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