: 55–60; 2025a: 782–787). Zum letztlich nicht auflösbaren Zusammenhang von journalistischer Qualität zwischen Handeln und Produkten vgl. Meier (2018: 242f.).

Journalismus-Konzept/ Berichterstattungsmuster	Rollenbild	Intention	Faktenpräsentation	Recherche
Informationsjournalismus	Vermittler	„Realität" in Fakten abbilden	Neutrale Faktizität	Verlautbarung
Interpretativer Journalismus	Erklärer, Analyst, Überprüfer	Orientierung stiften (durch Einordnung, Verständlichkeit und Fakten)	Erläuterte, analytische Faktizität	Recherche von Interpretationen
Investigativer Journalismus	Wachhund	Kontrolle/Kritik/Machtmissbrauch aufdecken	Beweisführend, zugespitzt	Unorthodox (Whistleblower)
Anwaltschaftlicher Journalismus	Anwalt, Betroffener	Verständnis, Solidarität schaffen	„Betroffenheits-Faktizität"	Inoffizielle Quellen
Erzählerischer Journalismus	Erzähler	Wirklichkeit abseits von blanken Fakten erfassen (über Erfahrungen, Gefühle, Handlungen)	Erzählend als „Geschichte" („Story")	Recherche über lange Zeiträume abseits der Nachrichtenfaktoren
Konstruktiver Journalismus	Dialog-Organisator	Lösungen für Probleme	Lösungs- und forumsorientiert	Aktionen, „ganzheitliches Bild", Lösungen

Abbildung 1: Journalistische Berichterstattungsmuster und die sie jeweils prägenden Dimensionen (eigene Darstellung verkürzt nach Meier 2018; 2019)

Vor dem theoretischen Hintergrund wird zum Beispiel angesichts von journalistischen Intentionen, die „Wirklichkeit abseits von blanken Fakten erfassen" (erzählerischer Journalismus) oder „Solidarität schaffen" (anwaltschaftlicher Journalismus) wollen, schnell deutlich, dass diese alternativen Berichterstattungsmuster – ähnlich wie der Boulevardjournalismus – den normativ am Informations- und Nachrichtenjournalismus ausgerichteten, „klassischen Qualitätsanforderungen überhaupt nicht gerecht werden [*können*]" (Jandura/Friedrich 2012: 405). In einem partizipativen Öffentlichkeitsmodell können diese Berichterstattungsmuster oder Journalismen jedoch alternative, oft subjektivere Varianten des ‚objektiven' Informationsjournalismus darstellen, die dessen informative Leistungspotenziale (zum Beispiel durch Meinung, Investigation, Emotion, Intervention, Partizipation etc.) ergänzen (vgl. Brinkmann 2025b) und müssen nicht als Formen eines „Pseudojournalismus" (Hohlfeld 2003) aus dem Feld des Qualitätsjournalismus exkommuniziert werden. Nicht zuletzt an diesem Punkt zeigt sich der immense Einfluss sprachlicher Zuschreibungen und damit verbundener Signifikations- und Legitimationsregeln (vgl. Lünenborg 2012): Die exklusive Verwendung eines journalistischen Qualitätsbegriffs, der darauf zielt, „bestimmte Formen des Journalismus an die Peripherie zu drängen und das Prädikat ‚Qualität' für bestimmte Spielarten zu reservieren" (Reineck 2024: 549), unterstreicht die Notwendigkeit eines inklusiveren, flexibleren Katalogs von Qualitätskriterien.

Hierfür bietet es sich an, spezifische Kriterien aus verschiedenen journalistischen Berichterstatttungsmustern deduktiv abzuleiten. Jene Kriterien können bei einer späteren empirischen Untersuchung auch als Indikatoren operationalisiert werden.

3. Vorschlag für ein flexibleres Kernset journalistischer Qualitätskriterien

Für eine gegenstandadäquate Bewertung von Berichterstattungsmustern als prägende journalistische Programme – und damit auch das in ihnen praktizierte Handeln, die dadurch produzierten Inhalte sowie deren Leistungsfähigkeit (vgl. Brinkmann 2024: 55–60) – kann ein flexibleres Set an Qualitätskriterien abgeleitet und dafür ein „dekompositorischer" (vgl. Reineck 2024: 544) Ansatz in Anschlag gebracht werden. Ausgehend von einem praxistheoretischen Verständnis von Journalismus (vgl. Meier 2018: 14) als „redaktionell unabhängiges Selektieren, Recherchieren, Strukturieren, Präsentieren, Verifizieren und Publizieren aktueller, faktischer und

relevanter Informationen über Medien an die Öffentlichkeit" (Brinkmann 2024: 22) wird dafür votiert, den Qualitätsdiskurs und die hervorgebrachten Kriterien nicht als bloße Konstruktionen im Sinne eines radikalen Konstruktivismus vollständig aufzulösen (vgl. Pörksen 2016: 254–255), sondern ausgehend von als zentral wahrgenommenen und etablierten „Kernqualitäten" – die sich zudem kritisch von „Objektivitätskriterien" wie Wahrheit, Neutralität, Ausgewogenheit etc. abgrenzen – „Orientierungspunkte" für die Analyse und Reflexion journalistischer Leistungen zu (re-)formulieren. Ein solches „Kernset" von Qualitätskriterien, ohne die Journalismus kein Journalismus (mehr) wäre beziehungsweise ohne die „letztendlich jede Art von Journalismus in seinen jeweiligen Eigenheiten zu Qualitätsjournalismus und der Qualitätsbegriff weitgehend zur phrasenhaften Leerformel [würde]" (Arnold 2016: 558), kann im Sinne der oben genannten Definition aus den Kriterien *Aktualität*, *Relevanz*, *Faktizität* sowie *Independenz* bestehen (vgl. Brinkmann 2024: 256; Meier 2018: 14).

Um von diesem basalen Katalog ausgehend nun weitere, für die jeweiligen Berichterstattungsmuster adäquate Qualitätskriterien zu identifizieren und zu einem flexiblen Set zu addieren – auch auf die Gefahr, damit bloß weitere „additive Indikatorenkataloge" (Jandura/Friedrich 2012: 403) zu produzieren, die potenziell „uferlos", „zusammenhangslos", „unklar" und/oder theoriefern sein können (vgl. Reineck 2024: 547) – wird eine funktional-professionelle Perspektive auf Qualität eingenommen, die insbesondere angesichts der oben beschriebenen Ausdifferenzierung fruchtbar erscheint. Obwohl *Information* als funktionaler Kern von Journalismus verstanden werden kann (beziehungsweise Nicht-Information als Abgrenzung im systemtheoretischen Sinne), ist Information nicht geeignet, um Journalismen nach innen beziehungsweise untereinander auszudifferenzieren: Alle Berichterstattungsmuster informieren (sonst wären sie keine Journalismen mehr), aber alle informieren unterschiedlich: Die Leitfrage für die Auseinandersetzung mit journalistischer Qualität beziehungsweise seiner Performanz (McQuail 1992) kann hier also lauten: Wie *informiert* Journalismus? Gedanklich angelehnt an das Konzept der „X-Journalisms" (Loosen et al. 2020) und anknüpfend an die diversen Berichterstattungsmuster wie Nachrichtenjournalismus sowie interpretativer, investigativer, anwaltschaftlicher, erzählerischer oder konstruktiver Journalismus informieren diese Journalismen dann zum Beispiel *nachrichtlich*, *interpretativ*, *investigativ*, *anwaltschaftlich*, *narrativ* oder *konstruktiv* und ergänzen das journalistische Leistungsspektrum dadurch jeweils (vgl. bereits Brinkmann 2025b; vgl. Abbildung 2).

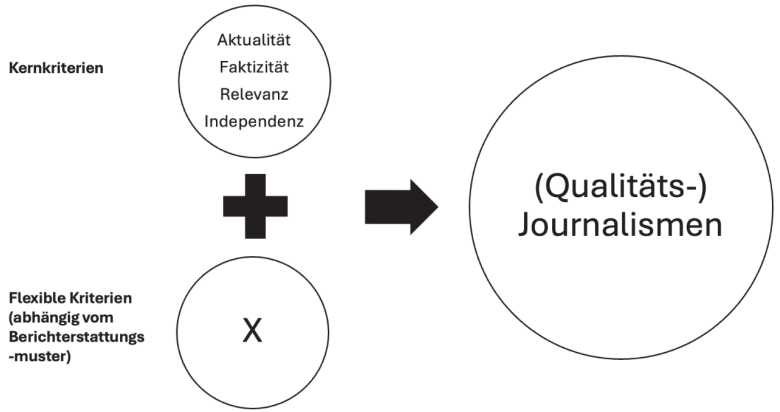

Abbildung 2: Ableitung von Kern- und flexiblen Kriterien für die Qualität unterschiedlicher Journalismen (eigene Darstellung)

Welche flexiblen Kriterien nun die Sets für eine gegenstandadäquatere Bewertung journalistischer Qualität ausdifferenzierter Berichterstattungsmuster jeweils ergänzen, kann zunächst deduktiv aus der rezenten Literatur zu diesen Journalismen (vgl. Brinkmann 2024: 77–84) gewonnen und anschließend für deren empirische Analyse operationalisiert werden.

So orientiert sich der traditionelle und in westlichen Mediensystemen noch immer dominante *Informations-* und *Nachrichtenjournalismus* eng am Ideal der Objektivität (vgl. Neuberger 2017), weswegen er auch als „objektive Berichterstattung" (Meier 2019: 106) bezeichnet wird. Dieses historisch zum Referenzpunkt des Journalismus gewachsene Berichterstattungsmuster, für das unter anderem die *ARD-Tagesschau* beispielhaft steht, vermittelt (aktuelle) Informationen so neutral, unparteiisch, passiv und faktisch wie möglich an das Publikum, wobei es Hintergründe und Interpretationen von Ereignissen weitgehend ausklammert. Als flexible Kriterien für die Bewertung der Qualität dieses journalistischen Berichterstattungsmuster bieten sich zum Beispiel *Neutralität/Überparteilichkeit*, *Ausgewogenheit* und *Vielfalt* (der Quellen und Perspektiven) an (vgl. Neuberger 1996; Schudson 2001; Schudson/Anderson 2009; Mothes 2016; Maras 2013; Meier 2019).

Im Gegensatz zum Informations- liefert der *interpretative Journalismus* nicht nur reine Fakten, sondern konsequent Kontext durch die Hinter-

gründe und Interpretationen von Ereignissen und Entwicklungen. Typisch sind Titelgeschichten in Nachrichtenmagazine wie dem *SPIEGEL*. Obwohl Nachrichten im Mittelpunkt stehen, geht es um deren Einordnung, wodurch die Grenze zum *Meinungsjournalismus*, der sich argumentativ und wertend zu Ereignissen positioniert (vgl. Degen 2004), verschwimmt. Entsprechend bieten sich als ergänzende Kriterien eines interpretativen Journalismus zum Beispiel *Meinung*/*Orientierung*, *Kontextualität*/*Einordnung*, *Verständlichkeit* und *Komplexitätsreduktion* an (vgl. Salgado/Strömbeck 2013; Fink/Schudson 2014; Olsson/Nord 2014; Brüggemann/Engesser 2014; Soontjes 2019).

Der *investigative Journalismus* stellt intensive, aufwändige Recherchen in den Mittelpunkt der journalistischen Arbeit. Dieses Berichterstattungsmuster begnügt sich nicht damit, Ereignisse zu berichten, einzuordnen oder zu bewerten, sondern soll Machenschaften aufdecken und übt damit die Kritik- und Kontrollfunktion als journalistischer *watchdog* besonders dezidiert aus (wie zum Beispiel der Rechecherverbund von *SZ*, *WDR* und *NDR* oder *correctiv*). Hier kommen ergänzende Qualitätskriterien wie zum Beispiel *Kritik*, *Exklusivität* und *Transparenz* (unter anderem der Quellen) in Frage (vgl. Lilienthal 2017; Haarkötter 2015; Ludwig 2014; Haller 2017; Cario 2006; Redelfs 1996; Hamilton 2016; de Burgh 2000).

Berichterstattungsmuster, die als *erzählerischer* beziehungsweise *narrativer Journalismus* deklariert werden, stellen konsequent journalistische Geschichten in den Mittelpunkt und machen das Thema über die Handlung der Akteure damit erlebnisstark. Neben Praktiken des *Storytelling* (vgl. Lampert/Wespe 2017) arbeitet der insbesondere in Reportagen praktizierte „Erzähljournalismus" (Haller 2020) mit konsequenter Personalisierung und Emotionalisierung und fokussiert eigenes, zumeist *subjektives* Erleben, Erfahren, Handeln und Fühlen. Beispiele hierfür sind unter anderem die großen Reportagen in den Magazinen der *Süddeutschen Zeitung* oder der *ZEIT*, aber auch junge Reportage-Formate wie *Y-Kollektiv* oder *STRG_F*. Narrativer Journalismus lässt sich ergänzend zum Beispiel durch *Emotionalität*, *Authentizität* oder *Narrativität* bewerten (vgl. van Krieken/Sanders 2017; 2019; Machill/Köhler/Waldhauser 2006; Neveu 2014; Shim 2014; Früh/Frey 2014; Köhler 2009; Köpke 2017; Tulloch 2014; Haller 2020).

Eine stark konturierte Opposition zum Informationsjournalismus nimmt auch der *anwaltschaftliche Journalismus* ein, der dessen Ideale neutraler, objektiver Berichterstattung zurückweist und sich stattdessen jener Themen annimmt, die für gesellschaftliche Minderheiten oder ‚machtlose' Mehrheiten Relevanz haben, womit er Betroffenheit und Solidarität erzeugen will

(vgl. Altmeppen 2016) und zugleich eine „Gegenöffentlichkeit" zu jenen „Mainstream-Medien" bildet, die im objektiven Journalismus verortet werden. Ergänzende Qualitätskriterien eines anwaltschaftlichen Journalismus sind zum Beispiel *Solidarität*, *Engagement* und *Parteilichkeit* (vgl. Altmeppen 2016; Harcup 2006; Fisher 2016; Ganella 2021; Ginosar/Reich 2020; Russell 2016).

Im Gegensatz zum reinen Informationsjournalismus, dessen klassische W-Fragen er bewusst um Lösungen und Perspektiven erweitert, setzt der *konstruktive Journalismus* auf lösungsorientierte Ansätze (zum Beispiel das Onlineangebot *Perspective Daily* oder „Plan B" des *ZDF*): Er endet nicht bei der Beschreibung von Ereignissen, Missständen und Skandalen, sondern richtet den Blick in die Zukunft. Als ergänzende Qualitätskriterien kommen zum Beispiel *Lösungsfokus*, *Perspektivenreichtum* und *Dialog* in Frage (vgl. Bonn Institute 2023) ebenso wie *Nutzwert* und *Reflexivität* (vgl. Kramp/Weichert 2020; Ahva/Hautagangas 2018; McIntyre 2019).

Auf diese Weise lassen sich deduktiv flexible Sets für die Bewertung der verschiedeneren Leistungspotenziale unterschiedlicher Berichterstattungsmuster ableiten (vgl. Abb. 3) – selbstverständlich auch für weitere, hier aus Platzgründen nicht berücksichtige Journalismen wie Partizipativer Journalismus oder Datenjournalismus ebenso wie für hybride Formen (was allerdings die Komplexität erhöht und potenziell neue Unschärfen produziert). Zudem können auch technologische oder distributive Besonderheiten von Journalismen berücksichtigt werden, zum Beispiel in Bezug auf Social Media wie YouTube oder Instagram (vgl. Hermida/Mellado 2020; Sehl/Eder/Kretzschmar 2022).

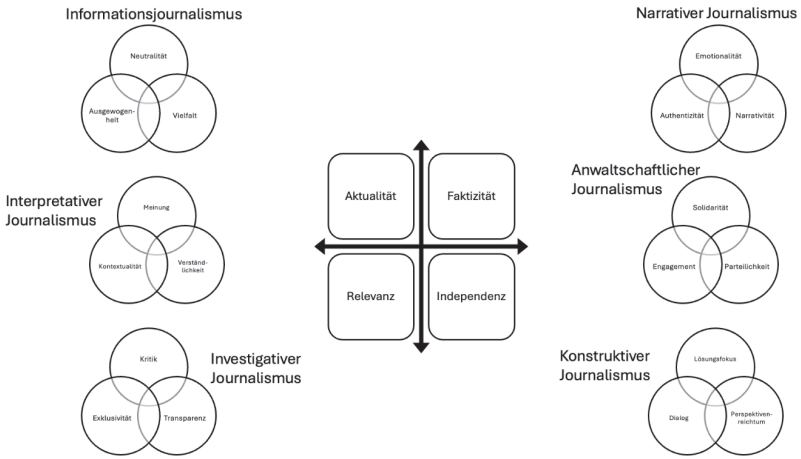

Abbildung 3: Kernset journalistischer Qualitätskriterien und die spezifischen Berichterstattungsmuster ergänzende Kriterien

Um diese flexibleren Sets journalistischer Qualitätskriterien für die empirische Analyse und die damit einhergehende Bewertung journalistischer Praktiken und Produkte nutzbar zu machen, müssen die Kriterien als Variablen operationalisiert werden, zum Beispiel über Indikatoren. Jandura und Friedrich (2012) schlagen dafür eine Systematisierung von Reinemann, Stanyer, Scherr und Legnante (2011) vor, um beispielsweise Boulevardjournalismus anhand der fünf Dimensionen Inhalt, Nachrichtenproduktion, Kontextualisierung, Wirkpotenzial und Stil zu erfassen und „aus dem partizipativen Öffentlichkeitsmodell theoretisch begründete Indikatoren boulevardesker Politikberichterstattung abzuleiten" (Jandura/Friedrich 2012: 409). Alternativ hat der Autor am Beispiel eines subjektiven Journalismus, der ein primär narratives Berichterstattungsmuster mit investigativen, interpretativen und partizipativen Elementen zu einer modernisierten Form des „New Journalism" hybridisiert (Brinkmann 2023: 22–24; 2025b), aufgezeigt, wie die Integration spezifisch subjektiver Kriterien zu einer gegenstandsadäquateren Bewertung eines solchen Journalismus führt, der die traditionellen Kriterien des ‚objektiven' Informationsjournalismus gar nicht erfüllen *kann* (und auch nicht erfüllen *will*): So waren beispielsweise bei den Presenter-Reportagen des öffentlich-rechtlichen Content-Netzwerks funk als subjektiv-journalistische Formate für junge Zielgruppen ergänzende ‚subjektive' Kriterien wie Authentizität (90,6 Prozent), Partizipativität

(82,9 Prozent), Emotionalität und Exklusivität (beide je 78,1 Prozent) sowie Narrativität (69,5 Prozent) mindestens „stark" ausgeprägt, während zentrale ‚objektive' Kriterien wie Relevanz oder Vielfalt in der Mehrheit der Beiträge nicht stark ausgeprägt waren (vgl. Brinkmann 2025b).[2] Dass subjektive Formate wie Y-Kollektiv oder STRG_F dennoch als neue Formen eines Qualitätsjournalismus verstanden werden können (vgl. Brinkmann 2023: 102ff.; 2025b) und keineswegs pauschal als „Pseudojournalismus" (vgl. Brodkorb 2023) geschmäht werden müssen, ist eine Erkenntnis, die aus der Erweiterung des journalistischen Qualitätsbegriff und seiner Kriterien resultiert.

4. Vorschlag Ausblick: Theoretische und praktische Implikationen für Journalistik und (Qualitäts-)Journalismus

Eine solche erweiterte Perspektive auf (Qualitäts-)Journalismus und seine Kriterien bringt sowohl Vor- als auch Nachteile für die Diskussion in Praxis und Forschung.

Einerseits können solche flexibleren Sets journalistischer Qualitätskriterien potenziell (1) die normativen Grundlagen von (Medien-)Kritik benennen und ausdifferenzieren (zum Beispiel traditioneller vs. experimenteller Journalismus), (2) damit die Kriterien und Praktiken journalistischer Arbeit transparenter und ggf. auch glaubwürdiger machen (zum Beispiel im Sinne von *Media Accountability*), (3) zu einer Versachlichung und Differenzierung teils pauschaler und populistisch-ideologischer Medienkritik im Diskurs über journalistische Qualität beitragen, (4) in Form von Handlungsempfehlungen in der journalistischen Praxis als Qualitätsmanagement beziehungsweise Praxistransfer umgesetzt werden (zum Beispiel durch Stakeholder wie Aufsichtsgremien, Redaktionen oder einzelne Journalist:innen), (5) durch (nicht-)Standardisierte Inhaltsanalysen Studien zu hybriden Journalismen anleiten, die den Trend in der Qualitätsforschung, vor allem publikumsbezogene (Befragungs-)Studien durchzuführen, methodisch ergänzen sowie (6) den starren Begriff des „Qualitätsjournalismus" zu einer Vielzahl von Qualitätsjournalismen aufbrechen und damit auch alternative Formen des Journalismus legitimieren.

2 Für eine kritische Auseinandersetzung und einen alternativen konzeptionellen und methodischen Ansatz zur Erforschung eines subjektiven Journalismus und seiner Qualitätsdimensionen vgl. Neuberger/Mayer 2026.

Andererseits kann diese Erweiterung des journalistischen Qualitätsbegriffs und seiner Kriterien ebenso (1) das Verständnis von Journalismus weiter verschwimmen lassen („*anything goes*") und zur Sinnentleerung beziehungsweise Inflation des Begriffs „Qualitätsjournalismus" führen, (2) potenziell „uferlose" und „zusammenhangslose" Indikatorenkataloge voller subjektivierter Adjektive produzieren (vgl. Reineck 2018: 7), (3) seine methodischen Schwächen (zum Beispiel subjektive, qualitative Entscheidungen über „Konstruktionen", wenig trennscharfe Indikatoren, begrenzte Informationen der Codier:innen) und Limitationen (explorativ, grobe Skalen, nur Näherungswerte) nicht hinreichend reflektiert und bei der Darstellung der Ergebnisse nicht erkennbar mitführen. Zudem droht (4) diese Perspektive durch die intendierte Passgenauigkeit der Qualitätskriterien für den jeweiligen Journalismus blind für dessen mögliche Schwächen zu werden (da vor allem Stärken empirisch gemessen werden).

Angesichts dieser Potenziale und Herausforderungen erscheint es umso nötiger, die Weiterentwicklung journalistischer Qualitätskriterien insbesondere aus medienethischer Perspektive zu voranzutreiben und kritisch zu begleiten.

Literatur

Ahva, Laura / Hautakangas, Mikko (2018): Why do we suddenly talk so much about constructiveness?, in: Journalism Practice 12 (6), S. 657–661.

Altmeppen, Klaus-Dieter (2016): Anwaltschaftlicher Journalismus, in: Jessica Heesen (Hg.), Handbuch Medien- und Informationsethik, Wiesbaden, S. 132–137.

Anderson, Chris, W. / Bell, Emily / Shirky, Clay (2014): Post Industrial Journalism. Adapting to the Present, New York.

Arnold, Klaus (2023): Qualität im Journalismus, in: Klaus Meier / Christoph Neuberger (Hg.), Journalismusforschung. Stand und Perspektiven, Baden-Baden, S. 93–110.

Bonn Institute (2023): Was ist konstruktiver Journalismus? (online unter: https://www.bonn-institute.org/was-ist-konstruktiver-journalismus – letzter Zugriff: 15.5.2025).

Brinkmann, Janis (2023): Journalistische Grenzgänger. Wie die Reportage-Formate von funk Wirklichkeit konstruieren (= Arbeitsheft 111 der Otto-Brenner-Stiftung), Frankfurt am Main.

Brinkmann, Janis (2024): Journalismus. Eine praktische Einführung, 2. Aufl., Baden-Baden.

Brinkmann, Janis (2025a): Subjektiver Journalismus. Theorie, Konzept, Praxis, Wiesbaden.

Brinkmann, Janis (2025b): Authentisch, emotional, partizipativ: Neue Qualitätskriterien eines neuen Journalismus, in: Vanessa Kokoschka et al. (Hg.), Nachhaltigkeit in der Medienkommunikation. Ethische Anforderungen und praktische Lösungsansätze, Baden-Baden, S. 133–152.

Brodkorb, Mathias (2023): 45 Millionen Euro pro Jahr für Pseudo-Journalismus, in: Cicero.de, 4. Juni 2023.

Brüggemann, Michael / Engesser, Sven (2014): Between Consensus and Denial: Climate Journalists as Interpretive Community, in: Science Communication 36 (4), S. 399–427.

Buschow, Christopher (2018a): Journalistik praxistheoretisch betreiben. Impulse für ein dynamisches Verständnis des Journalismus im Kontext seiner Neuordnung, in: Publizistik 63, S. 513–534.

Buschow, Christopher (2018b): Die Neuordnung des Journalismus. Eine Studie zur Gründung neuer Medienorganisationen, Wiesbaden.

Cario, Ingmar (2006): Die Deutschland-Ermittler: Investigativer Journalismus und die Methoden der Macher, Münster.

Creech, Brian / Nadler, Anthony M. (2018): Post-industrial fog: Reconsidering innovation in visions of journalism's future, in: Journalism 19 (2), S. 182–199.

de Burgh, Hugo (2000): Investigative Journalism, London.

Degen, Matthias (2004): Mut zur Meinung. Genres und Selbstsichten von Meinungsjournalisten, Wiesbaden.

Deuze, Mark / Witschge, Tamara (2018): Beyond Journalism. Theorizing the transformation of journalism, in: Journalism 19 (2), S. 165–181.

Engesser, Sven (2013). Die Qualität des Partizipativen Journalismus im Web: Bausteine für ein integratives theoretisches Konzept und eine explanative empirische Analyse, Wiesbaden.

Fink, Katherine / Schudson, Michael (2014): The rise of contextual journalism, 1950s-2000s, in: Journalism 15 (1), S. 3–20.

Fisher, Caroline (2016): The advocacy continuum: Towards a theory of advocacy in journalism, in: Journalism 17 (6), S. 711–726.

Fowler-Watt, Karen / Jukes, Stephen (2020): New Journalisms. Rethinking Practice, Theory and Pedagogy, London.

Früh, Werner / Frey, Felix (2014). Narration und Storytelling: Theorie und empirische Befunde, Köln.

Ganella, Gino (2021): Journalistic Power: Constructing the "Truth" and the Economics of Objectivity, in: Journalism Practice 17 (2), S. 209–225.

Ginosar, Avshalom / Reich, Zvi (2020): Obsessive–Activist Journalists: A New Model of Journalism?, in: Journalism Practice 16 (4), S. 660–680.

Haarkötter, Hektor (2015): Die Kunst der Recherche, Konstanz.

Habers, Frank (2016): Time to Engage. De Correspondent's redefinition of journalistic quality, in: Digital Journalism 4 (4), S. 494–511.

Haller, Michael (2020): Die Reportage. Theorie und Praxis des Erzähljournalismus, Köln.

Hamilton, James T. (2016): Democracy's Detectives: The Economics of Investigative Journalism, Cambridge.

Harcup, Tony (2006): "I'm Doing this to Change the World": Journalism in alternative and mainstream media, in: Journalism Studies 6 (3), S. 361–374.

Hermida, Alfred / Mellado, Claudia (2020): Dimensions of Social Media Logics: Mapping Forms of Journalistic Norms and Practices on Twitter and Instagram, in: Digital Journalism 8 (7), S. 864–884.

Hohlfeld, Ralf (2003): Vom Informations- zum Pseudojournalismus. Berichterstattungsmuster im Wandel, in: Communicatio Socialis 36 (3), S. 223–243.

Jandura, Olaf / Friedrich, Katja (2014): The quality of political media coverage, in: Carsten Reinemann (Hrsg.), Political Communication, Berlin, Boston, S. 351–374.

Jandura, Olaf / Friedrich, Katja (2012): Politikvermittlung durch Boulevardjournalismus Eine öffentlichkeitstheoretische Neubewertung, in: Publizistik 57, S. 403–417.

Köhler, Sebastian (2009): Die Nachrichtenerzähler: Zu Theorie und Praxis nachhaltiger Narrativität im TV-Journalismus, Baden-Baden.

Köpke, Wilfried (2017): Narrativer Fernsehjournalismus: rezeptions- und kommunikatorbezogene Begründung einer journalistischen Neuorientierung, in: Annika Schach (Hg.), Storytelling. Geschichten in Text, Bild und Film, Wiesbaden, S. 193–203.

Kramp, Leif / Weichert, Stephan (2020): Nachrichten mit Perspektive. Lösungsorientierter und konstruktiver Journalismus in Deutschland (= Arbeitsheft 101 der Otto-Brenner-Stiftung), Frankfurt am Main.

Lampert, Marie / Wespe, Rolf (2017): Storytelling für Journalisten. Wie baue ich eine gute Geschichte? 4. völlig überarb. Aufl., Köln.

Lilienthal, Volker (2017): Recherchejournalismus für das Gemeinwohl. Correctiv – eine Journalismusorganisation neuen Typs in der Entwicklung, in: M&K Medien- & Kommunikationswissenschaft 65 (4), S. 659–681.

Loosen, Wiebke (2016): Journalismus als (ent-)differenziertes Phänomen, in: Martin Löffelholz / Liane Rothenberger (Hg.), Handbuch Journalismustheorien, Wiesbaden, S. 177–190.

Loosen, Wiebke et al. (2020): ‚X Journalism'. Exploring journalism's diverse meanings through the names we give it, in: Journalism 23 (1), S. 39–58.

Ludwig, Johannes (2014): Investigatives Recherchieren, Köln.

Lünenborg, Margreth (2012): Qualität in der Krise?, in: Aus Politik und Zeitgeschichte 62 (29–31), S. 3–8.

Machill, Marcel / Köhler, Sebatian / Waldhauser, Markus (2006): Narrative Fernsehnachrichten: Ein Experiment zur Innovation journalistischer Darstellungsformen, in: Publizistik 51, S. 479–497.

Maras, Steven (2013): Objectivity in Journalism, Cambridge.

Maurer, Marcus / Haßler, Jörg / Jost, Pablo (2022): Die Qualität der Medienberichterstattung über den Ukraine-Krieg. Forschungsbericht zu ersten Befunden (= Arbeitsheft der Otto-Brenner-Stiftung), Frankfurt am Main.

Maurer, Marcus / Reinemann, Carsten / Kruschinksi, Simon (2021): Eine empirische Studie zur Qualität der journalistischen Berichterstattung über die Corona-Pandemie (= Studie der Rudolf Augstein Stiftung).

McIntyre, Karen (2019): Solutions Journalism: The Effects of Including Solution Information in News Stories About Social Problems, in: Journalism Practice 13 (8), S. 1029–1033.

McQuail, Denis (1992): Media Performance. Mass Communication and the Public Interest, London.

Meier, Klaus (2018): Journalistik, Konstanz.

Meier, Klaus (2019): Berichterstattungsmuster als Strategien der Komplexitätsreduktion, in: Beatrice Dernbach / Alexander Godulla / Annika Sehl (Hg.), Komplexität im Journalismus, Wiesbaden, S. 101–116.

Meier, Klaus / Neuberger, Christoph (2016): Einführung: Stand und Perspektiven der Journalismusforschung, in: Klaus Meier / Christoph Neuberger (Hg.), Journalismusforschung. Handbuch für Wissenschaft und Studium, Baden-Baden, S. 7–19.

Mothes, Cornelia (2014): Objektivität als professionelles Abgrenzungskriterium im Journalismus. Eine dissonanztheoretische Studie zum Informationsverhalten von Journalisten und Nicht-Journalisten, Baden-Baden.

Mothes, Cornelia (2016): Biased Objectivity: An Experiment on Information Preferences of Journalists and Citizens, in: Journals and Mass Communication Quaterly 94 (4), S. 1073–1095.

Neuberger, Christoph (1996): Journalismus als Problembearbeitung. Objektivität und Relevanz in der öffentlichen Kommunikation, Konstanz.

Neuberger, Christoph (2017): Journalistische Objektivität. Vorschlag für einen pragmatischen Theorierahmen, in: Uwe Hasebrink et al. (Hg.): Themenheft „Konstruktivismus in der Kommunikationswissenschaft. M&K Medien- & Kommunikationswissenschaft 65 (2), S. 406–431.

Neuberger, Christoph / Hohlfeld, Ralf (2024): Der russische Angriffskrieg gegen die Ukraine in den deutschen Medien: Kritik des Maßstabs „ausgewogene Bewertung" in Inhaltsanalysen, in: Publizistik 69, S. 455–493.

Neuberger, Christoph / Mayer, Anna-Theresa (2026): Was ist subjektiver Journalismus? Bedeutung und Relevanz für die Erfüllung des öffentlich-rechtlichen Auftrags. Eine Analyse am Beispiel des ARD/ZDF-Content-Netzwerks funk, in: Media Perspektiven, 3/2026, S. 1–24.

Neveu, Erik (2014): Revisiting Narrative Journalism as One of The Futures of Journalism, in: Journalism Studies 15 (5), S. 533–542.

Nielsen, Rasmus Kleis / Fletcher, Richard (2024): Public perspectives on trust in news, in: Nic Newman (Hg.), Reuters Digital News Report 2024, Oxford, S. 34–38.

Niggemeier, Stefan (2018): Der Fall Relotius: Der „Spiegel" und die gefährliche Kultur des Geschichten-Erzählens, in: Übermedien, 19. Dezember 2018 (online unter: https://uebermedien.de/33962/der-spiegel-und-die-gefaehrliche-kultur-des-geschichten-erzaehlens/ – letzter Zugriff: 15.5.2025).

Niggemeier, Stefan (2024): Was man aus dem Streit zwischen Rezo und „Strg_F" über guten Journalismus lernen kann, in: Übermedien, 16. Januar 2024 (online unter: https://uebermedien.de/91390/was-man-aus-dem-streit-zwischen-rezo-und-strg_f-ueber-guten-journalismus-lernen-kann/ – letzter Zugriff: 15.5.2025).

Olsson, Eva-Karin / Nord, Lars W. (2014): Paving the way for crisis exploitation: The role of journalistic styles and standards, in: Journalism 16 (3), S. 341–358.

Papacharissi, Zivi (2015): Toward New Journalism(s). Affective news, hybridity, and liminal spaces, in: Journalism Studies 16 (1), S, 27–40.

Pörksen, Bernhard (2016): Journalismus als Wirklichkeitskonstruktion, in: Martin Löffelholz / Liane Rothenberger (Hg.), Handbuch Journalismustheorien, Wiesbaden, S. 249–261.

Post, Senja (2015): Scientific objectivity in journalism? How journalists and academics define objectivity, assess its attainability, and rate its desirability, in: Journalism 16 (6), S. 730–749.

Quiring, Oliver et al. (2024): Zurück zum Niveau vor der Pandemie – Konsolidierung von Vertrauen und Misstrauen. Mainzer Langzeitstudie Medienvertrauen 2023, in: Media Perspektiven 9, S. 1–14.

Radü, Jens (2018): New Digital Storytelling. Anspruch, Nutzung und Qualität von Multimedia-Geschichten, Baden-Baden.

Redelfs, Manfred (1996): Investigative Reporting in den USA. Strukturen eines Journalismus der Machtkontrolle, Wiesbaden.

Reineck, Dennis (2018): Die soziale Konstruktion journalistischer Qualität. Fachdiskurs, Theorie und Empirie, Köln.

Reineck, Dennis (2024): Qualität des Journalismus, in: Martin Löffelholz / Diane Rothenberger (Hg.), Handbuch Journalismustheorien, Wiesbaden, S. 541–554.

Reinemann, Carsten et al. (2011): From hard vs. soft to a multi-dimensional approach. Towards a standardized definition and measurement of different types of news, in: Journalism 13 (2), S. 1–19.

Riedl, Andreas A. (2024): Nachrichtenqualität als journalistischer Prozess. Demokratietheoretisch fundierte Performanz zwischen Wollen, Sollen und Können, Köln.

Russell, Adrianne (2016): Journalism as Activism: Recoding Media Power, New Jersey.

Ruß-Mohl, Stephan (1992): Am eigenen Schopfe. Qualitätssicherung im Journalismus – Grundfragen, Ansätze, Näherungsversuche, in: Publizistik 37 (1/1992), S. 83–96.

Salgado, Susana / Strömbeck, Jesper (2011): Interpretive journalism: A review of concepts, operationalizations and key findings, in: Journalism 13 (2), S. 144–161.

Schatz, Heribert / Schulz, Winfried (1992). Qualität von Fernsehprogrammen. Kriterien und Methoden zur Beurteilung von Programmqualität im dualen Fernsehsystem, in: Media Perspektiven (11/1992), S. 690–712.

Schudson, Michael (2001): The objectivity norm in American journalism, in: Journalism 2 (2), S. 149–170.

Schudson, Michael / Anderson, Chris W. (2009): Objectivity, Professionalism, and Truth Seeking in Journalism, in: Karin Wahl-Jorgensen / Thomas Hanitzsch (Hg.), The Handbook of Journalism Studies, New York, S. 88–101.

Sehl, Annika / Eder, Maximilian / Kretzschmar, Sonja (2022): Journalismus auf Instagram Qualität neu definiert?, in: Jonas Schützeneder / Michael Graßl (Hg.), Journalismus und Instagram. Analysen, Strategien, Perspektiven aus Wissenschaft und Praxis, Wiesbaden, S. 45–58.

Shim, Hoon (2014): Narrative journalism in the contemporary newsroom: The rise of new paradigm in news format?, in: Narrative Inquiry 24 (2), S. 77–95.

Soontjes, Karolin (2019): The Rise of Interpretive Journalism. Belgian newspaper coverage, 1985–2014, in: Journalism Studies 20 (7), S. 952–971.

Tulloch, John (2014): Ethics, trust and the first person in the narration of long-form journalism, in: Journalism 15 (5), S. 629–638.

van Krieken, Kobie / Sanders, Jose (2017): Framing narrative journalism as a new genre: A case study of the Netherlands, in: Journalism 18 (10), S. 1364–1380.

van Krieken, Kobie / Sanders, Jose (2019): What is narrative journalism? A systematic review and an empirical agenda, in: Journalism 22 (6), S. 1393–1412.

Weischenberg, Siegfried (2001): Das Ende einer Ära? Aktuelle Beobachtungen zum Studium des künftigen Journalismus, in: Hans J. Kleinsteuber (Hg.), Aktuelle Medientrends in den USA. Journalismus, politische Kommunikation und Medien im Zeitalter der Digitalisierung, Wiesbaden.

Wyss, Vinzenz / Keel Guido (2010): Journalismusforschung, in: Heinz Bonfadelli / Otfried Jarren / Gabriele Siegert (Hg.), Einführung in die Publizistikwissenschaft, Bern, S. 337–378.

Verzeichnis der Autorinnen und Autoren

Dr. Mario Anastasiadis, Fachbereich Sozialpolitik und Soziale Sicherung, Hochschule Bonn-Rhein-Sieg

Prof. Dr. Stefanie Averbeck-Lietz, Institut für Politik- und Kommunikationswissenschaft (IPK), Universität Greifswald

Dr. Inga Bones, Department für Philosophie, Karlsruher Institut für Technologie

Prof. Dr. Janis Brinkmann, Institut für Kommunikationsmanagement, Hochschule Osnabrück

Prof. Dr. Bernhard Debatin, E.W. Scripps School of Journalism, Ohio University, Athens

Prof. Dr. Beatrice Dernbach, Forschungsprofessur für Nachhaltigkeits- und Wissenschaftskommunikation, Technische Hochschule Nürnberg

Dr. Henning Eichler, Hochschule für Technik, Wirtschaft und Kultur, Leipzig

Prof. Dr. Miriam Goetz, Professur für Medienmanagement, IST-Hochschule für Management, Düsseldorf

Prof. Dr. Petra Grimm, Hochschule der Medien, Stuttgart

Prof. Dr. Hektor Haarkötter, Fachbereich Sozialpolitik und Soziale Sicherung, Hochschule Bonn-Rhein-Sieg

Jana Hecktor, M.A., Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen

Prof. Dr. Jessica Heesen, Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen

Prof. Dr. Brigitte Huber, Professur für Marketing, IU Internationale Hochschule, München

Vanessa Kokoschka, M.Sc., Forschungsgruppe Human-Computer Interaction & Visual Analytics, Hochschule Darmstadt

Dr. Theresa Krampe, Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen

Susanne Kuhnert, Hochschule der Medien, Stuttgart

Prof. Dr. Julia Levasier, Fakultät für angewandte Geistes- und Naturwissenschaften, Technische Hochschule Augsburg

Prof. Dr. Michael Litschka, Forschungsgruppe Media Business, Institute for Creative Media Technologies, UAS St. Pölten

Laura Martena, M.A., Akademie für Politische Bildung

Dr. Jörg-Uwe Nieland, Institut für Kommunikationswissenschaft, Universität Münster

Prof. Dr. Dr. Matthias O. Rath, Institut für Philosophie, Pädagogische Hochschule Ludwigsburg

Prof. Dr. Christian Schicha, Institut für Theater- und Medienwissenschaft, Friedrich-Alexander-Universität Erlangen-Nürnberg

Marcel Schlegel, Hochschule der Medien, Stuttgart

Dr. Ingrid Stapf, Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen

Vanessa Wimmers, IST-Hochschule für Management, Düsseldorf