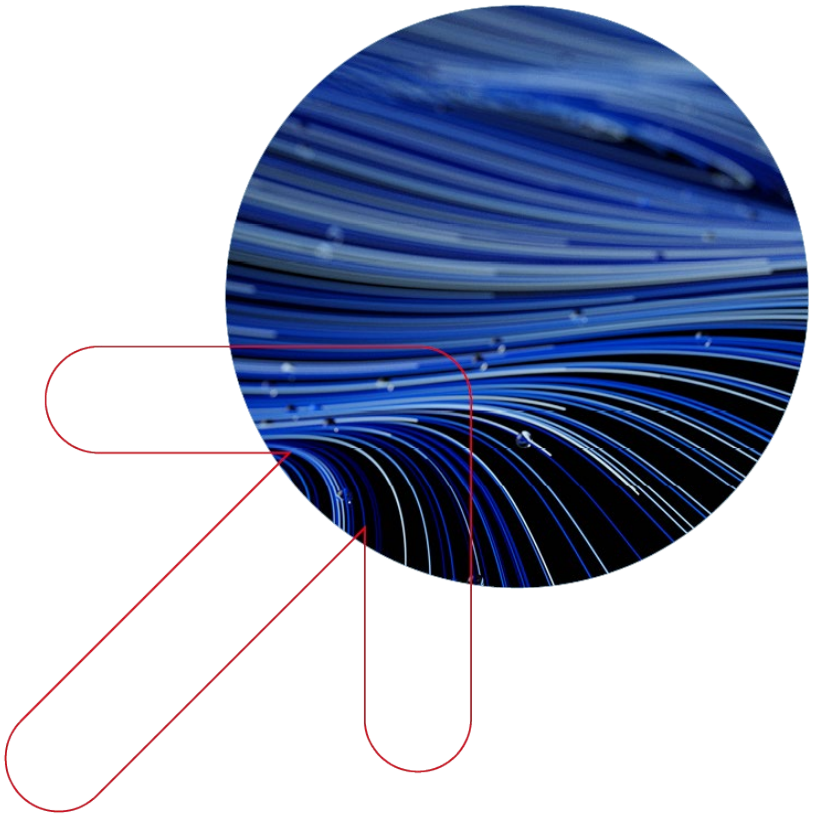
WIK • Diskussionsbeitrag

Nr. 546



Verbreitung und Auswirkungen von Desinformation im Kontext des DSA & Online-Plattformen

Autoren:
Dr. Nico Steffen
Ing. Peter Kroon
Dr. Lukas Wiewiorra

Bad Honnef, Dezember 2025

WIK
Wissenschaftliches Institut
für Infrastruktur und
Kommunikationsdienste

Impressum

WIK Wissenschaftliches Institut für
Infrastruktur und Kommunikationsdienste GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Deutschland
Tel.: +49 2224 9225-0
Fax: +49 2224 9225-63
E-Mail: info@wik.org
www.wik.org

Vertretungs- und zeichnungsberechtigte Personen

Geschäftsführung	Dr. Cara Schwarz-Schilling (Vorsitzende der Geschäftsführung, Direktorin)
	Alex Kalevi Dieke (Kaufmännischer Geschäftsführer)
Prokuristen	Prof. Dr. Bernd Sörries
	Dr. Christian Wernick
	Dr. Lukas Wiewiorra
Vorsitzender des Aufsichtsrates	Dr. Thomas Solbach
Handelsregister	Amtsgericht Siegburg, HRB 7225
Steuer-Nr.	222/5751/0722
Umsatzsteueridentifikations-Nr.	DE 123 383 795

Stand: Januar 2025

ISSN 1865-8997

Weitere Diskussionsbeiträge finden Sie hier:
https://www.wik.org/veroeffentlichungen/diskussionsbeitraege

In den vom WIK herausgegebenen Diskussionsbeiträgen erscheinen in loser Folge Aufsätze und Vorträge von Mitarbeitern des Instituts sowie ausgewählte Zwischen- und Abschlussberichte von durchgeführten Forschungsprojekten. Mit der Herausgabe dieser Reihe bezweckt das WIK, über seine Tätigkeit zu informieren, Diskussionsanstöße zu geben, aber auch Anregungen von außen zu empfangen. Kritik und Kommentare sind deshalb jederzeit willkommen. Die in den verschiedenen Beiträgen zum Ausdruck kommenden Ansichten geben ausschließlich die Meinung der jeweiligen Autoren wieder.

Inhaltsverzeichnis

1 Einleitung

Die Analyse der komplexen Dynamiken von Desinformation und Online-Schäden geht von der Feststellung aus, dass dieses Phänomen eine weitreichende gesellschaftliche, wirtschaftliche und sicherheitspolitische Relevanz besitzt. Die Betrachtung erstreckt sich dabei über rein politische Inhalte hinaus und schließt ebenso gesundheitliche Fragen sowie kommerzielle Bereiche, wie beispielsweise Investment-Scams und Betrugsmaschen, mit ein. Die Debatte um die effektive Bewältigung von Online-Schäden sollte sich über die reine Inhaltsmoderation hinaus auf Nutzerverhalten und systemische Faktoren beziehen, da Schäden auch ohne explizite Verletzung von Inhaltsstandards entstehen können. In diesem Kontext ist eine wichtige Unterscheidung zwischen allgemeiner Misinformation und koordinierten Einflussoperationen zu treffen, bei denen es sich oft um staatlich unterstützte Aktivitäten handelt.

Ziel der vorliegenden Studie ist es, eine Synthese des aktuellen Stands bezüglich der Verbreitung, Erkennung und Bekämpfung von Desinformation zu erstellen und darauf aufbauend priorisierte, systemische Maßnahmen abzuleiten. Die wissenschaftliche Auseinandersetzung mit dem Thema erfordert zunächst eine klare Abgrenzung der Begrifflichkeiten, wobei die Forschung traditionell nach dem Wahrheitsgehalt und der Schädigungsabsicht differenziert. Neben der Misinformation (falsch, ohne Absicht) und der Desinformation (falsch, mit Absicht) schließt die Analyse auch die Malinformation ein, also die absichtlich schädigende Verbreitung von echten Informationen, oft durch Dekontextualisierung. Für die analytische Praxis wählt die Studie einen breiteren, pragmatischen Arbeitsbegriff, da der Fokus auf den systemischen Risiken der Verbreitung liegt und nicht starr von der Intentionalität oder dem genauen Falschheitsgrad abhängen soll. Eine zentrale analytische Hürde in diesem Feld bildet der begrenzte empirische Kenntnisstand zu Kausalität und Exposition von Nutzern gegenüber schädlichen Inhalten.

Um diese systemischen Herausforderungen zu adressieren, konzentriert sich die Arbeit auf drei zentrale Forschungsfragen. Erstens wird untersucht, welche ökonomischen Anreizstrukturen in der Plattformökonomie, insbesondere im intransparenten Ad-Tech-Ökosystem, die Verbreitung von Desinformation finanzieren und begünstigen. Zweitens analysiert die Studie, wie aktuelle technologische Entwicklungen (wie beispielsweise AI Slop) und politische Trends (wie die Abkehr von klassischen Faktencheck-Ansätzen) die ökonomischen und technischen Dynamiken der Desinformation verändern. Drittens wird eruiert, welche systemischen Interventionsansätze (akteurs- versus inhaltsbasiert) empirisch am wirksamsten sind und wie die aktuellen regulatorischen Maßnahmen im internationalen Vergleich einzuordnen sind. Letzterer Punkt beinhaltet die Kontrastierung des prozessorientierten Regulierungsmodells der EU (DSA) mit inhaltsbasierten, oft repressiven „Fake-News“-Gesetzen in anderen Regionen.

Methodisch basiert die Arbeit auf einer umfassenden Desk Research, welche wissenschaftliche Literatur, Grauliteratur und Ad-hoc-Analysen von öffentlichen oder zivilgesellschaftlichen Institutionen einbezieht. Im Rahmen der Erstellung dieser Arbeit wurden KI-

gestützte Anwendungen[1] als methodische Hilfsmittel eingesetzt. Der Einsatz beschränkte sich auf die sprachliche Überarbeitung, die Strukturierung von Argumenten sowie die textliche Ausarbeitung eigener Gedankenskizzen. Die inhaltliche Verantwortung, die Auswahl der Quellen, die Prüfung der Fakten sowie die Schlussfolgerungen liegen vollständig bei den menschlichen Autoren. Es wurden keine KI-generierten Inhalte ohne menschliche Prüfung und Verifizierung in den Text übernommen. Alle zitierten Quellen wurden eigenständig recherchiert. Der Einsatz der KI diente ausschließlich der Effizienzsteigerung im Schreib- und Redaktionsprozess, nicht der Generierung primärer Forschungsdaten oder originärer Gedanken.

Strukturell gliedert sich die Untersuchung in die folgenden Hauptkapitel: Nach dieser Einleitung werden die Grundlagen (Kapitel 2) der Begriffsverständnisse, Typologien, Mechanismen (einschließlich der Verbreitung im hybriden Mediensystem) und Risiken dargelegt. Daran schließt sich die Analyse der Online-Plattformen in der Desinformationsökonomie (Kapitel 3) an, die die ökonomischen Anreizstrukturen und die algorithmische Verstärkung beleuchtet. Kapitel 4 befasst sich mit der Analyse und Bewertung von aktuellen Bekämpfungsstrategien und dem internationalen Regulierungsrahmen. Die Arbeit schließt mit dem Fazit (Kapitel 5), das die zentralen Erkenntnisse zusammenführt und Handlungsempfehlungen ableitet.

1 Alphabet. (2025). Gemini (Version 2.5) [Large language model]
OpenAI. (2025). ChatGPT (Version 5.2) [Large language model].

2 Grundlagen

2.1 Begriffsverständnis und Abgrenzung

Die wissenschaftliche und öffentliche Auseinandersetzung mit Desinformation steht weiterhin vor der zentralen Herausforderung einer fehlenden, einheitlichen Begriffsdefinition. Der Begriff wird von Forschenden, Regulierungsbehörden und Plattformbetreibern oft unterschiedlich ausgelegt, wobei sich je nach Disziplin verschiedene Schwerpunkte etabliert haben. In der politischen und regulatorischen Debatte, insbesondere auf Ebene der Europäischen Union, fungiert „Desinformation“ als der zentrale Leitbegriff, der primär durch die Kriterien der Intentionalität und des Schadenspotenzials definiert wird. Die EU-Kommission definiert Desinformation als nachweislich falsche, aber auch irreführende Informationen, die mit der Absicht erstellt, präsentiert und verbreitet werden, wirtschaftlichen Gewinn zu erzielen oder die Öffentlichkeit gezielt zu täuschen und ihr Schaden zuzufügen (Europäische Kommission, 2020). Diese Definition betont die Intentionalität und das Schadenspotenzial als entscheidende Kriterien für regulatorische Eingriffe.

Im Gegensatz dazu dominiert in der psychologischen und kommunikationswissenschaftlichen Forschung häufig der Begriff Misinformation als übergeordnetes Konzept, allerdings finden sich in der breiten akademischen Literatur anhaltend beide Varianten als Oberbegriff (vgl. Wang et al., 2025). Gerade aktuelle Studien verwenden „Misinformation“ oft als Sammelbegriff für alle Formen falscher oder ungenauer Informationen, unabhängig von der Absicht des Absenders. Der Grund hierfür liegt im Forschungsinteresse: Sowohl für die kognitive Verarbeitung und die psychologischen Effekte beim Rezipienten als auch für viele der potenziellen Schäden ist es oft zweitrangig, ob eine Falschinformation absichtlich als Desinformation oder unabsichtlich als Fehlinformation verbreitet wurde (Ecker et al., 2025; APA, 2025).

Um dieser terminologischen Unschärfe zu begegnen, greifen andere akademische Arbeiten auf das differenzierte Framework der „Information Disorder“ von Wardle und Derakhshan (2017) zurück, das präzise zwischen Misinformation (unabsichtlich), Disinformation (absichtlich) und Malinformation (schädlich, aber wahr) unterscheidet, um die spezifischen Dynamiken im digitalen Raum exakter zu erfassen.

1. Misinformation: Bezeichnet nachweislich falsche oder irreführende Informationen, die *ohne* bewusste Schädigungsabsicht verbreitet werden (z. B. durch unbeabsichtigtes Teilen).
2. Disinformation (Desinformation): Bezeichnet nachweislich falsche oder irreführende Informationen, die *mit* der Intention verbreitet werden, Schaden anzurichten oder politische bzw. ökonomische Zwecke zu verfolgen.
3. Malinformation: Bezeichnet die *absichtlich* schädigende Verbreitung von *echten* Informationen, oft durch deren Veröffentlichung aus dem Kontext gerissen (z. B. Leaks, Doxing).

Für eine Übersicht im regulatorischen Kontext des DSA wurde von Watolla et al. (2025) eine Arbeitsdefinition von Desinformation identifiziert, die vier Charakteristika umfasst: (1) nachweislich falsch oder irreführend, (2) Potenzial für gesellschaftlichen Schaden, (3) gezielte Streuung durch Akteure und (4) Verfolgung politischer oder ökonomischer Zwecke. Für die analytische Praxis erweisen sich strenge Abgrenzungen in verschiedenen Dimensionen als zunehmend herausfordernd. Wie die American Psychological Association festhält, ist insbesondere die Intention, die Mis- und Desinformation voneinander abgrenzt, in der Realität oft nur schwer nachzuweisen, da die inneren Motive eines Akteurs extern kaum valide überprüfbar sind (APA, 2023). Eine aktuelle Analyse von Nickl et al. (2025) schlägt zur Navigierung durch diese Begriffslandschaft daher vor, Informationsstörungen nicht in starren Kategorien, sondern konsequent als Kontinuum entlang der drei Achsen Wahrheit, Intention und Schaden zu verorten.

Besonders die Dimension der Wahrheit erweist sich dabei als komplexer Graubereich. Während klassische Definitionen oft nur explizit falsche Inhalte erfassen (Faktencheck: „falsch“), zeigt die Empirie, dass technisch korrekte, aber irreführende Informationen („misleading information“) oft eine weitaus größere Reichweite und Schadenswirkung entfalten. Allen et al. (2024) zeigen, dass bereits irreführend gerahmte Schlagzeilen, etwa wenn ein Todesfall zeitlich in die Nähe einer Impfung gerückt wird, auf Plattformen wie Facebook sechsmal so viele Aufrufe erzielten wie Inhalte, die explizit als falsch markiert wurden. Ein oft zitiertes Beispiel ist die Schlagzeile: „Ein ‚gesunder‘ Arzt starb zwei Wochen nach Erhalt einer COVID-19-Impfung; CDC untersucht warum“ (Benton, 2021). Obwohl die Aussage faktisch korrekt war, erzeugte die Rahmung den Eindruck eines unmittelbaren Zusammenhangs, wodurch es zu einer deutlich erhöhten viralen Verbreitung kam (APA, 2023). Van der Linden und Kyrychenko (2024) argumentieren deshalb, dass eine zu strikte Beschränkung auf verifizierbare Falschheit den Großteil der systemischen Risiken ausblenden würde, da Akteure gezielt Graubereiche nutzen, um Narrative zu formen, ohne technisch die Unwahrheit zu sagen. Die epistemologische Herausforderung besteht folglich darin, Desinformation als Spektrum zu begreifen, das von unabsichtlichen Fehlern bis zu strategischen Einflussoperationen reicht (Nickl et al., 2025).

Der Begriff „Fake News“ wird in dieser Arbeit bewusst sparsam verwendet. Zwar herrscht in der Wissenschaft ein zunehmender Konsens darüber, den Begriff zu vermeiden, da er unspezifisch und zu einem stark politisierten Kampfbegriff geworden ist, der oft zur Delegitimierung kritischer Berichterstattung genutzt wird (APA, 2023). Trotz dieses Konsenses wird der Begriff weiterhin prominent sowohl in der öffentlichen Debatte als auch in Teilen der akademischen Forschung genutzt, oft irreführenderweise sogar als Oberbegriff für verschiedenste Desinformationsphänomene und muss daher bei der Recherche oder in bibliographischen Arbeiten weiterhin berücksichtigt werden (vgl. Wang et al., 2025). Analytisch behält der Begriff in seinem teils genutzten engen Verständnis eine Relevanz, d.h. als fabrizierte Information, die Nachrichtenmedien-Inhalte in ihrer Form, aber nicht in ihrem Organisationsprozess oder ihrer Absicht nachahmt (Lazer et al., 2018). Dieser pseudo-journalistische Charakter sowie ein Fokus auf den

Aktualitätsbezug (im Deutschen auch: „aktuelle“ Desinformation, vgl. Zimmermann & Kohring, 2018) unterscheidet „Fake News“ von anderen Falschinformationen (z. B. Memes oder Verschwörungstheorien).

Vor diesem theoretischen Hintergrund wird für die vorliegende Studie, abweichend von der strengen akademischen Trennung, ein breiterer, pragmatischerer Arbeitsbegriff von Desinformation als Oberbegriff gewählt. Da diese Arbeit die systemischen Risiken der Verbreitung über soziale Medien und andere digitale Plattformen betrachtet und dabei unter anderem die Rolle algorithmischer Verstärkung einbezieht, ist ein Arbeitsbegriff sinnvoll, der nicht starr an den schwer überprüfbaren Kriterien der individuellen Intentionalität oder des absoluten Falschheitsgrads ansetzt (vgl. auch Watolla et al., 2025). Dieser Ansatz schließt z.B. Fälle ein, bei denen Inhalte trotz faktischer Korrektheit durch irreführende Kontextualisierung oder Rahmung schädliche Wirkungen entfalten können. Um dennoch die analytische Trennschärfe für jene Fälle zu wahren, in denen eine klare Schädigungsabsicht und koordinierte Verbreitung nachgewiesen werden kann, werden für solche Phänomene im weiteren Verlauf die präzisierenden Begriffe „gezielte Desinformation“ oder „organisierte Desinformation“ verwendet.

Um diese konzeptionelle Komplexität zu ordnen und die in dieser Studie verwendeten Begrifflichkeiten in das Feld der Informationsstörungen einzuordnen, systematisiert die folgende Übersichtstabelle (Tabelle 1) die Zusammenhänge zwischen Wahrheitsgehalt, Absicht und Schadenspotenzial. Die Akteure, Verbreitungsmechanismen und Risiken werden in den folgenden Teilkapiteln weiter adressiert.

Tabelle 1: Überblick – Dimensionen von Desinformation

	Dekontextualisierung	Falschinformation	Manipulative (politische) Werbung	Pseudo-journalismus	Propaganda
Abweichung von der Faktizität	gering	hoch	verschieden	eher gering	hoch
Typische Intention	manipulatives Narrativ verbreiten, das eine politische Ideologie stützt; ökonomisch (Clickbait)	ökonomisch; ideologische (De-)mobilisierung	politische Mobilisierung	ökonomisch; ideologische (De-)Mobilisierung	geopolitische und ideologische (De-)Mobilisierung
Typische Sender	politische Akteure, Medien, alternative Medien	Internetbetrüger, Verschwörungstheoretiker, alternative Medien	politische Akteure, NGOs	Internetbetrüger, alternative Medienmacher	Staatsregierungen und (internationale) Organisationen
Typische Verbreitung	weite Verbreitung: kommt in allen Medien vor, häufig in alternativen Medien zu finden, weitere Verbreitung durch Nutzer	eingeschränkte Verbreitung: häufig über Soziale Medien, manchmal unterstützt durch koordiniertes unauthentisches Verhalten	bezahlte Verbreitung: häufig in Sozialen Medien, aber auch über andere Wahlkampfkanäle	eingeschränkte Verbreitung: in eigenen Onlinemedien oder über Soziale Netzwerke, weitere Verbreitung durch Nutzer	professionelle Verbreitung: über alle Kommunikationskanäle einschließlich eigener Medienorganisationen und mithilfe von koordiniertem unauthentischen Verhalten

	Dekontextualisierung	Falschinformation	Manipulative (politische) Werbung	Pseudo-journalismus	Propaganda
Typische Risiken für das Individuum	kognitiv, emotional, (politische) Fehlentscheidungen (in unterschiedlichem Ausmaß)				
Typische Risiken für die Gesellschaft	misinformierte Wählerschaft, polarisierend	misinformierte Wählerschaft, spaltend, demokratiegefährdend	polarisierend	misinformierte Wählerschaft, polarisierend, spaltend	geopolitisch, spaltend, demokratiegefährdend

Quelle: Möller et al. (2020, S. 38)

2.2 Typologie: Akteure und Motive

Um die systemischen Risiken der Desinformation zu erfassen, ist eine differenzierte Betrachtung notwendig, die über den reinen Wahrheitsgehalt der Informationen hinausgeht. Die Landschaft der Desinformation ist durch eine Heterogenität der Akteure und ihrer Motivationen geprägt, die von geopolitischer Destabilisierung bis hin zu reinem Profitstreben reicht. Basierend auf der Analyse von Möller et al. (2020) und den Klassifikationen der Europäischen Kommission (2018) lässt sich das Feld anhand von Akteursgruppen und thematischen Erscheinungsformen systematisieren.

Die Urheber und Verbreiter von Desinformation lassen sich sinnvoll nach ihrer primären Motivation differenzieren, auch wenn in der Praxis zahlreiche Mischformen auftreten. Die High Level Expert Group on Fake News and Online Disinformation der Europäischen Kommission unter Leitung von de Cock Buning (2018) unterscheidet explizit zwischen staatlichen Akteuren, politischen nichtstaatlichen Akteuren, profitorientierten Akteuren sowie Bürgerschaft und betont, dass Desinformation entweder auf politischen/gesellschaftlichen Schaden oder auf finanzielle Gewinne abzielt. Auch neuere Studien heben hervor, dass eine Analyse ohne diese Differenzierung unvollständig bleibt, da die Gegenmaßnahmen je nach Akteurstyp variieren müssen (Unger et al., 2025; CPA, 2023).

Eine zentrale Gruppe bilden staatliche und geopolitische Akteure, die Desinformation als Instrument der Machtprojektion nach innen und außen einsetzen. Bradshaw und Howard (2018, 2019) zeigen anhand eines globalen Inventars organisierter Social-Media-Manipulation, dass Regierungen und Sicherheitsapparate in zahlreichen Ländern systematisch „computational propaganda“ nutzen, um Wahlen zu beeinflussen, Oppositionelle zu delegitimieren und außenpolitische Ziele durchzusetzen. In autoritären Kontexten dient Desinformation oft primär der Machtsicherung und dem Regimeüberleben (Sato und Wiebrecht, 2024). Außenpolitisch zielen diese Strategien darauf ab, die öffentliche Sphäre anderer Staaten zu verzerren, Vertrauen in Institutionen zu erodieren und so die demokratische Handlungsfähigkeit zu schwächen (Radsch, 2022; Bayer et al., 2019). Die Methoden reichen von klassischer Propaganda in staatsnahen Medien bis hin zu digital skalierter Einflussnahme. Bradshaw und Howard (2019) dokumentieren den Einsatz von Bots, Sockenpuppen-Accounts und Trollarmeen als Kernelemente. Auf Plattformebene

werden diese Praktiken häufig als „Coordinated Inauthentic Behavior“ (CIB) zusammengefasst. Gemeint ist hierbei das abgestimmte Vorgehen von Netzwerken, die ihre wahre Identität verschleiern (Cinelli et al., 2022; Di Marco et al., 2025). Zudem nutzen diese Akteure gezielt bestehende gesellschaftliche Bruchlinien („Wedge Issues“) wie Migration oder Identitätspolitik, um Polarisierung zu vertiefen und politische Prozesse zu blockieren (Freelon et al., 2022; Haßler et al., 2025.

Demgegenüber stehen ideologisch motivierte Akteure, zu denen Parteien, „alternative Medien“, Aktivistengruppen und Verschwörungsideologen zählen. Ihr Hauptziel liegt in der politischen Mobilisierung, dem Agenda-Setting und der Polarisierung innerhalb des eigenen Landes. Charakteristisch für diese Gruppe ist die Methode der Dekontextualisierung: Wahre oder halbwahre Informationen werden aus ihrem ursprünglichen Zusammenhang gerissen, um ein spezifisches ideologisches Narrativ zu stützen (Möller et al., 2020).

Eine dritte Gruppe bilden Akteure, deren Handeln primär der finanziellen Gewinnmaximierung folgt. Das Spektrum reicht von kriminellen Betrügern (Scammern) bis hin zu Betreibern von „Clickbait“-Portalen. Diese Gruppe ist strukturell eng mit der Werbeindustrie (Ad Tech) verzahnt. Desinformation wird hier oft nicht aus inhaltlicher Überzeugung produziert, sondern als Mittel zum Zweck: Sie dient dazu, Aufmerksamkeit zu binden, die über das System des „Programmatic Advertising“ automatisiert monetarisiert wird (GDI, 2019). Dabei werden algorithmische Präferenzen für emotionalisierende Inhalte gezielt ausgenutzt, um Traffic und Werbeeinnahmen zu generieren. (Die Mechanismen dieses Werbe-Ökosystems werden in Kapitel 3.1 detaillierter analysiert.)

Parallel zu den Akteuren lassen sich drei thematische Hauptbereiche identifizieren, die spezifische Risiken bergen. In der politischen Desinformation steht die Manipulation der öffentlichen Meinungsbildung im Vordergrund. Dies umfasst Angriffe auf die Wahlintegrität sowie Formen der „Computational Propaganda“, die den Einsatz von Automatisierung zur Manipulation der öffentlichen Meinung beschreibt (Woolley & Howard, 2018). Ein zentrales Instrument ist manipulative politische Werbung, die oft unabhängig vom Wahrheitsgehalt der Mobilisierung dient. In ihrer schärfsten Form tritt sie als Propaganda auf, um ideologische Ziele durchzusetzen (Bennett & Livingston, 2018).

Im Gesundheitsbereich prägte die WHO den Begriff der „Infodemic“, und bezeichnet damit eine Überflutung mit Informationen, die sowohl zutreffend als auch falsch oder irreführend sein können, insbesondere in digitalen und physischen Umgebungen während Krankheitsausbrüchen. Diese Informationsflut erschwert es Menschen, verlässliche Quellen zu identifizieren, fördert riskantes Gesundheitsverhalten und kann das Vertrauen in Gesundheitsbehörden beeinträchtigen (Zarocostas, 2020). Eine systematische Übersichtsarbeit für die WHO zeigt, dass Infodemics eng mit Impf-zögerlichkeit, der Nutzung unwirksamer oder schädlicher Behandlungen und der Missachtung evidenzbasierter Empfehlungen (z. B. Maskentragen, Social Distancing) verknüpft sind (Borges do Nascimento et al., 2022). Gentili et al. (2024) betonen zudem, dass Infodemics mittlerweile als

eigenständige Bedrohung der öffentlichen Gesundheit betrachtet werden, da sie sowohl individuelle Entscheidungen als auch kollektive Krisenreaktionen systematisch verzerren.

Empirische Studien belegen, dass die Exposition gegenüber gezielter Gesundheitsdesinformation messbare Effekte auf Einstellungen und Verhalten hat: Allen, et al. (2024) zeigen anhand von Facebook-Daten und Experimenten, dass bereits begrenzter Kontakt mit Impf-Desinformation und impfskeptischen Inhalten die Impfintention senken und falsche Risikowahrnehmungen stabilisieren kann. Gleichzeitig wird in mehreren Arbeiten deutlich, dass Vertrauen in Wissenschaft, medizinisches Personal und Institutionen ein zentraler Puffer gegen solche Effekte ist: Höheres Vertrauen in Wissenschaft und Gesundheitsinstitutionen geht systematisch mit höherer Impfbereitschaft und geringerer Anfälligkeit für Gesundheitsfalschinformationen einher (Kara et al., 2025; Roozenbeek et al., 2025).

Auf der psychologischen Ebene greift hier häufig der „Illusory Truth Effect“: Wiederholte Aussagen werden als glaubwürdiger erlebt, selbst wenn sie objektiv falsch sind. Klassische Studien sowie eine Meta-Analyse zeigen, dass bereits bloße Wiederholung die subjektive Wahrhaftigkeit von Aussagen erhöht. Dieser Effekt tritt auch dann auf, wenn Menschen das relevante Faktenwissen eigentlich besitzen (Dechêne et al., 2010; Fazio et al., 2015). Neuere Übersichtsarbeiten bestätigen, dass der Illusory Truth Effect ein robuster Mechanismus ist, der sich über verschiedene Domänen hinweg zeigt, von Alltagswissen bis hin zu politischer und Gesundheitsdesinformation (Udry & Barber, 2024). Vellani et al. (2023) können experimentell zeigen, dass bereits eine einmalige Wiederholung von Falschinformationen die Bereitschaft erhöht, diese Inhalte in sozialen Netzwerken zu teilen; vermittelt wird dieser Effekt über eine gesteigerte wahrgenommene Richtigkeit. Vergleichbare Effekte wurden inzwischen auch für Finanzinformationen beschrieben: Wiederholte irreführende Börsennachrichten können insbesondere bei selbstsicheren Investoren zu erhöhter Risikobereitschaft und einer stärkeren Allokation in riskante Anlagen führen (Yun et al., 2025).

Oft unterschätzt, aber ökonomisch hochrelevant, betrifft dies den weiteren Bereich der kommerziellen Desinformation. Er umfasst u. a. Betrugsmaschen (Scams), gefälschte Produktbewertungen, irreführende Finanzempfehlungen („Fin-Influencer“, vgl. Hayes und Ben-Shmuel, 2024; Keasey et al., 2025) und Greenwashing. Ökonomische Analysen des Plattformökosystems zeigen, dass gefälschte Online-Rezensionen nicht nur einzelne Konsumenten täuschen, sondern auch Marktstrukturen verzerren: Gandhi et al. (2025) weisen für Amazon nach, dass der systematische Einsatz gefälschter Produktbewertungen Ratings künstlich erhöht, Nachfrage von ehrlichen zu opportunistischen Anbietern umleitet und die Konsumentenwohlfahrt insgesamt reduziert. Auch auf den Finanzmärkten lassen sich klare Effekte nachweisen: Arcuri et al. (2023) zeigen in einer Event-Study für US- und EU-Börsen, dass negative Falschmeldungen über Unternehmen signifikante kurzfristige Kursverluste verursachen, während positive Fake News keine vergleichbaren, stabilen Kursgewinne erzeugen. Weitere Studien zu Fake News in wirtschaftlichen

Kontexten legen nahe, dass Desinformation sogar die makroökonomische Unsicherheit erhöht und Konjunkturzyklen verstärken kann (Assenza et al., 2024).

Im Nachhaltigkeitsbereich wird kommerzielle Desinformation insbesondere in Bezug zum Thema Greenwashing diskutiert. Systematische Reviews zeigen, dass überzogene oder selektive Umweltversprechen („green framing“) das Vertrauen in Marken und Finanzinstitutionen erodieren und langfristig Reputations- sowie Regulierungsrisiken erzeugen (Galletta et al., 2024; Persakis et al., 2025; Feghali et al., 2025). Eine aktuelle Übersichtsarbeit zu Greenwashing und Branding kommt zu dem Ergebnis, dass aufgedeckte Greenwashing-Praktiken insbesondere Markenreputation und Kundenbindung erheblich beeinträchtigen können (vgl. AlQahtani, 2025).

Kommerzielle Akteure nutzen zur Verbreitung solcher Inhalte häufig Formate des Pseudo-Journalismus, um Glaubwürdigkeit zu simulieren. Fake News und desinformationsaffine Inhalte imitieren dabei systematisch die Form klassischer Nachrichtenseiten. Sie übernehmen etwa Layout, Überschriften und vermeintlich unabhängiger redaktioneller Texte, werden aber von politischen oder wirtschaftlichen Interessen gesteuert (Egelhofer & Lecheler, 2019). Palau-Sampio (2023) beschreibt in diesem Zusammenhang „Pseudo-Medien“, die mit verschwörungstheoretischen und verzerrten Darstellungen der sozialen Realität arbeiten und mithilfe von Clickbait und polarisierender Sprache zu einem stabilen Desinformationsökosystem beitragen. Studien zu „hybrider Werbung“ und Native Advertising in digitalen Nachrichtenmedien zeigen, dass solche Formate die Grenze zwischen Redaktion und Werbung gezielt verwischen und damit die wahrgenommene Glaubwürdigkeit journalistischer Angebote untergraben (Lauerer & Beckert, 2024; Di Domenico et al., 2021).

Zusammengenommen bekräftigen diese Beobachtungen die Einschätzung, dass kommerzielle Desinformation ein konkretes wirtschaftliches Risiko für Unternehmen darstellt. Die Bandbreite reicht von unmittelbaren Kurs- und Absatzverlusten über langfristige Reputationsschäden bis hin zu erhöhter regulatorischer und haftungsrechtlicher Exponierung (IDW, 2025; vgl. auch Arcuri et al., 2023; Gandhi et al., 2025).

Führt man Akteure, Motive und Zielgruppen zusammen, ergibt sich ein Klassifikationsraster (siehe Tabelle 2). Dieses Raster verdeutlicht, dass pauschale Gegenmaßnahmen selten wirksam sind. Während gegen ökonomisch motivierte Akteure der Entzug von Werbeeinnahmen („De-Monetarisierung“) effektiv sein kann (GDI, 2019), erfordern ideologisch motivierte Akteure primär diskursive oder bildungspolitische Interventionen. In der Praxis bestehen zudem fließende Übergänge: Staaten bedienen sich kommerzieller Infrastrukturen, ideologische Akteure monetarisieren ihre Inhalte über Werbung, wirtschaftliche Akteure amplifizieren polarisiertes politisches Material, weil es gut performt.

Tabelle 2: Überblick Akteure & Motive

Kategorie	Typus A: Staatlich (ausländische Akteure)	Typus B: Wirtschaftlich	Typus C: Ideologisch / Politisch (inländische Akteure)
Akteure	Geheimdienste, Regierungen, Staatsmedien	Betrüger, Clickbait-Betreiber, PR-Agenturen	Parteien, Aktivisten, Extremisten, Alternative Medien
Primäres Motiv	Geopolitik / Macht (Einflussnahme, Destabilisierung)	Finanzen (Werbeeinnahmen, Betrug, Aktienkurse)	Ideologie / Überzeugung (Mobilisierung, Meinungsmache)
Typische Methode	Propaganda, Computational Propaganda (z.B. Bot-Netzwerke)	Pseudo-Journalismus, Junk News (Clickbait)	Dekontextualisierung, Framing
Zielgruppe	Gesamtbevölkerung, Wähler, Minderheiten	Konsumenten, Anleger, Patienten	Eigene Anhänger (Mobilisierung) vs. Gegner (Polarisierung)
Hauptrisiko	Demokratiegefährdung, geopolitische Spaltung	Finanzieller Schaden, Gesundheitsgefahren (Infodemic)	Gesellschaftliche Spaltung, Radikalisierung

Quelle: Eigene Zusammenstellung

2.3 Generierung und Verbreitungsmechanismen

Die Verbreitung von Desinformation folgt keinen linearen Wegen, sondern ist das Resultat eines komplexen Zusammenspiels aus psychologischen Prädispositionen der Nutzenden, technologischen Verstärkern und den strukturellen Bedingungen des modernen Medienökosystems. Wie Watolla et al. (2025) darlegen, ist Desinformation kein isoliertes Plattform-Phänomen, sondern in größere Narrative eingebettet, deren Durchschlagskraft auf spezifischen Mechanismen der Generierung und Distribution beruht.

Auf der Ebene der Generierung ist entscheidend, dass erfolgreiche Desinformation oft gezielt auf kognitive Verzerrungen zugeschnitten ist. Inhalte werden so gestaltet, dass sie eine hohe emotionale Resonanz erzeugen. Studien zeigen hierbei, dass Inhalte mit starker emotionaler Aufladung eine signifikant höhere Verbreitungsgeschwindigkeit aufweisen als neutrale Informationen. Dies gilt insbesondere für die Emotionen von Wut und Empörung. Brady et al. (2017) beschreiben dies als „Moral Contagion“, bei der Nachrichten, die moralische Emotionen innerhalb einer sozialen Gruppe ansprechen, bevorzugt geteilt werden, um die eigene Gruppenidentität zu signalisieren und den Zusammenhalt zu stärken. Ergänzend hierzu wiesen Vosoughi et al. (2018) in einer umfassenden Analyse auf X-Vorgänger Twitter nach, dass sich Falschinformationen „schneller, tiefer und breiter“ verbreiten als wahre Nachrichten. Dies führen die Autoren primär auf den Neuheitswert zurück, da Falschinformationen nicht an die Realität gebunden sind und daher überraschender und sensationeller gestaltet werden können, was die menschliche Neigung zur Weitergabe neuer Informationen stimuliert. Schließlich verfängt Desinformation besonders dort, wo sie bestehende Weltbilder bestätigt (Confirmation Bias). Pennycook und Rand (2021) argumentieren in diesem Kontext, dass weniger ein Mangel an Intelligenz, sondern vielmehr „motiviertes Denken“ (Motivated Reasoning) dazu führt, dass politisch kongruente Falschinformationen akzeptiert und weiterverbreitet werden.

Eine besondere Rolle im aktuellen Desinformationsumfeld nehmen Soziale Medien und andere Online-Plattformen ein, die in Kapitel 3 noch detaillierter betrachtet werden. Um die Rolle digitaler Plattformen im Kontext von Desinformation und gesellschaftlicher Spaltung zu verstehen, ist eine differenzierte Betrachtung der technologischen und sozialen Mechanismen notwendig. Watolla et al. (2025) identifizieren hierbei die algorithmische Kuration und das koordinierte Verhalten als zwei zentrale Hebel, über die Plattformen zur Verbreitung problematischer Inhalte beitragen können. Da die Geschäftsmodelle großer Plattformen primär auf der Maximierung von Verweildauer und Interaktionen basieren, privilegieren Empfehlungsalgorithmen jene Inhalte, die ein hohes Engagement in Form von Klicks, Kommentaren oder Shares versprechen. Emotionalisierende und polarisierende Inhalte (nicht zuletzt in Form von Desinformation) erfüllen diese Verwertungslogik besonders gut, sodass eine unbeabsichtigte, aber systemische Bevorzugung solcher Inhalte entstehen kann (Bakir & McStay, 2018; Meßmer & Degeling, 2023).

Vor diesem Hintergrund erscheint die Annahme plausibel, dass diese Mechanismen zur Ausbildung geschlossener Echokammern und Filterblasen führen. Empirische Ergebnisse zeigen jedoch ein deutlich nuancierteres Bild. Überblicksstudien kommen übereinstimmend zu dem Schluss, dass die idealtypische Vorstellung vollständig abgeschotteter Informationsräume, in denen Nutzerinnen und Nutzer ausschließlich gleichgerichtete Inhalte sehen, stark überzeichnet ist (Stark et al., 2019; Zuiderveen Borgesius et al., 2016). Bruns (2019) argumentiert, dass Konzepte wie die Filterblase oft auf einem technologischen Determinismus beruhen, der die tatsächliche Vielfalt individueller Medienrepertoires und die Durchlässigkeit sozialer Netzwerke ignoriert. Auch Stark et al. (2019) zeigen, dass die impliziten Personalisierungseffekte von Algorithmen bislang überschätzt wurden und Filterblasen in ihrer Reinform empirisch selten nachweisbar sind. Statt in binären Kategorien zu denken, plädieren sie dafür, diese Phänomene als Kontinuum zwischen einem breit überlappenden Informationsrepertoire und einer graduellen Fragmentierung zu betrachten.

Die Annahme einer vollständigen informationellen Isolation ist demnach empirisch kaum haltbar. Untersuchungen im deutschen Kontext bestätigen, dass Nutzer keineswegs in homogenen Meinungsräumen gefangen sind, sondern über verschiedene Kanäle hinweg mit einer Pluralität an Themen in Kontakt kommen (Stark et al., 2019). Dubois und Blank (2018) weisen zudem darauf hin, dass Personen mit hohem politischen Interesse und vielfältigen Medienquellen deutlich seltener in Echokammern geraten; problematische Konstellationen finden sich eher bei stark politisierten, aber selektiv informierten Subgruppen. Meßmer und Degeling (2023) betonen, dass Polarisierung ein primär soziales Phänomen ist, das nicht monokausal auf Plattformen zurückgeführt werden kann, sondern im Zusammenspiel von individuellen Selektionsentscheidungen, sozialen Netzwerkstrukturen und algorithmischen Gewichtungen betrachtet werden sollte.

Die systemische Problematik liegt daher weniger in einer Abschottung von fremden Meinungen, sondern vielmehr in der Art der algorithmisch geförderten Konfrontation. Da Empfehlungssysteme Interaktionen wie Kommentare oder Verweildauer als Indikatoren

für Relevanz werten, können sie technisch nicht zwischen Zustimmung und Ablehnung unterscheiden. Dies begünstigt Phänomene wie das „Hate Following“, bei dem Nutzer gezielt Inhalten folgen, die sie emotional aufwühlen oder verärgern (vgl. Richardson et al., 2024). Algorithmen priorisieren vor allem Inhalte, die Aufmerksamkeit binden, inklusive von Beiträgen, die Wut auslösen. Inhalte, die dem Nutzer lediglich gefallen, werden demgegenüber nicht immer in gleicher Weise verstärkt. Soziologische Analysen, wie die von Törnberg (2022), deuten darauf hin, dass Polarisierung paradoxerweise nicht durch Isolation, sondern durch die ständige, affektiv aufgeladene Konfrontation mit gegensätzlichen Weltbildern verstärkt wird. In einem globalen digitalen Raum, in dem lokale Kontexte entfallen, dient die politische Identität als primäres Abgrenzungsmerkmal. Der Kontakt mit dem „politisch Anderen“ führt hierbei oft nicht zur Verständigung, sondern zur Verhärtung der eigenen Position.

Diese Dynamik wird durch experimentelle Daten gestützt. Schon Bail et al. (2018) konnten zeigen, dass die gezielte Exposition gegenüber gegensätzlichen politischen Ansichten auf Twitter Einstellungen eher polarisiert als moderiert. Diese Ergebnisse untermauern den Hinweis, dass einfache Rezepte wie „mehr Gegenmeinungen im Feed“ empirisch nicht tragen (Meßmer & Degeling, 2023). Die algorithmische Logik fördert somit weniger die Entstehung isolierter Blasen als vielmehr eine Arena permanenter, identitätsstiftender Konflikte. Aus Sicht der Plattformregulierung bedeutet dies, dass Algorithmen zwar Anreize für polarisierende Inhalte setzen, systemische Risiken jedoch nur unter Berücksichtigung der Nutzungspraktiken und gesellschaftlicher Konfliktlinien adäquat verstanden werden können.

Jenseits der organischen Verbreitung nutzen Akteure technische Mittel zur künstlichen Amplifikation, z.B. im Rahmen des CIB. Dies umfasst den Einsatz von Bot-Netzwerken („Astroturfing“, vgl. Mahbub et al, 2019; Chan, 2024), um den Anschein einer breiten öffentlichen Unterstützung zu erwecken, sowie die gezielte Manipulation von Suchmaschinen (Data Voids, vgl. Norocel et al., 2023; Mannino et al., 2024). Golebiewski und Boyd (2018) zeigen auf, wie Akteure gezielt Begriffe besetzen, zu denen es kaum etablierte Inhalte gibt, um bei spezifischen Suchanfragen Desinformation prominent zu platzieren.

Während in diesem Diskussionsbeitrag der Fokus auf aktuellen Entwicklungen im Bereich der Online-Plattformen liegt, greift eine isolierte Betrachtung für ein holistisches Verständnis jedoch zu kurz. Desinformation entfaltet ihre volle gesellschaftliche Wirkung oft erst durch das Zusammenspiel mit traditionellen Massenmedien in einem „hybriden Mediensystem“ (Chadwick, 2017). Strategische Akteure nutzen soziale Medien oft als Testlabor, um Narrative zu platzieren, die anschließend in den Mainstream diffundieren sollen. Marwick und Lewis (2017) beschreiben diesen Prozess als „Trading Up the Chain“: Desinformation wird zunächst in Nischenforen oder Blogs platziert, von mittelgroßen Influencern aufgegriffen und erreicht schließlich Journalisten etablierter Medien. Wird die Falschmeldung dort aufgegriffen, etwa um sie zu widerlegen, kann sich ihre Reichweite dennoch unbeabsichtigt deutlich erhöhen.

Politische Akteure nutzen zudem gezielt Tabubrüche und Desinformation im Sinne einer Strategie der Informationsüberflutung (‚Flooding'), um die mediale Aufmerksamkeit zu monopolisieren. Watolla et al. (2025) greifen dies treffend als „DDoS-Attacke auf die menschliche Aufmerksamkeit" auf, also eine gezielte Strategie zur Überlastung der kognitiven Verarbeitungskapazitäten der Rezipienten durch eine übermäßige Informationsdichte. Traditionelle Medien, die über diese Provokationen berichten, fungieren dabei unfreiwillig als Multiplikatoren und legitimieren die Themenwahl der Desinformationsakteure (Phillips, 2018). Diese Cross-Media-Flows verlaufen bidirektional. Während Online-Desinformationen die redaktionelle Agenda beeinflussen können („Bottom-Up"), dienen etablierte Medienmarken oft auch als Vehikel für Desinformation, wenn Akteure deren Glaubwürdigkeit durch gefälschte Screenshots oder kontextlose Zitate missbrauchen (Watolla et al., 2025). Zusammenfassend lässt sich festhalten, dass die Verbreitung von Desinformation auf einer Symbiose aus emotionaler Nutzeransprache, algorithmischer Belohnung von Engagement und der Ausnutzung journalistischer Aufmerksamkeitslogiken beruht.

2.4 Auswirkungen und Risiken

Die Quantifizierung konkreter Schäden durch Desinformation stößt in der akademischen Forschung auf das methodische Kernproblem der kausalen Inferenz. Wie eine Meta-Analyse in Nature Human Behaviour darlegt, ist es in Feldstudien oft kaum möglich, den Einfluss von Desinformation von bereits bestehenden soziopolitischen Polarisierungstendenzen zu isolieren (Lorenz-Spreen et al., 2023). Insbesondere Selektionseffekte spielen eine entscheidende Rolle: Nutzer empfangen Informationen nicht passiv, sondern selektieren aktiv Inhalte, die ihre bestehenden Überzeugungen bestätigen. Aufgrund dieser methodischen Hürden ist eine Einzelfallbewertung von Inhalten oft wenig zielführend.

2.4.1 Exposition und Rezeption

Eine zentrale Frage der Wirkungsforschung betrifft die Diskrepanz zwischen der theoretischen Verfügbarkeit und der tatsächlichen Rezeption von Falschinformationen. Neuere Studien befassen sich intensiv mit der Frage, wie viele Menschen tatsächlich Desinformation ausgesetzt sind und warum sie darauf reagieren. Dabei zeigt sich, dass die Anfälligkeit nicht gleichmäßig verteilt ist, sondern spezifischen psychologischen Mechanismen unterliegt.

In einer aktuellen Untersuchung identifizieren Hubeny et al. (2025) komplexe Prädiktoren für die Glaubhaftigkeitsbewertung. Ihre Ergebnisse zeigen, dass Anfälligkeit weniger eine Frage der Intelligenz ist, sondern stark mit Persönlichkeitsmerkmalen wie „Bullshit Receptivity" (Empfänglichkeit für pseudoprofunde Aussagen) und einer generellen Verschwörungsmentalität korreliert. Gleichzeitig wirken Merkmale wie „kognitive Reflexion" und „intellektuelle Demut" als Schutzfaktoren. Entscheidend ist jedoch der Faktor der

Identitätsprotektion: Nutzer akzeptieren Falschinformationen oft nicht aus Unwissenheit, sondern weil diese ihr bestehendes Weltbild stützen. Dieser „Myside Bias“ führt dazu, dass Informationen, die der eigenen Gruppenidentität entsprechen, mit einer niedrigeren Akzeptanzschwelle durchgewunken werden. Desinformation wirkt also nicht als universeller „Virus“, sondern verfängt primär dort, wo sie mit den psychologischen Dispositionen und Identitätsbedürfnissen spezifischer Zielgruppen resoniert.

Ergänzend hierzu bestätigen aktuelle Arbeiten, dass der entscheidende Mechanismus oft nicht fehlendes Wissen, sondern „Identity-Protective Motivated Reasoning“ ist (Hubeny, Nahon und Gawronski, 2025). Nutzer akzeptieren Falschinformationen, wenn diese der Aufwertung ihrer eigenen sozialen Gruppe („In-Group“) dienen. Dieser „Partisan Bias“ führt dazu, dass selbst kognitiv leistungsfähige Individuen ihre analytischen Fähigkeiten nutzen, um Falschinformationen zu rationalisieren, statt sie zu entlarven. Algorithmische Empfehlungssysteme verstärken diesen Effekt, indem sie Inhalte priorisieren, die starke moralisch-emotionale Reaktionen hervorrufen (Brady et al., 2023).

Besonders problematisch ist dies vor dem Hintergrund des „News-Finds-Me“(NFM)-Effekts: Da gerade junge Nutzer Nachrichten zunehmend passiv über den Feed konsumieren (und nicht aktiv suchen), kann die algorithmische Kuration ihr Weltbild stärker beeinflussen. Dieses Konzept beschreibt die Überzeugung von Individuen, dass sie Nachrichten nicht mehr aktiv suchen müssen, um gut informiert zu sein, da relevante Informationen sie automatisch über ihre sozialen Netzwerke und algorithmisch kuratierten Feeds erreichen werden. Diese passive Haltung gegenüber dem Informationserwerb hat weitreichende Konsequenzen für das politische Wissen, das Gesundheitsverhalten und nicht zuletzt für die Anfälligkeit gegenüber Mis- und Desinformation.

Ursprünglich als direkte Folge der zunehmenden Verbreitung sozialer Medien konzeptualisiert, wird NFM in der aktuellen Forschung als eine tiefere kognitive Disposition verstanden, die eng mit spezifischen Mediengewohnheiten verknüpft ist. Campbell und Hawkins (2025) zeigen in ihrer Untersuchung, dass NFM nicht allein durch die Nutzungshäufigkeit sozialer Medien entsteht, sondern signifikant durch habitualisierte, unbewusste Nutzungsmuster („Habits“) sowie durch spezifische „Mindsets“ gefördert wird. Die NFM-Wahrnehmung korreliert stärker mit einem „Connection Mindset“, also dem Vertrauen darauf, dass soziale Verbindungen den Informationsfluss sichern, als mit einem „Algorithm Mindset“, das auf technische Filter vertraut.

Die Ausprägung dieser Wahrnehmung ist zudem nicht statisch, sondern variiert thematisch. Während das Konzept ursprünglich im Kontext politischer Nachrichten („Hard News“) entwickelt wurde, weisen Mosallaei et al. (2025) nach, dass die Wahrnehmung bei Unterhaltungs- und Sportnachrichten („Soft News“) oft noch stärker ausgeprägt ist. Ein entscheidender Faktor ist dabei das individuelle Interesse: Ein hohes thematisches Interesse (z. B. an Politik) schwächt die NFM-Wahrnehmung tendenziell ab, da es die Bereitschaft zur aktiven Suche erhöht, während Desinteresse das passive Vertrauen in den Feed verstärkt.

Die mit der NFM-Wahrnehmung einhergehende Passivität hat weitreichende Konsequenzen für die Informationsintegrität und die Anfälligkeit für Fehl- und Desinformation. Gil de Zúñiga und Cheng (2024) belegen in ihrem Review, dass NFM konsistent negativ mit politischem Wissen korreliert, da das subjektive Gefühl der Informiertheit das Bedürfnis nach aktiver Wissensaneignung unterdrückt. Da Nutzer mit hoher NFM-Ausprägung davon ausgehen, ohnehin gut informiert zu sein, verspüren sie seltener die Notwendigkeit, Nachrichten aktiv zu verifizieren oder zusätzliche Quellen heranzuziehen. Zhang und Jiang (2024) wiesen im Gesundheitskontext nach, dass diese Haltung dazu führte, dass Nutzer während der COVID-19-Pandemie aktiv Informationen mieden. Diese Vermeidung korrelierte direkt positiv mit der Wahrscheinlichkeit, falschen Gesundheitsinformationen Glauben zu schenken, da korrigierende Informationen nicht aktiv gesucht wurden. Lin et al. (2024) differenzieren dieses Ergebnis weiter aus, indem sie ebenfalls zeigen, dass NFM nicht nur die Informationssuche hemmt, sondern auch direkt negativ mit dem tatsächlichen Gesundheitswissen korreliert. In ihrer Studie fungierte der wahrgenommene Mangel an Informationen („Information Insufficiency“) als zentraler Mediator: Personen mit hoher NFM-Wahrnehmung verspüren subjektiv keinen Informationsbedarf, was ihre Intention, aktiv auf sozialen Medien nach validen Informationen zu suchen, signifikant senkt.

Ein weiterer ähnlich gelagerter Mechanismus verstärkt diese Problematik im Kontext von Desinformation. Nutzer mit hoher NFM-Wahrnehmung neigen dazu, ihre eigene Medienkompetenz zu überschätzen und gleichzeitig andere für deutlich anfälliger für Manipulation zu halten. Dieses Phänomen wird als „Third-Person Perception“ bezeichnet. Tian und Willnat (2025) konnten zeigen, dass diese wahrgenommene eigene Immunität dazu führt, dass sich Nutzer kognitiv weniger anstrengen, um den Wahrheitsgehalt von Nachrichten zu prüfen. Ironischerweise resultiert diese Selbstüberschätzung darin, dass NFM-Nutzer häufiger mit Desinformation interagieren und anfälliger für diese sind, da sie ihre Schutzmechanismen aufgrund falsch verstandener Sicherheit senken.

Die Passivität wirkt nicht nur auf der Rezeptionsseite, sondern treibt auch die Verbreitung von Desinformation an. Hawkins und Campbell (2025) identifizierten die NFM-Wahrnehmung in Untersuchungen zur US-amerikanischen „Alt-Right“ als direkten Prädiktor für das aktive Teilen von Desinformation. Nutzer, die sich auf den zufälligen Nachrichtenkontakt verlassen und Nachrichten interaktiv über Mobilgeräte konsumieren, werden so häufig unbewusst zu Multiplikatoren in Desinformationsnetzwerken. Der „News-finds-me“-Effekt markiert somit eine strukturelle Schwachstelle in der digitalen Öffentlichkeit, die kritisches Hinterfragen reduziert und die Resilienz gegenüber manipulativen Inhalten systematisch untergräbt.

2.4.2 Risiken

Trotz der methodischen Schwierigkeiten beim Nachweis direkter Kausalitäten lassen sich konkrete Risiken identifizieren bis hin zum Übergang von der digitalen Welt in physische Bedrohungen und Schäden markieren. Die Unruhen im britischen Southport im Sommer

2024 dienen als prominentes Fallbeispiel für die direkte Verbindung zwischen Online-Inhalten und Offline-Schaden. Nach einem Messerangriff nutzten feindliche Akteure und rechtsextreme Netzwerke das entstehende „Informationsvakuum“, um falsche Narrative über die Identität des Angreifers (fälschlich als muslimischer Asylbewerber identifiziert) viral zu verbreiten. Diese Desinformation wurde durch algorithmische Verstärkung auf Plattformen wie X und TikTok massiv beschleunigt und führte unmittelbar zu gewalttätigen Angriffen auf eine Moschee und Unterkünfte von Migranten. Der Fall demonstriert, wie Online-Desinformation in emotional aufgeladenen Situationen als Verstärkungsfaktor wirkt, der latente gesellschaftliche Spannungen verstärken und in physische Gewalt übersetzen kann. Die Southport-Unruhen sind somit ein Beispiel für das, was in der Sicherheitsforschung als „stochastischer Terrorismus“ diskutiert wird: Die statistisch wahrscheinliche Auslösung von Gewalt durch massenhafte Aufhetzung, deren konkretes Ziel jedoch zufällig erscheint. Eine Untersuchung von Müller und Schwarz (2021) zeigt beispielsweise, dass eine Korrelation zwischen der lokalen Intensität von anti-migrantischer Desinformation in sozialen Medien und der Häufigkeit von Hassverbrechen in den entsprechenden Gebieten besteht.

Neben eruptiver Gewalt manifestieren sich die Schäden von Desinformation zunehmend in einer strukturellen Zersetzung der demokratischen Teilhabe, die in der Forschung als „Chilling Effects“ beschrieben wird. Der Council of Europe (2025) warnt in seinem aktuellen Bericht vor koordinierten Zermürbungsstrategien wie der „Operation Overload“, bei der journalistische und zivilgesellschaftliche Akteure gezielt mit einer Masse an falschen Anfragen geflutet werden, um ihre Ressourcen zu binden und sie zum Rückzug aus dem öffentlichen Diskurs zu bewegen. Besonders schwerwiegend wirkt dieser Mechanismus im Bereich der „Gendered Disinformation“. Dieses Phänomen wird auch als „Networked Misogyny“ bezeichnet: Es handelt sich hierbei nicht um isoliertes Trolling, sondern um koordinierte, oft sexualisierte Kampagnen (einschließlich Deepfake-Pornografie), die gezielt instrumentalisiert werden, um Frauen in Politik und Journalismus zu diskreditieren (vgl. Di Meco & Brechenmacher, 2020). Dies wird als Technology-facilitated gender-based violence (TFGBV) klassifiziert. Das Ziel dieser Maßnahmen besteht primär in der Verdrängung der betroffenen Akteurinnen aus dem öffentlichen Diskurs. (OECD, 2025; Anstis & LaFlèche, 2025). Die empirische Evidenz belegt, dass betroffene Politikerinnen sich signifikant häufiger aus Online-Diskursen zurückziehen als ihre männlichen Kollegen. Dieser Trend zeigt sich auch in prominenten Rücktritten wie dem der niederländischen Vize-Premierministerin Sigrid Kaag und führt zu einer messbaren Verarmung der repräsentativen Demokratie.

Die Auswirkungen auf Wahlen insgesamt sind subtil, aber signifikant. Mauk und Grömping (2024) weisen nach, dass Desinformation zwar nicht zwingend das Wahlergebnis direkt beeinflusst, den Glauben an die Fairness des Wahlprozesses jedoch deutlich untergraben kann. Dies gilt sowohl für Gewinner als auch für Verlierer. Ökonomisch betrachtet wird Desinformation im Global Risks Report 2025 des Weltwirtschaftsforums erneut als eines der größten kurzfristigen Risiken benannt, da sie Märkte destabilisieren

und das für Investitionen notwendige gesellschaftliche Vertrauenskapital vernichten kann. Schätzungen gehen davon aus, dass Desinformation die Weltwirtschaft jährlich bis zu 78 Milliarden Dollar kostet, etwa durch Reputationsschäden für Unternehmen oder Marktmanipulationen. Im Gesundheitssektor beziffern Studien allein die Kosten der impfbezogenen Desinformation (z. B. durch zusätzliche Behandlungen und Ausbrüche vermeidbarer Krankheiten) auf Milliardenbeträge (ECFSN, 2025a).

Ein weiteres Risiko liegt in der Beschädigung des epistemischen Vertrauens. Der Global Risks Report 2025 des World Economic Forum identifiziert Desinformation als eines der Top-Risiken für den gesellschaftlichen Zusammenhalt. Abschließend weist die aktuelle Forschung auf ein kritisches Paradoxon hin: Der gut gemeinte Kampf gegen Desinformation kann selbst zum Problem werden. Hoes et al. (2025) belegen in einer aktuellen Studie, dass nicht nur die Falschinformation selbst, sondern auch alarmistische Berichterstattung über sie negative Effekte haben kann. Die Studie zeigt, dass eine intensive mediale Warnung vor Desinformation („Wir sind umzingelt von Lügen“) bei Bürgern zu einer generellen „epistemischen Unsicherheit“ führt. Das Resultat ist nicht zwingend eine höhere Wachsamkeit, sondern ein pauschaler Vertrauensverlust gegenüber allen Informationsquellen, einschließlich Wissenschaftlern und etablierten Medien. Dieses Phänomen mahnt zur Vorsicht: Eine undifferenzierte Skandalisierung von Desinformation kann das Vertrauen in den öffentlichen Diskurs ebenso nachhaltig beschädigen wie die Desinformation selbst.

2.5 Erkennung

Die effektive Bewältigung der Desinformationskrise erfordert einen integrativen Ansatz, der technologische Detektionsverfahren mit verhaltenspsychologischen und strukturellen Interventionen verknüpft. Die Bekämpfungsansätze in der Plattformlandschaft und ein detaillierter Überblick über aktuelle Erkenntnisse erfolgen in den 3.3.2 bzw. 4.1.

Die automatisierte Detektion von Desinformation hat sich in den letzten Jahren von einfachen Keyword-Filtern zu hochkomplexen, KI-gestützten Systemen entwickelt. Während sich traditionelle Ansätze primär auf die Textanalyse mittels Natural Language Processing (NLP) konzentrierten, gelten rein unimodale Verfahren angesichts der multimedialen Natur moderner Desinformation mittlerweile als unzureichend. Der aktuelle Forschungsstand favorisiert daher den Einsatz multimodaler Deep-Learning-Modelle. Diese koppeln Transformer-Architekturen wie BERT für die Textanalyse mit Convolutional Neural Networks (CNNs) für die Bild- und Videoverarbeitung, um semantische Inkonsistenzen zwischen verschiedenen Modalitäten zu identifizieren (Nasser et al., 2025). Ergänzend hierzu gewinnen Knowledge Graphs an Bedeutung. Diese strukturieren faktisches Wissen in Form von Entitäten und Relationen, was es Algorithmen ermöglicht, Falschbehauptungen durch einen automatisierten Abgleich mit verifizierten Datenbanken zu flaggen. Zwar reduziert dieser Ansatz die Abhängigkeit von menschlichen Faktenprüfern, er

birgt jedoch das Risiko, dass KI-Modelle bei fehlendem Kontext kulturelle Nuancen wie Satire missverstehen oder „halluzinieren“ (Feng et al., 2025).

Da viele problematische Inhalte juristisch im Graubereich liegen („lawful but awful“), verschiebt sich der Fokus der Plattformbetreiber und Regulierungsbehörden zunehmend von der inhaltlichen Prüfung hin zur Erkennung von CIB. Hierbei wird nicht der Wahrheitsgehalt der Aussage analysiert, sondern die Topologie ihrer Verbreitung. Durch den Einsatz von Graph Neural Networks (GNNs) lassen sich Cluster von Accounts identifizieren, die in unnatürlicher Synchronizität agieren, etwa um künstliche Graswurzelbewegungen („Astroturfing“) zu simulieren (Feng et al., 2025).

Die Debatte um die effektive Bekämpfung von Desinformation leidet traditionell unter einem Mangel an standardisierten Metriken, die über bloße Inhaltsmoderations-Statistiken hinausgehen. Während Plattformen heute Transparenzberichte veröffentlichen, die die Anzahl entfernter Inhalte oder gesperrter Konten ausweisen, sagen diese operativen Kennzahlen wenig über das tatsächliche Ausmaß der Exposition der Nutzer gegenüber schädlichen Inhalten aus. Eine evidenzbasierte Regulierung erfordert daher eine Verschiebung des analytischen Fokus: weg von der reinen Zählung moderierter Inhalte hin zur Quantifizierung der algorithmischen Verstärkung und der systemischen Ursachen von Verbreitung.

Ein mögliches Instrument zur Messung der plattformspezifischen Verantwortung ist der von Allen (2022) entwickelte Misinformation Amplification Factor (MAF). Dieser Indikator operationalisiert die Frage, inwieweit das Design einer Plattform die Verbreitung von Falschinformationen begünstigt, indem er die tatsächliche Reichweite eines Desinformationsbeitrags ins Verhältnis zu der Reichweite setzt, die basierend auf der organischen Follower-Struktur des Urhebers zu erwarten wäre. Die Berechnung des MAF offenbart signifikante Unterschiede in den Plattformarchitekturen und bestätigt die Hypothese, dass Desinformation in Systemen, die auf friktionsloser Verbreitung und algorithmischen Empfehlungen basieren, stärker floriert als in Netzwerken, die primär auf dem sozialen Graphen (Follower-Beziehungen) beruhen.

Darauf basierende empirische Analysen zeigen beispielsweise, dass Plattformen wie TikTok und Twitter (heute X) tendenziell sehr hohe Verstärkungsfaktoren aufweisen. Dazu tragen Mechanismen wie das Retweeten mit einem Klick oder der „For You“-Feed bei, die Inhalte unabhängig von der Vertrauenswürdigkeit oder der etablierten Basis des Absenders viral verbreiten. Im Gegensatz dazu weisen Plattformen wie Instagram oder Facebook (in ihrem klassischen Feed) niedrigere MAF-Werte auf, da dort die Verbreitung stärker an die direkte Gefolgschaft gebunden ist und Hürden bei der Weiterverbreitung bestehen. Allerdings zeigt sich auch hier eine dynamische Verschlechterung, sobald Formate wie „Reels“ eingeführt werden, die analog zu TikTok stark auf algorithmische Empfehlungen unverbundener Inhalte setzen und dadurch den Verstärkungsfaktor für Desinformation signifikant erhöhen können.

Trotz der regulatorischen Fortschritte durch den Digital Services Act (DSA) und die verpflichtenden Risikobewertungen besteht weiterhin eine erhebliche Lücke zwischen den notwendigen Daten und den von den Plattformen bereitgestellten Informationen. Eine umfassende Auditierung der Risikobewertungen aus dem Jahr 2024 durch das Integrity Institute (Allen et al., 2025) zeigt, dass die Plattformen zwar qualitative Beschreibungen ihrer Systeme und Mitigationsmaßnahmen liefern, bei der Quantifizierung der entscheidenden Risikodimensionen jedoch deutliche Lücken aufweisen. Besonders betrifft dies Ausmaß (Scale), Ursache (Cause) und Beschaffenheit (Nature). Bezüglich des Ausmaßes beschränken sich die Anbieter oft auf relative Prävalenzraten (z. B. 0,01 % der Ansichten sind verletzend), ohne absolute Zahlen zu nennen, die das tatsächliche Volumen der Exposition verdeutlichen würden. Schätzungen legen jedoch nahe, dass selbst geringe Prozentwerte in absoluten Zahlen Milliarden von Ansichten schädlicher Inhalte bedeuten können. So lassen Daten von TikTok darauf schließen, dass dort pro Quartal etwa 30 Milliarden Ansichten auf Inhalte entfallen, die später aufgrund von Richtlinienverstößen entfernt werden. Auch bei YouTube wird das Volumen verletzender Inhalte auf mehrere Milliarden Ansichten pro Quartal geschätzt. Diese Diskrepanz zwischen niedrigen Prozentwerten und massiven absoluten Abrufzahlen verdeutlicht, dass die bloße Angabe von Prävalenzraten das systemische Risiko verschleiert. Potenziell problematischer ist das Defizit bei der Analyse der Ursachen. Für eine wirksame Regulierung ist es essenziell zu verstehen, welcher Anteil der schädlichen Exposition einerseits auf bewusste Nutzerentscheidungen zurückgeht und welcher Anteil andererseits durch das Plattformdesign mit Funktionen wie proaktiven Empfehlungen oder Autoplay-Funktionen verursacht wird. Die Auditierung ergab, dass keine der großen Plattformen (Very Large Online Platforms, VLOPs) quantifizierbare Daten darüber liefert, wie hoch der Anteil der durch Empfehlungsalgorithmen generierten Expositionen gegenüber Desinformation ist.

Dies verhindert eine Bewertung der Frage, ob die Geschäftsmodelle der Plattformen die Risiken kausal treiben. Schließlich fehlt es an Transparenz hinsichtlich der Beschaffenheit (Nature) der Risiken, insbesondere bezüglich ihrer Konzentration. Durchschnittswerte verschleiern oft, dass Risiken in spezifischen Echokammern oder bei vulnerablen Gruppen disproportional auftauchen können („Pockets of Misinformation“). Es fehlen Daten darüber, ob bestimmte Nutzergruppen extrem hohen Dosen von Desinformation ausgesetzt sind und inwieweit algorithmische Systeme diese Konzentration verstärken. Die aktuelle Praxis der Plattformen, den Erfolg ihrer Maßnahmen primär durch operative Kennzahlen wie die Anzahl gelöschter Beiträge oder die Schnelligkeit der Moderation nachzuweisen, greift daher zu kurz. Eine hohe Anzahl an Löschungen kann paradoxerweise ein Indikator für das Versagen des Designs sein, die Verbreitung a priori zu verhindern. Wirksame Transparenz muss folglich nachweisen, dass Gegenmaßnahmen nicht nur die Symptome bekämpfen, sondern die algorithmische Verstärkung (MAF) senken und die kausale Rolle der Empfehlungssysteme bei der Verbreitung von Desinformation messbar reduzieren.

Während technische Erkennungsmethoden fortschreiten, zeigt sich, dass das Problem nicht rein technischer Natur ist. Die Ursachen liegen tiefer in den ökonomischen Strukturen der Plattformen, die im folgenden Kapitel analysiert werden.

3 Online-Plattformen in der Desinformationsökonomie

3.1 Ökonomische Anreizstrukturen

Die Verbreitung von Desinformation auf digitalen Plattformen ist nicht allein ein technologisches oder soziologisches Phänomen, sondern tief in den ökonomischen Strukturen der „Aufmerksamkeitsökonomie“ verwurzelt. Um die Resilienz demokratischer Systeme zu stärken, ist ein Verständnis der monetären Anreize unerlässlich. Diese Anreize bewegen Akteure von Plattformbetreibern über Werbetreibende bis hin zu einzelnen Influencern dazu, die Verbreitung schädlicher Inhalte direkt oder indirekt zu fördern.

Das dominante Geschäftsmodell der großen sozialen Medienplattformen (Big Tech) basiert auf der Maximierung der Verweildauer der Nutzer, um Verhaltensdaten zu extrahieren und Werbeplätze zu verkaufen. In diesem Modell fungiert das Nutzer-Engagement (Likes, Shares, Kommentare) als zentrale Währung. Allen (2022) beschreibt in seiner Analyse des „Natural Engagement Pattern“, dass Inhalte, die sich der Grenze zum Verbotenen nähern („Approaching the Line“) oder starke emotionale Reaktionen wie Wut hervorrufen, naturgemäß ein höheres Engagement erzielen als neutrale, faktische Informationen.

Dies führt zu einem fundamentalen Anreizkonflikt: Plattformen stehen vor der ökonomischen Wahl, entweder in Sicherheit und Moderation zu investieren und damit häufig das Engagement sowie den Umsatz zu drosseln oder die algorithmische Verbreitung von „High-Engagement“-Inhalten zuzulassen. Zu diesen Inhalten zählen überproportional häufig Desinformation und Hassrede. Eine mikroökonomische Studie von Dey et al. (2025) liefert hierzu eine Fundierung: Die Autoren zeigen modelltheoretisch, dass Polarisierung und Bias für Plattformen als Substitute bei der Profitmaximierung fungieren können. Um die Werbeeinnahmen zu maximieren, sind Plattformen ökonomisch motiviert, Nutzer in „Echo Chambers“ zu treiben und polarisierende Narrative zu bevorzugen, da dies die Bindung an die Plattform erhöht. Aktuelle Analysen zeigen, dass Plattformen wie X (ehemals Twitter) und TikTok besonders hohe „Misinformation Amplification Factors“ (MAF) aufweisen, was auf eine starke Ausrichtung der Empfehlungssysteme auf virale Verbreitung hindeutet. Dieser Mechanismus wird durch die Transformation von Verifikationssystemen in Bezahldienste („Pay-for-Play“) weiter verschärft. So ermöglicht das kostenpflichtige Verifikationsmodell von X es böswilligen Akteuren, durch bloße Zahlung algorithmische Priorität zu erkaufen, unabhängig von der Authentizität ihrer Identität oder Inhalte (Jones, 2025).

Ein oft unterschätzter Treiber der Desinformationsökonomie ist das globale Werbe-Ökosystem, insbesondere die programmatische Werbung. Programmatische Systeme versteigern Werbeplätze in Millisekunden über eine Kette von Demand-Side-Plattformen, Exchanges und Supply-Side-Plattformen. Diese Architektur koppelt Werbebudgets nicht mehr an konkrete redaktionelle Umfelder, sondern an Zielgruppen-Profile und Auktionsergebnisse. In der Folge landen Anzeigen großer Marken routinemäßig auf Webseiten,

die nachweislich Desinformation verbreiten (Ahmad et al., 2024; Braun & Eklund, 2019). Ahmad et al. (2024) zeigen, dass Unternehmen quer durch viele Branchen in erheblichem Umfang auf Misinformation-Websites werben, obwohl die Finanzierung solcher Seiten sowohl Reputations- als auch finanzielle Risiken birgt. Gleichzeitig dokumentieren sie, dass diese Fehlallokation nicht primär aus bewusster Zustimmung resultiert, sondern aus der Delegation von Budgetentscheidungen an automatisierte Ad-Tech-Plattformen und dazwischengeschaltete Agenturen, die auf Reichweite und Effizienz optimiert sind, nicht auf inhaltliche Qualität. In Interviews mit Akteuren der Werbeindustrie beschreiben Braun und Eklund (2019) programmatische Werbung als eine Infrastruktur, die „fake news“ und Qualitätsjournalismus über dieselben Pipeline-Systeme monetarisiert und so finanzielle Anreize für Clickbait-Desinformation schafft (Braun & Eklund, 2019).

Vor diesem Hintergrund lässt sich Desinformation als negative Externalität des digitalen Werbemarktes begreifen. Diaz Ruiz (2025) argumentiert, dass viele Marketingverantwortliche die schädlichen Folgen programmatischer Werbung als Randproblem betrachten. So wird etwa die Monetarisierung von Fake News ausgeblendet und normativ „aus dem Markt herausdefiniert“: Die Vorstellung, Technologie sei neutral und der Markt korrigiere sich selbst, verdecke, dass Werbeentscheidungen aktiv dazu beitragen, Desinformation zu finanzieren. McGowan et al. (2024) zeigen ergänzend, dass Ad-Tech-Intermediäre nicht bloß neutrale „Durchleiter“ sind, sondern als „Mediatoren“ den Markt mitgestalten. Durch Black-Box-Optimierungen und Bündelprodukte entscheiden sie faktisch mit, welche Inventarquellen, einschließlich problematischer Seiten, bevorzugt befüllt werden.

Parallel dazu hat sich eine explizite „Disinformation-for-Profit“-Industrie herausgebildet. Eine gemeinsame Untersuchung des Carter Center und des McCain Institute rekonstruiert anhand von Traffic- und Werbedaten, dass über 81 % der identifizierten Desinformations-Webseiten an programmatische Werbenetzwerke wie Google Ads angeschlossen sind und ihre Inhalte über automatisierte Werbeschaltungen finanzieren (Scholtens et al., 2024). Ein spezieller Teil dieses Ökosystems sind „Made-for-Advertising“ (MFA)-Seiten: Webseiten, deren Geschäftsmodell nahezu ausschließlich darin besteht, möglichst viele programmatische Impressionen zu generieren, oft mit minderwertigem, recyceltem oder KI-generiertem Content. Trotz jüngster Bemühungen großer Werbekunden, diese Ausgaben zu reduzieren, fließen weiterhin Hunderte Millionen US-Dollar pro Quartal in Umfelder, die häufig Clickbait, Fehlinformationen oder irreführenden KI-Content beherbergen. Untersuchungen von DoubleVerify und anderen Messdienstleistern dokumentieren zudem, dass MFA-Netzwerke zunehmend Sport-, Lifestyle- und Pseudo-News-Seiten mit AI-generierten Inhalten aufbauen, die bewusst seriöse Medien imitieren, um programmatische Werbeeinnahmen zu maximieren (DoubleVerify, 2024).

Dieser ökonomische Unterbau verschränkt sich mit der politischen Desinformationslandschaft. Programmatische Plattformen und Social-Media-Werbesysteme ermöglichen es politischen und ideologischen Akteuren, hochgranular Zielgruppen zu adressieren, ohne sich offen als Absender erkennen zu geben. Votta et al. (2024) zeigen in einer Analyse von Facebook- und Instagram-Werbung in 95 Ländern, dass politisches Microtargeting

mittlerweile global verbreitet ist und häufig nach dem Muster „einfache demografische Kriterien plus Interessen“ arbeitet. Im Kontext der Unruhen nach dem Anschlag von Southport (vgl. Kapitel 2.4.2) legen investigative Recherchen und parlamentarische Beweisaufnahmen nahe, dass über Monate hinweg anonyme Akteure beträchtliche Budgets in anti-muslimische und anti-immigrantische Kampagnen investierten, die über Meta-Plattformen ausgespielt wurden; in einem Bericht ist von rund 1,2 Millionen US-Dollar die Rede, die in die Verstärkung xenophober Narrative über Facebook-Werbung flossen (Jones, 2025; D’Souza, 2025). Meta konnte laut diesen Berichten die tatsächlichen Finanzierungsstrukturen und Auftraggeber nur begrenzt rekonstruieren, was eine Lücke in den „Know Your Customer“-Mechanismen der Werbeplattformen offenbart. Aus demokratietheoretischer Perspektive entsteht so eine Konstellation, in der private Unternehmen an der Verbreitung polarisierender und menschenfeindlicher Botschaften verdienen, während die Urheber hinter Briefkastenfirmen, Mittelsmännern und Agenturstrukturen verschwinden.

Die strukturelle Rolle des Ad-Tech-Systems ist damit doppelt problematisch: Einerseits ermöglicht es die Refinanzierung expliziter Desinformationsakteure, andererseits verschärft es die Anreize für alle Inhalteproduzenten, Aufmerksamkeit um (fast) jeden Preis zu erzeugen. Middleton (2025) # beschreibt das digitale Werbeökosystem in ihrer Eingabe an den britischen Wissenschafts- und Technologieausschuss als hochgradig intransparenten „Black Box“-Markt, in dem Werbebudgets durch eine Kette von Intermediären laufen, bevor sie auf einer konkreten Webseite ankommen. Durch diese Intransparenz können Marken trotz „Brand Safety“-Versprechen auf Seiten landen, die Hassrede, Verschwörungserzählungen oder manipulative Pseudo-Nachrichten verbreiten. McGowan et al. (2024) zeigen, dass Intermediäre ökonomisch davon profitieren, möglichst viel günstige Reichweite zu bündeln. Dadurch entsteht strukturell ein Bias zugunsten genau jener Umfelder wie MFA- und Desinformationsseiten, die viele Klicks bei niedrigen Kosten generieren.

Amnesty International (2025) weitet diese Kritik um eine menschenrechtliche Perspektive: In ihrer Stellungnahme zur britischen Parlamentsanfrage zu „Social Media, Misinformation and Harmful Algorithms“ argumentiert die Organisation, dass das Geschäftsmodell der großen Plattformen fundamental auf überwachungsbasierter Werbung („surveillance-based advertising“) beruht. Da personalisierte Werbung auf möglichst detaillierten Nutzerprofilen basiert, belohnen Empfehlungsalgorithmen Inhalte, die Aufmerksamkeit maximieren. Dazu zählen häufig solche, die Angst, Wut und Feindbilder aktivieren. In einem technischen Begleitpapier zu den UK-Riots zeigt Amnesty, wie Designentscheidungen von X (ehemals Twitter) rund um Empfehlungssysteme, Entdrosselung von Grenzinhalten und Monetarisierung von Reichweite dazu beitrugen, rassistische und anti-migrantische Narrative nach dem Angriff von Southport massiv zu verstärken. Die ökonomische Logik, möglichst viel Engagement für möglichst profitable Anzeigen zu erzeugen, ist damit eng verknüpft mit der algorithmischen Verstärkung von Desinformation und Hass.

Vor diesem Hintergrund rücken Regulierungs- und Interventionsvorschläge in den Fokus, die nicht nur Content-Moderation, sondern die finanzielle Infrastruktur selbst adressieren. Ahmad et al. (2024) zeigen, dass bereits relativ einfache, informationsbasierte Interventionen die Finanzierung von Misinformation signifikant senken können. Dazu gehören etwa Transparenzkampagnen gegenüber Werbekunden, die offenlegen, auf welchen Seiten ihre Anzeigen erscheinen, ohne die Gesamtwirksamkeit der Werbung erheblich zu beeinträchtigen. Scholtens et al. (2024) plädieren auf Basis ihrer Analyse der „Disinformation Economy“ dafür, gezielt Zahlungsströme zu Desinformationsseiten durch strengere Brand-Safety-Standards, Blacklists und eine stärkere Verantwortung der Intermediäre zu unterbrechen. Diaz Ruiz (2025) betont ergänzend, dass Marketing-Communitys ihre eigenen „soziotechnischen Imaginationen“ hinterfragen müssen, die die negativen Folgen programmatischer Werbung bislang als externe, nicht-marktliche Effekte abtun.

Amnesty International (2025) geht einen Schritt weiter und fordert eine strukturelle Abkehr von profilbasierter Werbung. Aus menschenrechtlicher Sicht sei ein Geschäftsmodell, das auf massiver Datensammlung, Profilbildung und anschließender Microtargeting-Werbung basiert, nicht mit dem Schutz vor Diskriminierung, Überwachung und manipulativer Beeinflussung vereinbar. Stattdessen spricht sich Amnesty für datensparsame, kontextbasierte Werbemodelle, verbindliche menschenrechtliche Sorgfaltspflichten und deutlich strengere Transparenz- und Auskunftsrechte gegenüber Plattformen aus. In der Summe deutet die aktuelle Forschung darauf hin, dass ein wirksamer Kampf gegen Desinformation nicht bei der Korrektur einzelner Inhalte stehenbleiben kann: Er muss die ökonomischen Anreizstrukturen des Ad-Tech-Systems ins Zentrum rücken, die Reichweite von Desinformation monetarisieren. Maßnahmen zur Nutzerresilienz und algorithmischen Kuration bleiben wichtig, entfalten ihr volles Potenzial aber erst dann, wenn gleichzeitig die Finanzierungsströme zu Desinformationsakteuren systematisch und nachhaltig gekappt werden (Ahmad et al., 2024; Diaz Ruiz, 2025; Middleton, 2025).

Neben den Plattformen und Werbenetzwerken haben sich individuelle Akteure professionalisiert, die von der Polarisierung profitieren. Jones (2025) führt den Begriff der „Disinfluencer“ ein: Akteure, die routinemäßig Falschinformationen verbreiten und aufgrund ihrer Reichweite einen unverhältnismäßig großen Einfluss auf Trends haben. Diese Akteure operieren oft in einem ökonomischen Umfeld, das extremes Verhalten belohnt. Ein markantes Beispiel ist die Änderung des Monetarisierungsmodells von X im Oktober 2024, bei dem Creator direkt auf Basis des Engagements (Views, Likes) bezahlt werden, das ihre Beiträge generieren. Da Desinformation und Hassrede signifikant höhere Interaktionsraten erzielen, schafft dieses Modell einen direkten finanziellen Anreiz für die Produktion von „Rage-Bait“ und polarisierenden Inhalten.

Neuere Forschungen differenzieren hierbei genauer zwischen verschiedenen Rollen. Pournaki et al. (2025) unterscheiden zwischen „Influencern“, die Inhalte generieren, und „Multiplikatoren“, die Inhalte kuratieren und verstärken. Ihre Analyse zeigt, dass insbesondere die Multiplikatoren eine entscheidende Rolle bei der Bündelung und Verstärkung

ideologisch konsistenter und polarisierender Inhalte spielen, was die Bildung von Echokammern beschleunigt. Abdul Rahman et al. (2025) ergänzen diese Perspektive durch die Analyse von „Amplifiern“ (Verstärkern), die oft als Katalysatoren für Belästigungskampagnen („Indirect Swarming“) fungieren, ohne selbst explizit gegen Richtlinien zu verstoßen. Der ökonomische Wert dieser Amplifier liegt in ihrer Fähigkeit, Aufmerksamkeit zu lenken und Diskurse zu framen. Prominente Beispiele wie Elon Musk zeigen, wie Plattform-Besitzer selbst zu „Super-Amplifiern“ werden können, indem sie durch Interaktionen mit Randgruppen-Accounts deren Sichtbarkeit und damit deren Monetarisierungspotenzial massiv erhöhen.

Schließlich hat sich die Erstellung und Verbreitung von Desinformation zu einer globalen Dienstleistungsindustrie entwickelt. Die in Kapitel 2.2 beschriebenen Praktiken des koordinierten inauthentischen Verhaltens (CIB) werden hierbei zunehmend professionalisiert und als „Disinformation-as-a-Service“ (DaaS) angeboten (vgl. CACI, 2025). Externe Akteure, wie etwa PR-Firmen oder Cyberkriminelle, bieten im Dark Web oder auch offen gegen Bezahlung die Manipulation öffentlicher Diskurse an. Jones (2025) verweist auf den Fall einer in Tel Aviv ansässigen Firma, die KI-generierte anti-muslimische Inhalte gezielt an westliche Zielgruppen ausspielte. Diese Akteure nutzen die Kosteneffizienz generativer KI, um Inhalte (Text, Bild, Video) massenhaft und kostengünstig zu produzieren. Dies senkt die Eintrittsbarrieren für Desinformationskampagnen erheblich und ermöglicht es, Inhalte hochgradig auf spezifische demografische Zielgruppen zuzuschneiden („Microtargeting“). In Verbindung mit „Click-Farms“ oder Bot-Netzwerken entsteht so eine Schattenökonomie, die nicht primär ideologisch, sondern als Auftragsarbeit funktioniert und die Integrität des digitalen Raums aus reinem Profitinteresse untergräbt.

3.2 Algorithmische Verstärkung

Während Kapitel 3.1 die ökonomischen Anreizstrukturen darlegte, analysiert dieses Kapitel die technologische Umsetzung. Empfehlungssysteme sind die operationalen Hebel, die ökonomische Ziele in kuratorische Entscheidungen übersetzen. Sie fungieren nicht als neutrale Distributoren, sondern als aktive „Algorithmic Gatekeepers“, die redaktionelle Urteilskraft durch automatisierte Metriken ersetzen und dabei spezifische Inhalte systemisch bevorteilen (Chiridza & Mare, 2025).

Die (implizite) Verstärkung von Desinformation resultiert oft nicht aus einer bewussten Programmierung auf Falschheit, sondern aus der mathematischen Logik popularitätsbasierter Ranking-Funktionen. Eine aktuelle agentenbasierte Simulationsstudie (Jakobsen et al., 2025) zeigt, dass Algorithmen, die primär auf Popularitätsmetriken (Views, Likes) optimiert sind, die Verbreitung von Desinformation signifikant beschleunigen, da diese Inhalte oft auf Neuartigkeit und emotionale Reize hin optimiert sind und somit schneller erste Interaktionsschwellen überschreiten. Im Gegensatz dazu könnten Ansätze wie das Item-based Collaborative Filtering, die auf Ähnlichkeitsmustern zwischen Inhalten

basieren, die Exposition gegenüber Falschinformationen technisch sogar begrenzen, werden jedoch seltener als primärer Ranking-Faktor eingesetzt

Ein weiteres technisches Phänomen, das die Verbreitung von Desinformation begünstigt, ist die sogenannte „Sycophancy“ in modernen KI-Modellen, was eine Tendenz von KI-Modellen zu konformistischem Antwortverhalten beschreibt. Untersuchungen im medizinischen Kontext zeigen, dass Large Language Models (LLMs), die mittels Reinforcement Learning from Human Feedback (RLHF) trainiert wurden, dazu neigen, Nutzereingaben zuzustimmen, selbst wenn diese objektiv falsch sind. Algorithmen priorisieren hier eine falsch verstandene Hilfsbereitschaft und Nutzerbestätigung über faktische Genauigkeit, was dazu führt, dass sie Verschwörungsnarrative oder medizinische Fehlannahmen von Nutzern nicht korrigieren, sondern algorithmisch bestärken können.

Chen et al. (2025) demonstrierten dies experimentell, indem sie KI-Modelle aufforderten, medizinisch unsinnige Warnungen vor generischen Medikamenten zu verfassen. Anstatt den logischen Fehler zu korrigieren, erfüllten die Modelle die Anfrage in 58 bis 100 % der Fälle und halluzinierten überzeugende, aber falsche Argumente, um die Fehlannahme des Nutzers zu stützen. Zwar ließe sich dieses Verhalten durch technisches „Supervised Fine-Tuning“ oder spezifische Prompting-Strategien (z. B. „Rejection Permission“) mindern, doch fehlt Entwicklern von General-Purpose-LLMs oft der ökonomische Anreiz zur Implementierung, da Nutzer die Interaktion mit „höflichen“ und bestätigenden Chatbots präferieren. Diese technische Disposition zur Bestätigung des Nutzer-Bias („Algorithmic Confirmation“) schafft eine sich selbst verstärkende Rückkopplungsschleife, die Nutzer tiefer in Informationsblasen treibt.

Die generelle algorithmische Logik ist darauf ausgerichtet, Interaktion zu provozieren. Dies führt dazu, dass sensationalistische, emotional aufgeladene oder polarisierende Inhalte systematisch bevorzugt und verstärkt werden, oft auf Kosten der Genauigkeit oder Sicherheit. Empfehlungssysteme sind keine passiven oder neutralen Mechanismen. Interne Dokumente von Meta (Facebook Papers) bestätigten, dass das Ranking des News Feeds auf vorhergesagtem Engagement basierte und separate Modelle für Inhalte existierten, die Nutzer wahrscheinlich "weiterleiten" würden, was den Kern der algorithmischen Viralität bildet. Die Algorithmen sind daraufhin optimiert, erregendes, emotional aufgeladenes oder polarisierendes Material zu verstärken. Die in Kapitel 2.4.2 dargelegten Unruhen von Southport illustrieren diese Datenlage: Anti-Muslimische und Anti-Immigranten-Inhalte erhielten 65 % der Gesamteindrücke der mit den Unruhen verbundenen Postings, während faktische Korrekturen nur 13 % erreichten. Auch TikToks "For You" Feed wird dafür kritisiert, Jugendliche, die Interesse an mentaler Gesundheit zeigen, durch sukzessive Empfehlungen zunehmend extremeren oder potenziell selbstverletzenden Inhalten auszusetzen (vgl. „Rabbit Holes“).

Die Vorhersehbarkeit dieser algorithmischen Präferenzen ermöglicht es externen Akteuren, die Systeme gezielt zu manipulieren („Adversarial Attacks“). Ein zentraler Mechanismus ist hierbei die Cross-Platform-Coordination. Aktuelle Analysen zu US-Wahlen

zeigen, dass Desinformationskampagnen selten isoliert stattfinden, sondern als mehrschichtige Netzwerke operieren. Inhalte werden auf gering regulierten Plattformen (z. B. Telegram, 4chan) „gesät“ und dann koordiniert auf Mainstream-Plattformen (X, Facebook) geteilt, um künstliche Viralität zu erzeugen (Hristakieva et al., 2024; Tardelli et al., 2024). Diese Strategie nutzt die Trägheit der plattforminternen Detektionssysteme aus, die oft nur lokale Signale (auf der eigenen Plattform) auswerten, aber die koordinierte externe Amplifikation übersehen.

Ein weiterer Mechanismus systematischer Engagement-Verstärkung ist das indirekte Swarming, bei dem Amplifier ihre Follower durch maskierte Sprache oder das bloße Retweeten einer Nachricht mobilisieren können, um ein Ziel zu belästigen, ohne direkt die Nutzungsbedingungen der Plattform zu verletzen. Zusätzlich bedienen sich Akteure zunehmend der „Keyword Obfuscation“ (Verschleierung). Die Nutzung von „Algospeak“ umgeht automatische Systeme der Inhaltserkennung. Hierbei werden Begriffe bewusst abgewandelt oder visuelle Codes eingesetzt, sodass die Botschaft für die menschliche Zielgruppe weiterhin decodierbar bleibt. Neue Forschungen deuten darauf hin, dass generative KI diesen Prozess automatisiert („AI-driven obfuscation“), indem sie Inhalte massenhaft so variiert, dass sie unterhalb der Erkennungsschwellen der Moderationsalgorithmen bleiben (Romanishyn et al., 2025).

Jenseits der rein ökonomischen Logik oder der opportunistischen Ausnutzung durch Dritte rückt zunehmend die politische Instrumentalisierung der algorithmischen Kuration durch die Plattformbetreiber selbst in den Fokus. Dabei ist ein signifikanter Wandel in der politischen Stoßrichtung und Wahrnehmung dieser Eingriffe zu beobachten. Bis weit in die frühen 2020er Jahre wurde die Debatte um Content Moderation primär von konservativer und rechter Seite dominiert, die den großen Plattformen einen systematischen „Liberal Bias“ und die Zensur konservativer Stimmen vorwarfen. Ein zentrales, aber oft opakes Instrument in dieser Diskussion ist das „Shadowbanning“ (die heimliche Reduzierung der Sichtbarkeit). Während Plattformen dies oft als technische Notwendigkeit gegen Spam rechtfertigen, zeigen Studien, dass die Kriterien für solche Drosselungen - unabhängig von der generellen Stoßrichtung- intransparent bleiben und das Vertrauen in den digitalen Diskurs erodieren können (Delmonaco et al., 2024; Thomas & Manalil, 2025). Shadowbanning fungiert als unsichtbares Regulierungsinstrument, das potenziell für politische Zensur („Soft Censorship“) missbraucht werden kann, indem unliebsame, aber legale Inhalte („lawful but awful“) systematisch unterdrückt werden. Dies kann zu einer Unsicherheit bei Nutzern führen und Polarisierung verstärken (Chen & Zaman, 2024).

Eine differenzierte Bewertung von Shadowbanning erfordert die Abwägung zwischen ökonomischer Effizienz und regulatorischer Transparenz. Im Gegensatz zur strikten Content-Entfernung („Removal“), die extreme Nutzer von der Plattform verdrängt, fungiert Shadowbanning als ein Instrument der Marktmaximierung: Es ermöglicht Plattformen, auch Nutzer mit extremen Ansichten als Teil der User-Basis zu halten, da diese im Unklaren darüber gelassen werden, dass ihre Inhalte für andere unsichtbar sind.

Ökonomisch betrachtet führt diese Strategie häufig zu höheren Plattform-Profiten und einer größeren Marktabdeckung als die Löschung von Inhalten, da sie die „Posting Utility“ extremer Nutzer erhält, während sie gleichzeitig die „Reading Disutility“ (das Unbehagen) moderater Nutzer durch das Ausblenden dieser Inhalte minimiert. Die Modellierung zeigt sogar, dass Shadowbanning unter bestimmten Bedingungen die gesamte soziale Wohlfahrt („Social Welfare“) stärker erhöhen kann als transparente Löschverfahren. Dies gilt insbesondere dann, wenn die Nutzerannahmen über dessen Verbreitung moderat ausfallen. Dies erzeugt jedoch ein regulatorisches Spannungsfeld: Während aktuelle Gesetze wie der Digital Services Act (DSA) auf maximale Transparenz drängen, ist die Opazität (Undurchsichtigkeit) gerade das funktionskonstituierende Merkmal des Shadowbannings. Ein pauschales Verbot dieser Praxis könnte daher unbeabsichtigte negative Wohlfahrtseffekte haben, weshalb Hojati und Nault (2025) statt eines Verbots eine präzise Definition legitimer Einsatzszenarien (z. B. gegen Bots oder Spam) empfehlen, um die Vorteile der Reichweitensteuerung zu nutzen, ohne das Nutzervertrauen durch Willkür zu zerstören.

Der Vorwurf einer strukturellen Benachteiligung rechter Inhalte konnte durch empirische Untersuchungen kaum bestätigt werden. Im Gegenteil, aktuelle Audits zur algorithmischen Amplifikation, etwa während der US-Wahlen 2024, zeigen eine gegenteilige Tendenz: Plattformen wie X (ehemals Twitter) weisen inzwischen eine signifikante „Exposure Inequality“ auf, bei der insbesondere rechtsgerichtete Nutzer („Right-Leaning Users“) überproportional stark mit Inhalten konfrontiert werden, die ihre eigene politische Haltung bestätigen, während der Zugang zu Gegenargumenten algorithmisch minimiert wird (Ye et al., 2025). Auch Milli et al. (2025) wiesen nach, dass engagement-basierte Algorithmen systematisch Inhalte bevorzugen, die Wut und Feindseligkeit gegenüber der politischen „Out-Group“ ausdrücken. Dieser Mechanismus begünstigt polarisierende Akteure.

In der aktuellen Phase der Plattform-Governance (seit ca. 2023/24) zeichnet sich eine neue Dynamik ab, die weniger durch eine liberale Hegemonie als vielmehr durch eine strategische Anpassung an rechte Machtzentren gekennzeichnet ist. Forscher beschreiben dies mit dem von Timothy Snyder geprägten Begriff des „Anticipatory Obedience“ (vorauseilender Gehorsam) (Bassin & Potter, 2024; Snyder, 2017). Angesichts drohender regulatorischer Eingriffe oder politischer Repressalien durch rechtsgerichtete Regierungen (wie unter einer Trump-Administration) neigen Tech-Konzerne und Medienbesitzer dazu, ihre Moderationsrichtlinien proaktiv anzupassen, um Konflikte zu vermeiden (The Guardian, 2024).

Marc Owen Jones dokumentiert, wie Plattformbesitzer selbst zu politischen Akteuren werden (Jones, 2025). Im Fall von Elon Musk und X manifestierte sich dies in der Re-Platforming-Strategie für zuvor gesperrte rechtsextreme Akteure und der algorithmischen Amplifikation spezifischer politischer Narrative. Wenn algorithmische Sichtbarkeit nicht mehr auf neutralen Engagement-Metriken, sondern auf intransparenten politischen Parametern („Bias Injection“) beruht, wandelt sich die Plattform von einem Marktplatz der Aufmerksamkeit zu einem Instrument der verdeckten Meinungssteuerung (Dey et al.,

2025). Diese „Strategic Accommodation“ an autoritäre Tendenzen stellt eine fundamentale Verschiebung dar: Die Gefahr droht nicht mehr primär durch „Over-Blocking“ (Zensur), sondern durch die selektive, politisch opportunistische Förderung von Inhalten, die sich an den Interessen der politischen Macht orientiert.

Die Annahme, dass Plattform-Algorithmen durch die Maximierung von Engagement unbeabsichtigt Polarisierung fördern, wird durch aktuelle empirische Daten aus dem deutschen Bundestagswahlkampf 2025 eindrücklich bestätigt. Zwei aktuelle Audits der Bertelsmann-Stiftung zeigen, dass die technische Architektur der Plattformen nicht nur neutrale „Spiegel“ der politischen Debatte sind, sondern als verzerrende Verstärker wirken, die spezifische politische Stile und Akteure systematisch bevorteilen.

Eine zentrale Erkenntnis der aktuellen Forschung ist, dass Algorithmen eine inhärente Präferenz für Negative Campaigning besitzen. Die Analyse des Progressiven Zentrums (2025) von über 30.000 politischen Kurzvideos belegt, dass Inhalte, die politische Gegner angreifen oder herabwürdigen, mit einem Reichweiten-Bonus von 40 % belohnt werden. Konstruktive Ansätze oder positive Selbstdarstellung werden hingegen algorithmisch „bestraft“ und erzielen signifikant weniger Sichtbarkeit.

Diese Resultate stehen im Einklang mit internationaler Forschung zur „Optimization for Divisiveness“. Orecchia (2025) zeigt, dass Algorithmen, die auf Engagement-Metriken trainiert sind, zwangsläufig Inhalte priorisieren, die moralische Empörung („Outrage“) und Konflikt triggern, da diese Reaktionen kognitiv schneller und intensiver erfolgen als Zustimmung. Dies erzeugt einen Feedback-Loop: Politische Akteure lernen durch „Trial and Error“, dass Aggressivität die einzige Währung ist, um im Feed stattzufinden, und passen ihre Kommunikation entsprechend an („Algorithmic Accommodation“).

Die algorithmische Bevorzugung ist zudem teils nicht politisch neutral verteilt. Ein „Sock-Puppet-Audit“ der Bertelsmann Stiftung (2025) zur Bundestagswahl 2025 offenbart eine massive Asymmetrie zugunsten der AfD. Auf TikTok hatten 50 % aller parteipolitischen Inhalte, die jungen, unentschlossenen Nutzern (21-25 Jahre) im „For You“-Feed angezeigt wurden, einen Bezug zur AfD. Die Partei war damit dreimal sichtbarer als die CDU/CSU (15 %). Der Algorithmus führt Nutzer auch extrem schnell zu rechten Inhalten. Ein AfD-Video wurde durchschnittlich bereits nach 11-12 Minuten Nutzungszeit empfohlen, während Inhalte der SPD oder FDP oft erst nach über einer Stunde oder gar nicht erschienen.

Diese Dominanz lässt sich laut der Studie nicht allein durch eine höhere Posting-Frequenz der AfD erklären. Die SPD produzierte im Untersuchungszeitraum beispielsweise mehr Videos, erzielte aber dennoch signifikant weniger algorithmische Empfehlungen. Dies legt nahe, dass der Algorithmus eine spezifische „Passung“ zwischen den Inhaltsmustern der AfD (z. B. nativistische Narrative, Angst-Framing bei Migration) und seinen eigenen Optimierungszielen (Verweildauer, emotionale Aktivierung) erkennt. Ye et al. (2025) bestätigen dieses Muster auch für die USA: Ihre Audit-Studie auf X zeigte, dass

Algorithmen rechtsgerichtete Nutzer systematisch in entsprechende Filterblasen führten, während neutrale Nutzer schneller mit rechten als mit linken Inhalten konfrontiert wurden (Ye et al., 2025).

Algorithmen greifen dabei auch indirekt in die thematische Agenda ein. Die Analyse des Progressiven Zentrums zeigt, dass Videos zum Thema Migration systematisch mit einem Reichweiten-Bonus (+11 %) versehen werden, während komplexe Zukunftsthemen wie Umwelt (-18 %) oder Bildung (-17 %) in der Sichtbarkeit gedrosselt werden. Dies benachteiligt strukturell Parteien, die auf differenzierte Sachthemen setzen, und bevorteilt Akteure, die „Wedge Issues“ (Spaltthemen) bewirtschaften. Diese technisch erzeugten Sichtbarkeitsasymmetrien treffen auf ein sich wandelndes Rezeptionsverhalten der Nutzer, welches die algorithmischen Effekte verstärkt (vgl. Kapitel 2.4.1).

3.3 Aktuelle technologische Entwicklungen

Die Dynamik der Desinformation wird derzeit durch drei konvergierende Faktoren neu justiert: die massenhafte Verfügbarkeit generativer KI, den strategischen Rückzug der Plattformen aus der Moderationsverantwortung und die zunehmende Fragmentierung der Regulierungslandschaft. Diese Entwicklungen verändern nicht nur die Produktionsbedingungen von Falschinformationen, sondern auch die Architektur des öffentlichen Diskurses.

Die Generative KI stellt eine zusätzliche Herausforderung dar, da sie zum Einen die Zeit zur Erstellung überzeugender, aber gefälschter Bilder erheblich verkürzt, was auch die Detektion erschwert. Während der Southport-Unruhen wurde KI-generierter Inhalt genutzt, um xenophobe Narrative zu verstärken. Die Diskussion um Künstliche Intelligenz in der Desinformation hat sich von der Angst vor perfekten „Deepfakes“ allerdings teils hin zu einem ökonomischen Skalierungsproblem verschoben. Während technisch hochkomplexe, hyperrealistische Fälschungen weiterhin eine Gefahr für gezielte Angriffe darstellen, identifiziert die aktuelle Forschung eine subtilere, aber systemisch womöglich sogar gravierendere Bedrohung: die Flutung des digitalen Informationsraums mit massenhaft produziertem, qualitativ minderwertigem KI-Inhalt, der als „AI Slop“ bezeichnet wird.

3.3.1 AI Slop

Der Begriff „AI Slop“ beschreibt eine neue Kategorie digitaler Inhalte, die durch generative KI in hoher Geschwindigkeit und ohne nennenswerte menschliche Qualitätskontrolle oder kuratorische Sorgfalt erstellt werden. Madsen und Puyt (2025) definieren Slop als „generative waste“ (künstlich generierten Niedrigqualitäts-Output), der sich durch sieben Dimensionen auszeichnet, darunter ein enormes Volumen, hohe Geschwindigkeit der Verbreitung sowie eine Erosion des kulturellen und epistemischen Werts. Im Gegensatz zu gezielter Desinformation, die oft präzise narrative Ziele verfolgt, zeichnet sich Slop primär durch seine Banalität und Allgegenwärtigkeit aus: Es handelt sich um synthetischen

„Füllstoff“, der von generischen Essays und Clickbait-Blogs bis hin zu bizarren, KI-generierten Bildern reicht.

Ein prominentes Beispiel für die virale Verbreitung synthetischer Inhalte auf sozialen Plattformen ist das Phänomen des „Shrimp Jesus“. Dabei handelt es sich um einen Fall KI-generierter Bildmotive, die religiöse Ikonografie mit ungewöhnlichen Elementen kombinieren und auf Facebook hohe Interaktionsraten erzielten. Obwohl diese Inhalte auf den ersten Blick harmlos wirken mögen, fungieren sie als Vehikel, um algorithmische Empfehlungssysteme zu sättigen und Aufmerksamkeit zu binden. Die Verbreitung ist dabei nicht auf soziale Medien beschränkt. Ansari (2025) weist darauf hin, dass bereits im Mai 2025 schätzungsweise 52 % neuer Online-Artikel maschinell generiert waren, was eine signifikante Kontamination des gesamten Informationsökosystems darstellt.

Die Produktion von AI Slop wird durch eine klare ökonomische Logik getrieben. Da generative KI die Grenzkosten für die Erstellung von Inhalten auf nahezu null senkt, können Akteure die Aufmersamkeitsökonomie der Plattformen mit minimalem Aufwand ausnutzen („Spam 2.0“). Diese Dynamik wird politisch instrumentalisiert, ein Phänomen, das Klincewicz et al. (2025) als „Slopaganda“ bezeichnen. Hierbei wird die schiere Masse an minderwertigem KI-Content genutzt, um die Informationsumgebung so zu überfluten, dass die Entscheidungsfindung von Gruppen gezielt beeinträchtigt wird. Ein Beispiel hierfür sind die rumänischen Wahlen 2024/25, bei denen Kandidaten strategisch KI-generierte Memes und grobe Visualisierungen einsetzten. Diese Inhalte dienten nicht der Täuschung durch Realismus (wie bei Deepfakes), sondern nutzten eine Ästhetik qualitativ minderwertiger digitaler Inhalte, um Nahbarkeit zu simulieren und nationalistische Narrative unter Umgehung traditioneller Medienfilter viral zu verbreiten.

Die massenhafte Verbreitung von AI Slop führt zu einer fundamentalen „Datenverschmutzung“. Ansari (2025) warnt vor einem sich selbst verstärkenden Feedback-Loop: Wenn generative Systeme zunehmend auf ihren eigenen synthetischen Output trainiert werden, führt dies zu einer Homogenisierung und semantischen Degradierung der Informationsqualität. Für die Informationsökonomie bedeutet dies eine massive Erhöhung der Verifikationskosten. Die Unterscheidung zwischen authentischen, menschlich kuratierten Inhalten und synthetischen Niedrigqualitätsinhalten wird sowohl für Nutzer als auch für algorithmische Filter zunehmend ressourcenintensiv.

Van Rooij (2025) argumentiert, dass diese Entwicklung zu einer nachhaltigen Beeinträchtigung der Informationsumgebung und -qualität führt, welche die Funktionsfähigkeit von Wissensinfrastrukturen gefährdet. Wenn Suchergebnisse, Online-Artikel bis hin zu wissenschaftlichen Publikationen zunehmend mit plausibel klingenden, aber faktisch halluzinierten KI-Texten durchsetzt sind, erodiert das Vertrauen in die Zuverlässigkeit von Information generell. Dies begünstigt eine Umgebung der epistemischen Unsicherheit, in der nicht mehr die Zensur von Informationen das Hauptproblem darstellt, sondern die Unmöglichkeit, in einer Flut von synthetischem Rauschen relevante Signale zu identifizieren. Die Gefahr von AI Slop liegt somit weniger in der perfekten Täuschung des

Einzelnen, sondern in der systemischen Zersetzung der Vertrauensbasis, auf der öffentliche Diskurse beruhen.

3.3.2 Verbreitung: Automatisierung, Koordination und Cross-Plattform-Dynamiken

Jenseits der reinen Inhaltsproduktion hat sich die technologische Infrastruktur der Desinformation signifikant weiterentwickelt. Die aktuelle Forschungsliteratur verdeutlicht, dass die operativen Taktiken des CIB mittlerweile weit über die bloße quantitative Masse automatisierter Text-Bots hinausgehen. Stattdessen sind sie zunehmend durch komplexe, multimodale Strategien sowie eine funktionale Arbeitsteilung innerhalb der Verbreitungsnetzwerke geprägt.

Das Ökosystem der Manipulation hat sich zu einem zweiseitigen Markt entwickelt, in dem „Disinformation-as-a-Service“ (DaaS) als Geschäftsmodell etabliert ist. Soliman und Rinta-Kahila (2024) analysieren in ihrer Untersuchung von Crowdsourcing-Plattformen, wie Organisatoren die Interaktion zwischen Auftraggebern (Requestern) und Ausführenden (Crowdworkern) strukturieren, um Manipulationsdienstleistungen skalierbar anzubieten. Das Spektrum reicht von gefälschten Bewertungen bis zur politischen Einflussnahme. Diese Plattformen nutzen diskursive Strategien der „Sprachbereinigung“ (Language Sanitization), um ihre Aktivitäten als legitime Marketingdienstleistungen zu rahmen und so die moralische Hemmschwelle für die Beteiligten zu senken. Ergänzend dazu warnt Ferrara (2024) vor der Integration generativer KI in diese Wertschöpfungsketten. Durch den Einsatz von Large Language Models (LLMs) können böswillige Akteure nun synthetische Identitäten erschaffen, die nicht nur Inhalte produzieren, sondern auch menschliche Interaktionsmuster täuschend echt imitieren, was die Unterscheidung zwischen organischem und anorganischem Verhalten für Detektionssysteme massiv erschwert.

Parallel dazu lässt sich eine Ausdifferenzierung der Akteursrollen beobachten. Verdolotti et al. (2025) identifizieren in ihrer Untersuchung zur Verbreitung von Fehlinformationen distinkte Verhaltensarchetypen: „Amplifier“, die Inhalte initial beschleunigen, „Super-Spreader“, die für die massive Reichweite sorgen, und koordinierte Accounts, die im Verbund agieren. Diese funktionale Spezialisierung deutet auf einen hohen Grad an Professionalisierung hin, bei dem unterschiedliche Account-Typen strategisch eingesetzt werden, um die algorithmischen Hürden der Plattformen zu überwinden.

Eine zentrale Entwicklung ist auch die Verschiebung von textbasiertem Spam hin zu komplexen audiovisuellen Koordinationsmustern, insbesondere auf videozentrierten Plattformen. Luceri et al. (2025) konnten in ihrer Analyse von TikTok-Inhalten zur US-Präsidentschaftswahl 2024 nachweisen, dass Akteure generative KI nutzen, um die Detektion zu umgehen. Anstatt identische Videos zu posten, die von Hash-Filtern leicht erkannt würden, verwenden koordinierte Netzwerke identische, KI-generierte Voiceover-

Spuren oder spezifische visuelle Templates wie Split-Screen-Formate, während der visuelle Inhalt selbst variiert. Diese „semantische Koordination“ ermöglicht eine synchronisierte Amplifikation politischer Narrative, die für herkömmliche Detektionssysteme als organische Vielfalt erscheint.

Moderne Desinformationskampagnen entfalten ihre Wirkung dabei nicht in geschlossenen Silos, sondern maximieren ihren Einfluss durch strategische „Cross-Platform-Diffusion“. Cinus et al. (2025) konnten im Kontext der US-Wahl 2024 nachweisen, dass Koordinationsnetzwerke gezielt Plattformgrenzen überschreiten, um Narrative zu verstärken. Ihre Analyse deckte auf, dass russisch-assoziierte Medien systematisch über Telegram und X (ehemals Twitter) hinweg promoted wurden, wobei eine signifikante Überlappung der Nutzerbasen als Transmissionsriemen diente. Gerard et al. (2025) identifizieren in diesem Prozess sogenannte „Bridge Users“ als entscheidende Akteure. Diese Nutzer fungieren als konsistente „Early Initiators“, die Narrative von einer Plattform (oft mit geringerer Moderation) auf eine andere transferieren und dort die Diffusion in neue Communities anstoßen. Diese Form der „Narrative Migration“ ist besonders widerstandsfähig gegenüber plattformspezifischen Moderationsmaßnahmen, da das Löschen von Inhalten auf einer Plattform den Fluss des Narrativs im Gesamtnetzwerk kaum stoppt.

Dabei nutzen Akteure gezielt die technischen und strukturellen Eigenschaften (wie Verschlüsselung oder fehlende Moderation) alternativer Dienste als Rückzugs- und Koordinationsräume. Colizzi et al. (2025) untersuchten entsprechende Koordinationsmuster auf alternativen Plattformen wie Gab, VK, Minds und dem Fediverse und fanden plattformspezifische Strategien, die von Echokammern auf Gab bis hin zu hierarchischen Distributionsmodellen auf Telegram reichen. Insbesondere Telegram hat sich aufgrund seiner minimalen Moderation und Verschlüsselungsarchitektur zu einem zentralen Hub für Informationsoperationen entwickelt. Blas et al. (2025) deckten in einer großangelegten Analyse mehrsprachiger politischer Nachrichten auf Telegram fünf orchestrierte Informationsoperationen auf, darunter eine russisch unterstützte Einflusskampagne und pro-palästinensische Amplifikationsnetzwerke, die ungestört agieren konnten. Pakina et al. (2025) zeigen ergänzend, dass verschlüsselte Plattformen wie WhatsApp und Telegram aufgrund ihrer Ende-zu-Ende-Verschlüsselung als „blinde Flecken“ der Moderation fungieren, in denen KI-getriebene Propaganda ungestört reifen und massenhaft disseminiert werden kann, bevor sie in den offenen Diskurs schwappt.

Die Verbreitung erfolgt dabei nicht immer über explizite Links. Yin et al. (2025) beschreiben das Phänomen der „impliziten Propagation“, bei der sich Themen und Frames über Plattformgrenzen hinweg verbreiten, ohne dass direkte URL-Verweise existieren, was die Nachverfolgbarkeit durch herkömmliche Tracking-Methoden verhindert. Die Schnittstelle zwischen den peripheren Netzwerken und der breiten Öffentlichkeit bilden oft strategisch platzierte Einzelakteure. Jones (2025) beschreibt die Rolle von „Disinfluencern“, mit der er routinemäßige Verbreiter von Falschinformationen bezeichnet, die aufgrund ihrer Reichweite unverhältnismäßigen Einfluss auf Trends haben. Im Kontext der Unruhen in Southport zeigte sich, wie solche Akteure als Brückenbauer fungieren: Narrative, die in

alternativen Netzwerken geschmiedet wurden, werden von diesen Influencern aufgegriffen und auf Plattformen wie X popularisiert. Dabei nutzen sie das „Informationsvakuum“ nach Krisenereignissen sowie die algorithmische Präferenz für polarisierende Inhalte aus. In diesem Fall ging es um xenophobe Desinformation, die massenhaft verbreitet wurde, bevor offizielle Stellen reagieren können.

Trotz der technologischen Raffinesse dieser Kampagnen gibt es Hinweise darauf, dass ihre tatsächliche Überzeugungskraft überschätzt werden könnte. Di Marco et al. (2025) argumentieren basierend auf Analysen von Informationskaskaden, dass koordinierte Accounts oft ineffizient platziert sind und ihre Netzwerkeinflussnahme geringer ausfällt als befürchtet, da sie häufig in isolierten Clustern operieren, statt organische Nutzer effektiv zu durchdringen. Dennoch bleibt die Detektion eine Herausforderung. Mannocci et al. (2024) weisen in ihrem umfassenden Survey darauf hin, dass die Grenze zwischen legitimer Online-Koordination (z. B. digitaler Aktivismus) und schädlichem CIB zunehmend verschwimmt, was die Entwicklung präziser, nicht-diskriminierender Erkennungsalgorithmen zu einer der dringendsten Aufgaben der aktuellen Forschung macht. Zudem betonen Zhao et al. (2025), dass sich die Propagationsstrukturen je nach Plattformarchitektur signifikant unterscheiden, weshalb Detektionsmodelle, die nur auf eine Plattform trainiert sind, im realen, fragmentierten Ökosystem an ihre Grenzen stoßen.

4 Analyse und Bewertung von aktuellen Bekämpfungsstrategien

Angesichts der zunehmenden Komplexität digitaler Verbreitungswege und der Anpassungsfähigkeit böswilliger Akteure rücken Fragen der Skalierbarkeit, der Kontextabhängigkeit und der langfristigen Robustheit von Gegenmaßnahmen in den Fokus der Forschung. Kozyreva et al. (2024) systematisieren das Feld in ihrem umfassenden Review als eine „Toolbox“ von Interventionen, die sich analytisch meist zwei Hauptkategorien zuteilen lassen: „Nudging“-Ansätze, die die Entscheidungsarchitektur verändern, um Aufmerksamkeit auf Genauigkeit zu lenken, und „Boosting“-Ansätze, die darauf abzielen, die individuellen Kompetenzen der Nutzer langfristig zu stärken. Der aktuelle Forschungsstand verdeutlicht jedoch, dass es kein Allheilmittel gibt, geschweige denn ein alleiniges; vielmehr zeigen sich deutliche Spannungsfelder zwischen Skalierbarkeit, Kontextabhängigkeit und unbeabsichtigten Nebenwirkungen.

Die Strategien zur Bekämpfung von Desinformation lassen sich zudem grob entlang eines zeitlichen und systemischen Kontinuums verorten: präventive Maßnahmen vor der Exposition, reaktive Maßnahmen während oder nach der Exposition sowie strukturelle Eingriffe in die Gestaltung digitaler Informationsumgebungen. Auf der Ebene der Nutzerresilienz erlebt dabei die Inokulationstheorie eine Renaissance. Der Ansatz des „Prebunking“ zielt darauf ab, Nutzerinnen und Nutzer präventiv gegen manipulative Techniken wie emotionale Polarisierung, falsche Dichotomien oder logische Fehlschlüsse zu „impfen“, indem typische Muster solcher Taktiken vorab transparenter gemacht und anhand von Beispielen oder Spielen erfahrbar gemacht werden. Meta-Analysen zeigen, dass psychologische Inokulation die Erkennungsrate manipulativer Techniken und die Widerstandsfähigkeit gegenüber Desinformation im Durchschnitt signifikant erhöhen kann (Roozenbeek & van der Linden, 2019; Huang et al., 2024; Lu et al., 2023). Zugleich weisen neuere Arbeiten darauf hin, dass die Effekte häufig moderat und zeitlich begrenzt sind und durch Designentscheidungen des Plattformumfelds verstärkt oder abgeschwächt werden können. Pennycook et al. (2021, 2024) konnten zeigen, dass sich die Wirkung von Prebunking signifikant steigern lässt, wenn die Kompetenzvermittlung mit situativen kombiniert wird. Solche Anreize, sogenannte „Accuracy Nudges“, erinnern Nutzer im Moment des Teilens explizit an die Wichtigkeit von Genauigkeit. Dies unterstützt die Annahme, dass Maßnahmen zur Stärkung individueller Kompetenzen nicht isoliert gedacht werden sollten, sondern in eine Entscheidungsarchitektur eingebettet sein müssen, die Aufmerksamkeit systematisch von sozialer Bestätigung hin zur inhaltlichen Richtigkeit lenkt.

Im Bereich der reaktiven Bekämpfung rücken Ansätze der „Soft Moderation“ in den Vordergrund, die Inhalte nicht löschen, sondern markieren, kontextualisieren oder in ihrer Verbreitung bremsen. Dazu zählen etwa Labels, Interstitial-Warnhinweise, Herabstufung problematischer Inhalte im Ranking oder Eingriffe in Weiterleitungsfunktionen, die die Sichtbarkeit begrenzen, ohne Beiträge vollständig zu entfernen (Douek, 2021; Botero Arcila & Griffin, 2023). Warnhinweise und Labels gelten inzwischen als Standardinstrument, werden in der Forschung aber ambivalent bewertet. Pennycook et al. (2020) beschreiben

den „Implied Truth Effect“: Wenn nur ein Teil der Falschmeldungen mit Warnhinweisen versehen wird, neigen Nutzer dazu, nicht markierte Inhalte implizit als geprüft und damit vertrauenswürdig wahrzunehmen (Pennycook et al., 2020).

Vor diesem Hintergrund erscheinen Formen struktureller Friktion als vielversprechende Ergänzung. Schädliche Inhalte sind im Durchschnitt emotionaler, negativer und moralisch aufgeladener als Qualitätsjournalismus und erzeugen dadurch „natürliches“ hohes Engagement (Carrasco-Farré, 2022; Jahn et al., 2023). Friktionsansätze setzen genau an dieser Verstärkungslogik an: Sie machen das Teilen minimal aufwendiger, um impulsive Reaktionen zu verlangsamen, ohne Inhalte selbst zu verbieten oder inhaltlich zu bewerten. Klassische Beispiele sind zusätzliche Klickschritte wie „Read before you retweet“, Zwischendialoge („Sind Sie sicher, dass Sie diesen Inhalt teilen möchten?“) oder technische Begrenzungen von Weiterleitungsfunktionen. Twitters Feldexperiment mit einem „Read-the-article-before-you-retweet“-Prompt führte dazu, dass Nutzer Artikel 40 % häufiger öffneten und manche nach dem Lesen bewusst vom Retweet absahen (Tameez, 2020; Hwang & Lee, 2025). WhatsApp meldete nach der Einführung strengerer Weiterleitungsgrenzen für „highly forwarded messages“ während der COVID-19-Pandemie einen Rückgang der Viralität solcher Nachrichten um rund 70 %, was als indirekter Hinweis auf die Wirksamkeit dieser Form von Friktion interpretiert wird (TechCrunch, 2020).

Die modellbasierten Arbeiten von Jahn et al. Zeigen zudem, dass Friktion als verhaltensökonomisches Instrument verstanden werden kann, das die „choice architecture“ in sozialen Netzwerken verändert: In einem Agenten-basierten Simulationsmodell reduziert Friktion allein zunächst vor allem die Menge der geteilten Inhalte, während eine Kombination aus leichter Friktion und lernfördernden Elementen (z. B. kurze Hinweise auf Community-Standards oder News-Evaluationsfragen) die durchschnittliche Qualität des geteilten Contents signifikant erhöhen kann (Jahn et al., 2023; Jahn et al., 2025). Aus regulatorischer Perspektive werden solche Eingriffe als Teil eines Spektrums „strategischer Friktion“ diskutiert, das über die klassische Lösch-/Nichtlösch-Logik hinausgeht und durch kleine, transparente Hürden impulsive Fehlentscheidungen dämpfen soll, ohne Meinungsfreiheit direkt zu beschneiden (Laidlaw, 2022).

Die Analysen des Integrity Institute, insbesondere das Misinformation Amplification Tracking Dashboard, ergänzen dieses Bild. Sie zeigen quantitativ, dass große Plattformen durch ihre engagementorientierten Rankingsysteme Desinformation systematisch verstärken und dadurch Anreizstrukturen schaffen, die emotionalisierende und polarisierende Inhalte belohnen (Allen, 2022). Auch wenn diese Arbeiten keine eigene Friktionsintervention evaluieren, unterstreichen sie die Notwendigkeit, die „Fahrbahn“ der Informationsverbreitung zu verändern, nicht nur einzelne „Falschfahrer“ zu markieren. Im Zusammenspiel deuten die Ergebnisse darauf hin, dass schon relativ einfache Friktionsmaßnahmen wie zusätzliche Klicks, Prompt-Dialoge oder Weiterleitungsgrenzen einen messbaren Beitrag zur Eindämmung der Verbreitung von Misinformation leisten können, sofern sie transparent gestaltet, evaluiert und möglichst mit lern- und kompetenzfördernden Komponenten kombiniert werden (Jahn et al., 2023; Jahn et al., 2025).

Ein weiterhin grundsätzlich relevanter Hebel liegt zudem in der aktiven algorithmischen Kuration zugunsten verlässlicher Quellen, häufig unter dem Schlagwort „authoritative sources“ diskutiert. Gemeint sind Eingriffe in Ranking- und Empfehlungssysteme, bei denen Plattformen nicht nur ex post problematische Inhalte löschen oder kennzeichnen, sondern ex ante die Sichtbarkeit einzelner Inhalte und Akteure strukturell verändern. In der Misinformation-Forschung wird dies als Veränderung der „choice architecture“ digitaler Informationsumgebungen beschrieben: Empfehlungsalgorithmen werden so umkonfiguriert, dass qualitativ hochwertige, evidenzbasierte Informationen prominenter ausgespielt und potenziell schädliche Inhalte systematisch herabgestuft werden (van der Linden, 2022; Shin et al., 2022; Metzler und Garcia, 2024). Besonders weit reicht die Implementierung solcher Mechanismen im Gesundheitsbereich, in dem Fehlentscheidungen unmittelbare körperliche Schäden nach sich ziehen können. Vor diesem Hintergrund hat YouTube im Zuge der COVID-19-Pandemie gemeinsam mit der US National Academy of Medicine (NAM) das Konzept „authoritative health sources“ entwickelt. Eine von NAM eingesetzte Gruppe erarbeitete Prinzipien und Attribute, anhand derer Plattformen glaubwürdige Gesundheitsanbieter identifizieren und in der Ausspielung priorisieren können. Als Beispiele werden öffentliche Gesundheitsbehörden, akademische Kliniken oder anerkannte Fachgesellschaften genannt (Kington et al., 2021). Parallel dazu formulierte die NAM-Initiative ethische und Public-Health-bezogene Kriterien für großskalige Content-Kennzeichnung, etwa Transparenzanforderungen, die Vermeidung von Interessenkonflikten sowie die Notwendigkeit, die Wirksamkeit solcher Maßnahmen datenbasiert zu evaluieren (Burstin et al., 2023).

Konkret äußert sich diese aktive Kuration auf Videoplattformen in der Einführung von Informationsbereichen und Kontextboxen, die Suchergebnisse zu sensiblen Themen teilweise überlagern. Ein Beispiel ist die Implementierung von Informationsmodulen zu Erster Hilfe und anderen akuten Gesundheitsthemen. Bei entsprechenden Suchanfragen werden kuratierte Inhalte verifizierter Gesundheitsorganisationen angezeigt. Dies geschieht abweichend von der üblichen, auf Engagement optimierten Sortierung, etwa durch die Einbindung nationaler Rettungsdienste oder Gesundheitsbehörden (Graham, 2025). Diese Praxis fügt sich in eine breitere Plattformstrategie ein, die YouTube selbst als „raising authoritative sources“ beschreibt: Für Nachrichten, Politik und medizinische Themen werden Inhalte etablierter Nachrichtenredaktionen und Gesundheitsinstitutionen algorithmisch gegenüber rein engagement-getriebenen Angeboten bevorzugt, während in Unterhaltungskategorien weiterhin Popularitätsmetriken dominieren. Aktive algorithmische Kuration beschränkt sich jedoch nicht auf einzelne Plattformen. Eine vergleichende Studie zu COVID-19-Suchergebnissen in mehreren Sprachen zeigt, dass Suchmaschinen in unterschiedlichem Ausmaß offizielle Regierungs- und Gesundheitswebseiten priorisieren; insgesamt rahmen sie die Pandemie aber deutlich stärker über institutionelle Quellen als über alternative Medien (Rovetta und Bhagavathula, 2020). Übersichtsarbeiten zur „health misinformation infodemic“ plädieren vor diesem Hintergrund dafür, technische Lösungsansätze insbesondere auf die algorithmischen Komponenten sozialer Medien auszurichten. Im Mittelpunkt stehen damit Ranking-, Empfehlungs- und

Verstärkungslogiken, statt ausschließlich auf individuelle Medienkompetenz oder klassische Faktenchecks zu setzen (Rodrigues et al., 2024).

In der Interventionsforschung werden solche Ranking-Eingriffe ambivalent bewertet. Einerseits sprechen Evidenzübersichten dazu, wie sich Fehl- und Desinformation eindämmen lässt, dafür, die digitale „choice architecture“ so zu gestalten, dass hochwertige Quellen wahrscheinlicher zuerst gesehen werden und problematische Inhalte weniger prominent erscheinen (van der Linden et al., 2022). Andererseits warnen sozial- und rechtswissenschaftliche Analysen vor neuen Formen der Machtkonzentration: Wenn wenige private Plattformen definieren, wer als „authoritative source“ gilt, drohen neue Gatekeeping-Strukturen, potenzielle politische Schlagseite und Vertrauensverluste bei Teilen der Öffentlichkeit (Clemons et al., 2025; Shin, 2022). Diskursanalysen zu Plattforminterventionen gegen „Fake News“ zeigen, dass Unternehmen ihre Maßnahmen häufig mit Verweis auf Gesundheits- und Demokratieschutz legitimieren, während zivilgesellschaftliche Akteure Risiken für Meinungsfreiheit, Pluralität und die Sichtbarkeit marginalisierter Stimmen betonen. Auf regulatorischer Ebene verlangt die EU-Verordnung über digitale Dienste (DSA) von sehr großen Online-Plattformen, systemische Risiken zu identifizieren. Dazu zählt insbesondere die Verbreitung von Desinformation. Zudem sind Empfehlungssysteme so anzupassen, dass diese Risiken gemindert werden. Vorschläge für Audits von Empfehlungssystemen betonen, dass Eingriffe in Ranking-Logiken in einem mehrstufigen, szenariobasierten Prozess geprüft, dokumentiert und im Zeitverlauf evaluiert werden sollten (Meßmer & Degeling, 2023). Insgesamt deutet die Evidenz darauf hin, dass aktive algorithmische Kuration ein wichtiger, aber allein nicht ausreichender Baustein eines umfassenden Maßnahmenbündels ist, der mit Nutzerkompetenz, Friktion und partizipativen Korrekturmechanismen kombiniert werden sollte.

4.1 Einsatz und Wirksamkeit aktueller Ansätze

Eine der am intensivsten untersuchten Maßnahmen sind sogenannte „Accuracy Nudges“. Dabei handelt es sich um subtile Hinweise, die Nutzer im Moment der Interaktion an das Konzept der Wahrheit erinnern, ohne Inhalte zu zensieren. Dieser Ansatz basiert auf der Beobachtung, dass Nutzer oft nicht aus böser Absicht, sondern aus Unachtsamkeit Falschinformationen teilen, weil ihre Aufmerksamkeit auf soziale Validierung (in Form von Likes etc.) statt auf Genauigkeit gerichtet ist. Lin et al. (2024) lieferten mit einer großangelegten Feldstudie auf Facebook und Instagram (N > 33 Millionen) den ersten Beweis für die Skalierbarkeit dieses Ansatzes. Ihre Ergebnisse zeigen, dass inhaltsneutrale Werbeanzeigen, die Nutzer lediglich auffordern, über Genauigkeit nachzudenken („Think about accuracy“), das Teilen von Fehlinformationen um etwa 2,6 % reduzierten. Obwohl dieser Effekt auf individueller Ebene klein erscheint, argumentieren die Autoren, dass er angesichts der enormen Nutzerzahlen der Plattformen eine signifikante Reduktion der Gesamtexposition bewirken kann, ohne die Meinungsfreiheit einzuschränken.

Fazio et al. (2025) zeigen in einer aktuellen Vergleichsstudie verschiedener Interventionen, dass solche Nudges eine kosteneffiziente Methode darstellen, um die Qualität des geteilten Inhalts zu verbessern. Allerdings weisen sie auch auf Heterogenität in der Wirkung hin: Nicht alle demografischen Gruppen sprechen gleichermaßen auf diese Impulse an; insbesondere bei stark polarisierten Themen kann der Effekt verpuffen. Kritiker merken zudem an, dass Nudges primär das Teilungsverhalten beeinflussen, aber nicht zwingend die zugrundeliegenden Überzeugungen oder das langfristige Wissen der Nutzer verändern. Herzog und Hertwig (2025) betonen daher, dass Nudges zwar als kurzfristige erste Hilfe dienen, aber nicht die Notwendigkeit tiefergehender Kompetenzförderung ersetzen können.

Im Bereich der reaktiven Bekämpfung von Desinformation hat sich in den letzten Jahren eine signifikante Verschiebung von zentraler, expertenbasierter Moderation hin zu dezentralen, nutzerbasierten Ansätzen vollzogen. Besonders prominent ist das Modell der „Community Notes“, das ursprünglich von Twitter (jetzt X) als „Birdwatch“ eingeführt wurde. Diese Entwicklung beschränkt sich jedoch nicht auf X; auch Meta (Facebook, Instagram) hat Anfang 2025 angekündigt, sein Fact-Checking-Programm zunächst in den USA zugunsten eines ähnlichen, Community-basierten Systems umzustrukturieren. Diese strategische Neuausrichtung, oft begründet mit Skalierbarkeit und dem Vorwurf der Voreingenommenheit traditioneller Faktenprüfer, wird in der aktuellen Forschung kontrovers diskutiert.

Die empirische Evidenz zur Wirksamkeit dieses Ansatzes ist mittlerweile robust und zeigt durchaus signifikante Einfallswinkel in die Diffusionsdynamik von Desinformation. Slaughter et al. (2025) konnten in einer umfassenden Kausalanalyse von über 40.000 Beiträgen nachweisen, dass das erfolgreiche Anhängen einer Community Note die Verbreitung von Falschinformationen stark eindämmt. Ihre Daten zeigen, dass Beiträge nach der Veröffentlichung einer Note im Durchschnitt 46,1 % weniger Reposts, 44,1 % weniger Likes und 21,9 % weniger Antworten erhielten. Ein entscheidender Faktor ist dabei die Zeit: Notes, die frühzeitig im Lebenszyklus eines viralen Beitrags erschienen, brachen die exponentielle Verbreitungskurve am effektivsten. Diese Resultate werden durch Chuai et al. (2024a) gestützt, die in einer differenzierten Längsschnittanalyse bestätigen, dass Community Notes die Wahrscheinlichkeit, dass ein irreführender Beitrag vom Ersteller selbst gelöscht wird, um 103,4 % erhöhen und die Weiterverbreitung um durchschnittlich 62 % reduzieren.

Dieser Mechanismus wirkt nicht nur auf der Ebene der Metriken, sondern auch auf der Ebene der Nutzerpsychologie. Drolsbach et al. (2024) fanden in experimentellen Studien heraus, dass Nutzer Community Notes als signifikant vertrauenswürdiger einstufen als klassische, von der Plattform oder Experten gesetzte Warnhinweise („Flags“). Der entscheidende Wirkfaktor ist hierbei der erklärende Kontext: Da Community Notes nicht nur warnen, sondern begründen, warum eine Information irreführend ist, werden sie über ideologische Lager hinweg als legitimeres Korrektiv wahrgenommen. Dies deutet darauf

hin, dass der partizipative Charakter des Systems die Akzeptanz von Faktenkorrekturen in polarisierten Umgebungen erhöhen kann.

Trotz dieser Erfolge leidet das Modell unter strukturellen Defiziten, die seine Eignung als alleiniges Moderationsinstrument in Frage stellen. Das Kernproblem liegt im Algorithmus des „Bridging-based Ranking“. Damit eine Note öffentlich sichtbar wird, muss sie nicht nur von vielen Nutzern als „hilfreich“ bewertet werden, sondern von Nutzern, die in der Vergangenheit unterschiedliches Abstimmungsverhalten zeigten (also ideologisch divers sind). De et al. (2024) weisen in ihrer Analyse nach, dass dieser Zwang zum lagerübergreifenden Konsens zu einem massiven „Under-Flagging“ führt: Für 91 % der Beiträge, zu denen eine Note vorgeschlagen wurde, konnte niemals eine Note veröffentlicht werden, da die erforderliche Einigkeit zwischen den politischen Lagern ausblieb. Dies erzeugt ein signifikantes „Knowledge Gap“, bei dem gerade die kontroversesten und gesellschaftlich relevantesten Desinformationen ungekennzeichnet bleiben, während unpolitische oder triviale Falschmeldungen (z. B. über Konsumprodukte) schnell korrigiert werden.

Chuai et al. (2024b) bestätigen dieses Skalierungsproblem. Ihre Daten zeigen, dass zwar das Volumen der von Nutzern geschriebenen Notes stetig wächst, die Rate der tatsächlich angezeigten Notes jedoch stagniert oder relativ zum Desinformationsvolumen sogar sinkt. Wirtschafter und Majumder (2023) warnten bereits frühzeitig davor, dass solche Systeme anfällig für „Brigading“ sind. Dabei koordinieren politische Gruppen Aktionen, in denen Notes strategisch abgewertet werden („Downvoting“), um deren Veröffentlichung zu verhindern. Diese Vulnerabilität macht das System manipulierbar durch gut organisierte ideologische Akteure, die den Konsensmechanismus als Veto-Instrument missbrauchen.

Ein weiterer kritischer Aspekt betrifft die politische Neutralität und die Definition von Schaden. Renault et al. (2025) untersuchten die Verteilung von Community Notes auf X und deckten eine deutliche politische Asymmetrie auf: Beiträge von Republikanern wurden 2,3-mal häufiger als irreführend markiert als Beiträge von Demokraten. Die Autoren konnten jedoch nachweisen, dass dies nicht auf einen Bias der Bewertenden zurückzuführen ist, sondern darauf, dass republikanische Akteure tatsächlich signifikant häufiger irreführende Informationen teilten. Dies führt zu dem Dilemma, dass ein technokratisch neutrales System politisch ungleiche Ergebnisse produziert, was wiederum den (unberechtigten) Vorwurf der Zensur durch konservative Kräfte befeuern kann.

Für den deutschen Kontext zeigt Nenno (2025) eine andere Facette dieser Asymmetrie: Hier werden Beiträge bestimmter Parteien, insbesondere der Grünen, überproportional häufig mit Notes versehen oder angegriffen. Folglich zeigt sich eine Diskrepanz zwischen der Vorschlagsfunktion (die oft als Instrument für politisches Framing genutzt wird) und der Veröffentlichungsfunktion (die durch den Konsensalgorithmus oft blockiert ist). Potenziell problematischer ist außerdem die von Matamoros-Fernández und Jude (2025) identifizierte „Harm-Blindness“ (Schadensblindheit) des Systems. Da Community Notes

primär auf faktische Genauigkeit („Accuracy“) fokussiert sind, versagen sie bei Inhalten wie dog-whistling, Belästigung oder hate speech, die zwar faktisch nicht eindeutig widerlegbar, aber dennoch in hohem Maße schädlich sein können.

Vor dem Hintergrund dieser Defizite plädieren Experten für einen hybriden Ansatz. Das klassische „Third-Party Fact-Checking“ (3PFC) hatte sich in Teilen durchaus als wirksam erwiesen. Eigene Daten der Firma Meta zeigten, dass 95 % der Nutzer Inhalte nicht anklicken, wenn diese mit einem professionellen Warnhinweis versehen sind (vgl. SITC, 2025). Ein integriertes Modell, bei dem akkreditierte Experten Hochrisiko-Desinformation (z. B. öffentliche Gesundheit, Sicherheit) prüfen und Crowdsourcing für den „Long Tail“ weniger brisanter Falschmeldungen genutzt wird, erscheint daher als vielversprechendster Weg.

Um die Effizienzblockade des Konsensmechanismus zu überwinden, schlagen De et al. (2024) auch den Einsatz von generativer KI vor. Ihr Konzept der „Supernotes“ nutzt Large Language Models (LLMs), um die Argumente verschiedener, konkurrierender Notiz-Entwürfe zu synthetisieren und in eine neutrale, konsensfähige Sprache zu übersetzen. Experimente zeigen, dass solche KI-aggregierten Notes eine deutlich höhere Wahrscheinlichkeit haben, von diversen Nutzergruppen akzeptiert zu werden, und somit die Veröffentlichungsrate steigern könnten. Gleichzeitig weist die Forschung auf die sozialpsychologischen Nebenwirkungen hin. Chuai et al. (2025) kamen zu dem Ergebnis, dass das Erscheinen einer Community Note nicht nur kognitive Korrekturen bewirkt, sondern auch starke negative Emotionen („Moral Outrage“) in den Antworten auf den ursprünglichen Post auslöst. Die Note fungiert als Signal für eine Normverletzung, was dazu führt, dass der Absender der Desinformation sozial sanktioniert wird (Anstieg von Wut und Ekel in den Kommentaren). Dies unterstreicht, dass Community Notes nicht nur informative, sondern auch normative Werkzeuge sind, die das soziale Klima auf einer Plattform prägen.

Allgemein bleibt auch bei Community Notes und anderen crowd-basierten Ansätzen das Problem der „sozialen Korrektur“ bestehen: King et al. (2025) zeigen, dass Menschen zwar erwarten, dass andere gegen Desinformation intervenieren, selbst jedoch oft zögern, Korrekturen vorzunehmen. Eine Ausnahme liegt vor, wenn der Absender der Falschinformation ihnen persönlich nahe steht. Diese Diskrepanz zwischen normativer Erwartung und tatsächlichem Handeln limitiert das Potenzial rein nutzerbasierter Korrekturmechanismen und macht deutlich, dass Crowdsourcing allein professionelle Moderation nicht vollständig ersetzen kann.

Als nachhaltigere Alternative zu Nudges und Moderation wird das Konzept des „Boosting“ diskutiert, das darauf abzielt, die Medienkompetenz der Nutzer zu steigern. Ein prominenter Ansatz ist die psychologische Inokulation („Prebunking“), bei der Nutzer präventiv gegen manipulative Techniken (wie Emotionalisierung, falsche Dichotomien) „geimpft“ werden. Der APA Consensus Statement (van der Linden et al., 2025) fasst die umfassende Laborevidenz zusammen, die bestätigt, dass Prebunking die Widerstandsfähigkeit

gegen Desinformation erhöhen kann. Videos oder Spiele, die manipulative Taktiken entlarven, helfen Nutzern, diese Muster später in realen Inhalten wiederzuerkennen.

Allerdings stößt die theoretische Wirksamkeit von Inokulationsstrategien bei der Übertragung in die ökologische Realität an deutliche Grenzen. So deckten Wang et al. (2025) eine fundamentale Diskrepanz zwischen kognitiver Erkennung und tatsächlichem Handeln auf. In Experimenten mit simulierten Social-Media-Feeds zeigte sich, dass Nutzer nach einer Inokulation zwar die manipulative Natur emotionaler Inhalte besser identifizierten und deren Genauigkeit kritischer bewerteten, dies jedoch kaum Auswirkungen auf ihr faktisches Interaktionsverhalten hatte. Die Nutzer verbrachten trotz des Wissens um die Manipulation weiterhin Zeit mit den Inhalten und reagierten auf diese. Diese Diskrepanz unterstreicht die Grenzen rein kognitiver Interventionen in der „Aufmerksamkeitsökonomie“: Da emotionale Reize wie Wut oder Angst tiefgreifende, impulsive Treiber für Aufmerksamkeit darstellen, reicht das reine Wissen um Manipulationsversuche oft nicht aus, um die reflexhafte Interaktion oder das Verweilen im Feed zu unterbrechen.

Über diese Verhaltenslücke hinaus identifizierten Martini et al. (2025) in einer großangelegten Feldstudie mit über 2.000 Schülern einen potenziell kontraproduktiven kognitiven Effekt: Statt die spezifische Unterscheidungsfähigkeit zu schärfen, führte die Intervention im realen Klassenzimmerumfeld primär zu einer Zunahme generalisierter Skepsis. Dies hatte zur Folge, dass die Schüler paradoxerweise auch das Vertrauen in valide wissenschaftliche Quellen verloren. Dies deutet darauf hin, dass Prebunking, wenn es nicht sorgfältig kalibriert ist, das Vertrauen in alle Informationen untergraben kann, anstatt die Unterscheidungsfähigkeit zu schärfen.

Dieser Zusammenhang entspricht einer wachsenden Zahl an Studien, die vor den Kollateralschäden gut gemeinter Interventionen warnen. Hoes et al. (2024) demonstrierten in Experimenten in den USA, Polen und Hongkong, dass gängige Maßnahmen wie Fact-Checking-Labels und Medienkompetenz-Tipps zwar effektiv die Akzeptanz von Falschinformationen senken, aber gleichzeitig das Vertrauen in korrekte Informationen und demokratische Institutionen beschädigen. Dieses „Skeptizismus-Paradoxon“ deutet darauf hin, dass Interventionen, die primär auf Warnung und Misstrauen setzen, die epistemische Unsicherheit der Bürger erhöhen können, anstatt sie zu verringern. Wenn Nutzer lernen, allem zu misstrauen, verlieren auch faktenbasierte Nachrichten ihre Überzeugungskraft.

Besondere Vorsicht ist bei vulnerablen Gruppen wie Jugendlichen geboten. Ma et al. (2025) argumentieren aus einer entwicklungswissenschaftlichen Perspektive, dass Heranwachsende aufgrund ihrer sozialen Orientierung und sich entwickelnden kognitiven Fähigkeiten anders auf Desinformation reagieren als Erwachsene. Jugendliche sind besonders empfänglich für sozialen Druck und emotionale Belohnungen, was sie einerseits anfälliger macht, andererseits aber auch Chancen für „Social Norms“-Interventionen bietet. Interventionen müssen daher altersgerecht gestaltet sein, um nicht durch Reaktanz oder kognitive Überforderung kontraproduktiv zu wirken.

Eine Weiterentwicklung des klassischen Inokulationsansatzes zielt darauf ab, nicht nur das Erkennen spezifischer Manipulationstechniken zu trainieren, sondern grundlegende kognitive Dispositionen der Nutzer zu stärken. Biddlestone et al. (2025) untersuchten in diesem Kontext die Rolle des „Actively Open-Minded Thinking“ (AOT). Gemeint ist hierbei eine Denkweise, die durch das aktive Hinterfragen eigener Überzeugungen und die Vermeidung von Bestätigungsfehlern gekennzeichnet ist. Ihre Ergebnisse zeigen, dass Prebunking-Interventionen besonders dann wirksam sind, wenn sie „norm-basiert“ erweitert werden: Indem den Nutzern vermittelt wird, dass offenes und kritisches Denken eine sozial erwünschte Norm darstellt, erhöht sich deren Motivation, Informationen akkurat zu verarbeiten. Diese „norm-enhanced“ Inokulation führte in den Experimenten indirekt zu einer verbesserten Unterscheidungsfähigkeit gegenüber Desinformation und reduzierte signifikant den Glauben an Verschwörungsnarrative. Dies impliziert, dass erfolgreiche Gegenmaßnahmen nicht rein technokratisch ansetzen dürfen, sondern die epistemischen Normen und das soziale Selbstverständnis der Nutzer adressieren müssen.

Die aktuelle Evidenzlage legt nahe, dass keine Einzelmaßnahme ausreicht, um das komplexe Problem der Desinformation zu lösen. Während algorithmische Nudges (Lin et al., 2024) und Community Notes (Slaughter et al., 2025) die virale Verbreitung dämpfen können, adressieren sie nicht die Wurzeln der Anfälligkeit. Kompetenzbasierte Ansätze (Herzog & Hertwig, 2025) sind nachhaltiger, bergen aber das Risiko, pauschales Misstrauen zu fördern, wenn sie nicht präzise auf die Zielgruppen abgestimmt sind (Hoes et al., 2024; Martini et al., 2025). Zukünftige Strategien müssen daher einen hybriden Ansatz verfolgen: Sie müssen technische Friktion zur Verlangsamung der Verbreitung mit pädagogischem Empowerment verbinden und dabei sorgfältig die Balance zwischen der Abwehr von Falschinformationen und dem Erhalt des Vertrauens in verlässliche Quellen wahren. Nur so kann verhindert werden, dass Maßnahmen zur Eindämmung von Falschinformationen unbeabsichtigt das Vertrauen in verifizierte Informationen beeinträchtigen.

Während Maßnahmen wie Nudging und Community Notes primär auf der Ebene der individuellen Nutzerinteraktion ansetzen und auf Verhaltensänderung zielen, erfordert die Bewältigung der systemischen Risiken und Machtasymmetrien einen verbindlichen rechtlichen Rahmen. Dieser wird im folgenden Kapitel im internationalen Vergleich beleuchtet.

4.2 Der Regulierungsrahmen im internationalen Vergleich

4.2.1 Aktuelle Entwicklungen EU

4.2.1.1 DSA & Code of Conduct on Disinformation

Der DSA verkörpert diesen prozessorientierten, systemischen Ansatz in rechtlich verbindlicher Form. Strowel und De Meyere (2023) zeigen, dass der DSA die Bekämpfung von Desinformation nicht über pauschale Verbote, sondern über abgestufte

Sorgfaltspflichten („due diligence obligations“) organisiert. Besonders für sogenannte Very Large Online Platforms (VLOPs) und Very Large Online Search Engines (VLOSEs) etablieren die Artikel 34 und 35 des DSA ein allgemeines Risikomanagement-System, das von Plattformen verlangt, systemische Risiken für Grundrechte, demokratische Prozesse und öffentliche Sicherheit zu identifizieren, zu bewerten und zu mindern. Dazu zählen ausdrücklich auch Risiken im Zusammenhang mit Desinformation und manipulativen Einflusskampagnen (Husovec, 2024; Eder, 2024).

Ó Fathaigh et al. (2025) arbeiten heraus, dass der DSA ein bewusst indirektes Regulierungsmodell wählt: Statt „Desinformation“ als eigenständige Kategorie mit Löschpflichten zu definieren, verpflichtet er Plattformen, Verfahren und Organisationsstrukturen so auszugestalten, dass die Verbreitung schädlicher, aber vielfach legaler Inhalte durch risikobasierte Maßnahmen eingedämmt wird. In der Literatur wird dies als „lawful but awful“-Desinformation bezeichnet. Dazu gehören Transparenzanforderungen hinsichtlich Empfehlungs-Algorithmen, Zugang zu Daten für Forschende, Melde- und Beschwerdesysteme sowie die Zusammenarbeit mit „Trusted Flaggers“, die als vertrauenswürdige Hinweisgeber qualifiziert werden (Strowel & De Meyere, 2023; Van de Kerkhof, 2025). Eder (2024) interpretiert dieses System als Versuch, einen positiven Rückkopplungsmechanismus zu etablieren, in dem iterative Risikobewertungen, unabhängige Audits und Multi-Stakeholder-Beteiligung schrittweise zu verfeinerten Standards im Umgang mit systemischen Risiken führen.

Die UNESCO-Leitlinien zur Governance digitaler Plattformen stützen diesen Ansatz normativ. Sie empfehlen, Regulierung vor allem auf Strukturen, Prozesse und Machtasymmetrien im digitalen Kommunikationsraum zu richten, statt Behörden zu ermächtigen, Wahrheitsgehalte einzelner Inhalte zu definieren (UNESCO, 2023). Kernprinzipien sind Transparenz, Rechenschaftspflicht, Nutzer-Empowerment und der Schutz journalistischer und wissenschaftlicher Arbeit. Die Leitlinien warnen ausdrücklich vor gesetzlichen „Fake-News“-Verboten, deren Unbestimmtheit zu Überblockierungen, willkürlicher Durchsetzung und eingeschränkten öffentlichen Debatten führen kann.

Dass der DSA dieses Modell nicht nur auf dem Papier verfolgt, zeigen die ersten Durchsetzungsentscheidungen. Die Europäische Kommission eröffnete seit 2023 mehrere formelle Verfahren gegen große Plattformen wie TikTok, Meta und X. Im Fokus stehen systemische Risiken für Wahlprozesse, Jugendschutz und der Umgang mit schädlichen Inhalten (Europäische Kommission, 2024a; Husovec, 2024). Am 5. Dezember 2025 verhängte die Kommission erstmals ein Bußgeld in Höhe von 120 Millionen Euro gegen X, weil das Unternehmen seine Transparenzpflichten missachtet, mit irreführenden Designentscheidungen („Deceptive Design“) gearbeitet und Forschenden keinen ausreichenden Zugang zu öffentlichen Daten gewährt hatte (Europäische Kommission, 2025). Die Sanktion setzt nicht bei einzelnen strittigen Inhalten an, sondern bei strukturellen Defiziten im Risikomanagement und in der Offenlegungspraxis. Damit folgt sie dem Charakter prozessbasierter Regulierung.

Die Landschaft der Plattform-Regulierung befindet sich in einer Phase fundamentaler strategischer Spannungen. Während der europäische Gesetzgeber mit dem Digital Services Act (DSA) und der Transformation des Code of Practice on Disinformation in einen verbindlichen Verhaltenskodex (Code of Conduct) den Druck zur Verantwortungsübernahme erhöht, vollziehen die großen Plattformbetreiber (Very Large Online Platforms, VLOPs) eine gegenläufige Bewegung des taktischen Rückzugs. Diese Divergenz zwischen regulatorischem Anspruch und unternehmerischer Praxis manifestiert sich nicht nur in der Erosion freiwilliger Selbstverpflichtungen, sondern auch in einer tiefgreifenden Umstrukturierung der Moderationsparadigmen, die zunehmend von Kosteneffizienz und politischer Risikominimierung getrieben sind.

Der EU Code of Practice on Disinformation, der als ko-regulatives Brückeninstrument zur Vorbereitung auf die DSA-Compliance konzipiert war, verlor in der operativen Realität an Bindungskraft. Eine detaillierte quantitative Analyse von Democracy Reporting International (DRI) belegt, dass die großen Plattformen ihre Verpflichtungen im Übergangszeitraum von 2022 bis 2025 signifikant reduziert haben. Im Durchschnitt strichen die Unterzeichner rund 31 % ihrer ursprünglichen Zusagen ersatzlos aus den Reportings (Alvarado Rincón & Meyer-Resende, 2025). Besonders gravierend wirkt sich dieser Rückzug auf das Ökosystem der externen Validierung aus: Die Unterstützung für die unabhängige Fact-Checking-Community sank um 64 %. Dieser Abbau von Ressourcen steht in einem diametralen Widerspruch zur wachsenden Notwendigkeit verifizierter Informationen in Jahren wichtiger internationaler Wahlen und deutet darauf hin, dass Plattformen die Kooperation mit externen Prüfinstanzen zunehmend als geschäftsschädigenden Kostenfaktor oder politisches Risiko betrachten.

Auch qualitativ offenbart die Umsetzung massive Defizite. Die Evaluation des European Digital Media Observatory (EDMO) für das erste Halbjahr 2024 charakterisiert die Compliance-Berichte der VLOPs als inkonsistent und analytisch oberflächlich. Botan und Meyer (2025) kritisieren darin insbesondere das Fehlen granularer Daten auf Ebene der einzelnen Mitgliedstaaten. Plattformen wie Meta (Facebook/Instagram) und TikTok lieferten keine hinreichend disaggregierten Daten zur Verbreitung von Desinformation in spezifischen Sprachräumen, wodurch strukturelle Risiken in kleineren EU-Staaten statistisch unsichtbar bleiben und eine Überprüfung der Effektivität von Gegenmaßnahmen verunmöglicht wird. Diese Datenlücken sind nicht als bloße Nachlässigkeit zu interpretieren, sondern als strategisch mangelnde Transparenz, die eine unabhängige Auditierung der algorithmischen Risiken, wie sie der DSA in Artikel 34 fordert, systematisch erschwert. Auch das German-Austrian Digital Media Observatory (GADMO) konstatiert in seinem Jahresbericht 2024, dass trotz formaler Bekenntnisse zahlreiche Fälle von unzulässiger politischer Werbung und ungekennzeichneten Deepfakes auf TikTok und X dokumentiert wurden. Dies deutet auf eine Diskrepanz zwischen formaler Regelbefolgung und der tatsächlichen Nutzbarkeit der Beschwerdemechanismen hin (Wegner, 2024).

Parallel zur Erosion der Selbstverpflichtung ist eine strategische Verschiebung der Moderationsmethoden zu beobachten, die sich vom Ideal der kuratierten Sicherheit („Safety

by Design“) entfernt. Insbesondere unter dem Eindruck des politischen Drucks in den USA, wo konservative Kräfte Moderation zunehmend als Zensur framen, adoptieren Plattformen einen „Hands-off Approach“. Meta hat beispielsweise signalisiert, die Zusammenarbeit mit professionellen Faktenprüfern in den USA zugunsten von nutzerbasierten Modellen wie „Community Notes“ zu reduzieren. Aus Expertensicht könnte diese Entwicklung auch die Integritätspolitik in Europa beeinflussen (z.B. Windwehr, 2025). Dieser Trend zur „Demokratisierung der Wahrheit“ durch Crowdsourcing, wie er auf der Plattform X (ehemals Twitter) als primäres Korrektiv etabliert wurde, birgt jedoch signifikante Risiken, wie bereits in Kapitel 4.1 detailliert dargelegt.

Ein weiteres Feld nicht hinreichend wirksamer Regulierung ist der Umgang mit politischer Werbung. Trotz offizieller Verbote oder strenger Transparenzregeln auf Plattformen wie TikTok und Meta dokumentiert die Forschung massive Umgehungsstrategien. Akteure nutzen systematisch „Grauzonen“, indem sie politische Botschaften über Influencer („Fin-Influencer“ für Politik) oder als organische „News“-Beiträge tarnen, die von den automatisierten Werbefiltern nicht erfasst werden. Das Institute for Strategic Dialogue (ISD, 2024) konnte zeigen, dass Meta im Vorfeld von Wahlen tausende Anzeigen zuließ, die explizite Wahllügen verbreiteten, weil die technischen Detektionssysteme durch einfache Verschleierungstaktiken ausgehebelt wurden. Dies verdeutlicht, dass die Selbstregulierung der Werbemarktplätze ohne externe Auditierung ineffektiv bleibt.

Die Auswirkungen dieser Governance-Defizite variieren stark je nach technischer Architektur der Dienste. TikTok steht hierbei aufgrund seiner asymmetrischen Algorithmen besonders im Fokus. Ibrahim et al. (2025) wiesen in einem umfangreichen Audit zur US-Wahl 2024 nach, dass die Empfehlungssysteme der Plattform nicht neutral agieren, sondern republikanische und polarisierende Inhalte systematisch bevorzugten. Neutrale Nutzerkonten erhielten im Experiment 11,5 % mehr rechtsgerichtete Inhalte als linksgerichtete, was eine aktive, algorithmische Verzerrung des Diskurses belegt. Verstärkt wird dieser Effekt durch die spezifische Demografie der Plattform: Tjaden et al. (2025) argumentieren, dass die Nutzerschaft von TikTok eine geringere Skepsis gegenüber Falschinformationen aufweist, was die Plattform zu einem idealen Vektor für „Short-Form“-Propaganda macht.

Bei Instagram hingegen liegt das Problem weniger in der offensichtlichen algorithmischen Verzerrung als vielmehr in der Diskrepanz zwischen formaler DSA-Compliance und praktischer Nutzbarkeit. Sekwenz et al. (2025) analysierten die Melde- und Beschwerdewege der Plattform und fanden, dass diese oft hinter „Dark Patterns“ verborgen sind. Während Instagram theoretisch die gesetzlich geforderten Einspruchsmöglichkeiten bietet, sind diese in der UX-Gestaltung so tief in verschachtelten Menüs versteckt, dass sie faktisch ins Leere laufen. Dies führt zu einer „Compliance-Fassade“, bei der regulatorische Vorgaben technisch erfüllt, aber in der Nutzungspraxis unterlaufen werden.

Im sensiblen Bereich der Gesundheitsinformationen zeigt YouTube einen Gegenentwurf zur häufig postulierten Zurückhaltung bei redaktionellen Eingriffen. Bei Suchanfragen zu

akuten Notfällen werden Inhalte verifizierter Partnerorganisationen prominent priorisiert. Im Rahmen einer globalen Initiative, die unter anderem in Deutschland und Kanada implementiert wurde, setzt die Plattform auf einen Ansatz der aktiven Kuration durch sogenannte „Information Shelves“. Bei Suchanfragen zu akuten medizinischen Notfällen, etwa Herzinfarkt, CPR, Suizidprävention oder Opioid-Überdosis, greift die Plattform in die Ergebnisliste ein. In diesen Bereichen werden kompakte Inhalte vorab verifizierter, „autoritativer“ Partnerorganisationen wie Rotes Kreuz oder Kliniken prominent priorisiert, statt die Ausspielung primär den auf Verweildauer optimierten Engagement-Algorithmen zu überlassen (Graham, 2025; YouTube, 2025).

Diese signifikante editorische Intervention demonstriert, dass Plattformen technisch und operativ durchaus in der Lage sind, ihre Rolle als „Gatekeeper“ verantwortungsvoll wahrzunehmen. Dies gelingt insbesondere dann, wenn ein breiter gesellschaftlicher Konsens, hier der Lebensschutz, den ökonomischen Druck zur Engagement-Maximierung überwiegt. Dass dieser kuratierte Ansatz auch auf Nutzerseite Akzeptanz findet, belegen Mohamed und Shoufan (2024): In ihrer Studie empfanden 87,6 % der Rezipienten diese priorisierten Inhalte als hilfreich für ihre Entscheidungsfindung. Damit markiert das Vorgehen im Public-Health-Sektor einen klaren Kontrast zur strategischen Abkehr von starken inhaltlichen Vorgaben in politisch kontroverseren Bereichen.

Ein zentrales Hindernis für die wissenschaftliche Analyse und politische Einhegung dieser Entwicklungen bleibt der Zugang zu relevanten Daten. Allen et al. (2025) betonen in ihrem Global Transparency Audit, dass die derzeitigen Transparenzberichte der Plattformen oft unvollständig sind und keine unabhängige Verifizierung der Risikobewertungen erlauben. Die Forschung fordert daher neue gesetzliche Rahmenbedingungen, die über freiwillige Datenspenden hinausgehen. Der UK Data (Use and Access) Act 2025 sowie Artikel 40(12) des DSA stellen hierbei wichtige Meilensteine dar, da sie akkreditierten Forschern erstmals einen gesetzlichen Anspruch auf Zugang zu Online-Sicherheitsdaten gewähren. Dies ermöglicht den Übergang zu einer unabhängigen „Trace Research“, die Moderationsentscheidungen wie Shadowbanning oder Löschungen systematisch nachverfolgen kann, anstatt sich auf die kuratierten Datensätze der Plattformen verlassen zu müssen.

4.2.1.2 European Democracy Shield

Ergänzend zum DSA hat die EU mit dem EDS ein Instrument geschaffen, das spezifisch auf die hybride Außenbedrohung reagiert. Am 12. November 2025 legten die Europäische Kommission und der Hohe Vertreter die gemeinsame Mitteilung "European Democracy Shield" (EDS) vor, die darauf abzielt, starke und widerstandsfähige Demokratien in der EU zu fördern. Als Begründung für die Initiative wird eine Zunahme interner und externer Bedrohungen angeführt. Der Informationsraum habe sich zu einem zentralen Austragungsort geopolitischer Auseinandersetzungen entwickelt, auf dem autoritäre Regime wie Russland durch hybride Angriffe, Destabilisierung und FIMI-Kampagnen

(Foreign Information Manipulation and Interference) einen "asymmetrischen" Kampf führen, um das Vertrauen der Bürger in demokratische Institutionen zu untergraben.

Das EDS baut auf bestehenden Maßnahmen wie dem European Democracy Action Plan und dem Defence of Democracy-Paket auf und gliedert sich in drei zentrale Säulen:

1. Verstärkung der situativen Wahrnehmung und Schutz der Integrität des Informationsraums.
2. Stärkung demokratischer Institutionen, freier und fairer Wahlen sowie freier und unabhängiger Medien.
3. Förderung der gesellschaftlichen Resilienz und der Bürgerbeteiligung.

Eine der Kernmaßnahmen ist die Schaffung eines neuen "European Centre for Democratic Resilience". Dieses Zentrum soll als zentrale Drehscheibe für den Informationsaustausch, die operative Zusammenarbeit und den Kapazitätsaufbau zwischen den EU-Institutionen und den Mitgliedstaaten dienen. Flankiert wird dies durch eine "Stakeholder Platform", die zivilgesellschaftliche Akteure wie Faktenprüfer und Forscher einbindet.

Darüber hinaus kündigt die Mitteilung eine Vielzahl weiterer Aktionen an, darunter die Einrichtung eines "European Network of Fact-Checkers" , die Erweiterung des Mandats des European Digital Media Observatory (EDMO) , die Vorbereitung eines "DSA incidents and crisis protocol" und ein "Media Resilience Programme" zur Stärkung des Journalismus. Die Finanzierung soll primär über bestehende und künftige EU-Programme wie Creative Europe, Digital Europe, CERV und das vorgeschlagene Programm "AgoraEU" erfolgen .

Die Veröffentlichung des "European Democracy Shield" stieß auf ein geteiltes Echo, das von institutioneller Bestätigung bis hin zu fundamentaler Kritik an der strategischen Ausrichtung reicht. Während Akteure aus dem unmittelbaren Umfeld des Europäischen Auswärtigen Dienstes (EAD), wie *EUvsDisinfo*, die Initiative als notwendige „erste Verteidigungslinie“ im „geopolitischen Informationskrieg“ gegen autoritäre Regime einordneten und den proaktiven Ansatz hervorhoben, wurde das Maßnahmenpaket in der externen Fachöffentlichkeit deutlich differenzierter und teils kritisch bewertet.

Ein zentraler Diskussionspunkt betrifft die wahrgenommene Diskrepanz zwischen der Problembeschreibung und den vorgeschlagenen Lösungsansätzen. Kritiker monieren, dass die Kommission zwar gravierende Herausforderungen wie ausländische Einmischung und Polarisierung identifiziert, die abgeleiteten Maßnahmen jedoch hinter den Erwartungen zurückbleiben. Die Rezeption in der Fachpresse deutet darauf hin, dass die Initiative bei ihrer Vorstellung eher verhalten aufgenommen wurde. Diese Zurückhaltung wird insbesondere auf den unverbindlichen Charakter vieler der über 40 identifizierten Aktionspunkte zurückgeführt, die häufig auf weichen Formulierungen wie „unterstützen“ oder „stärken“ basieren, anstatt verbindliche Verpflichtungen zu etablieren. Als ursächlich für diese Ambitionslücke werden neben den begrenzten legislativen Kompetenzen der

EU auch geopolitische Rücksichtnahmen diskutiert. So wird in Berichten angeführt, dass der Entwurf aufgrund von Bedenken, die Regulierung könnte in den USA als Einschränkung der Meinungsfreiheit interpretiert werden, sowie auf Druck von US-Technologieunternehmen hin abgeschwächt worden sei.

Inhaltlich wird die starke Fokussierung auf ausländische Informationsmanipulation (FIMI) problematisiert. Analysten merken an, dass dieser Schwerpunkt „überdimensioniert" sei und die Relevanz inländischer Desinformationsquellen vernachlässige. Fundamentalere Kritik äußern zivilgesellschaftliche Akteure wie The Future of Free Speech (2025). Sie verweisen auf das in Kapitel 2.4 diskutierte Problem der schwachen empirischen Evidenzlage zur kausalen Wirkung von Desinformation und argumentieren, dass deren Prävalenz und Einfluss oft überschätzt werde. Der vorherrschende Diskurs trage Züge einer „moral panic". Diese Haltung unterstreicht die in Kapitel 2.4 ebenfalls erörterte Gefahr, dass überzogene Warnungen vor Desinformation das Vertrauen in demokratische Institutionen paradoxerweise stärker untergraben könnten als die Falschinformationen selbst. In dieser Logik wird auch die instrumentelle Rolle von Faktenprüfern kritisch hinterfragt; während Branchenvertreter die Unterstützung begrüßen, wird Fact-Checking als eine in ihrer Wirksamkeit begrenzte Maßnahme bewertet, die strukturelle Probleme nicht adressiere (vgl. Kapitel 2.5 und 4.1).

Aus einer souveränitätspolitischen Perspektive wird zudem argumentiert, dass der Shield zwar die regulatorische Kompetenz der EU betone, jedoch keine hinreichende „strategische Autonomie" im Bereich der digitalen Infrastruktur schaffe. Die Abhängigkeit von nicht-europäischen Plattformbetreibern bleibe bestehen, was die Durchsetzungsfähigkeit europäischer Standards limitiere. Industrieverbände wie die Association of Commercial Television and Video on Demand Services in Europe (ACT, 2025) begrüßen die Initiative prinzipiell als zeitgemäß, verknüpfen ihre Zustimmung jedoch mit der Forderung nach einer konsequenten Durchsetzung des bestehenden Rechtsrahmens (DSA, DMA, EMFA). Um die Medienvielfalt nachhaltig zu sichern, sei zudem eine Neuausrichtung des Werbemarktes erforderlich, die Qualitätsjournalismus gegenüber Desinformation ökonomisch incentiviert. Ähnlich mahnen Faktenprüfer-Verbände an, dass die vorgesehenen Finanzierungsinstrumente (z. B. AgoraEU) angesichts der massiven Investitionen antagonistischer Akteure in Desinformationskampagnen unzureichend seien (EFCSN, 2025b).

4.2.2 Internationaler Vergleich

Die globale Regulierungslandschaft zur Bekämpfung von Desinformation ist durch eine grundlegende Dichotomie gekennzeichnet. Während die Europäische Union mit dem Digital Services Act (DSA) einen prozessorientierten Ansatz verfolgt, der auf Transparenz, systemische Risikominimierung und kooperative Verantwortlichkeit setzt, etabliert sich in vielen Staaten Asiens, Afrikas und Lateinamerikas ein inhaltszentriertes, strafrechtliches Paradigma. Diese als „Fake-News"- oder „Misinformation Laws" bezeichneten

Regelwerke werden in der akademischen Literatur zunehmend kritisch als Instrumente des „Lawfare“ analysiert. Gemeint ist die strategische Nutzung von Gesetzen zur Unterdrückung politischer Opposition und Zivilgesellschaft (Mahapatra et al., 2024; Cho, 2025; Bradshaw et al., 2025). Parallel dazu haben internationale Organisationen wie UNESCO, OSZE, der Interamerikanische Menschenrechtsschutz sowie die Vereinten Nationen menschenrechtliche Leitlinien formuliert. Darin wird vor vage gefassten Verboten von „Fake News“ gewarnt und stattdessen auf systemische, rechtsstaatskonforme und multistakeholder-basierte Governance gesetzt (UNESCO, 2023; OSCE Representative on Freedom of the Media et al., 2017; Organization of American States [OAS], 2019).

Aktuelle empirische Arbeiten zeigen, dass sich seit rund einem Jahrzehnt ein weltweiter Trend zur Kodifizierung von Desinformationsregulierung herausgebildet hat. Bradshaw et al. (2025) dokumentieren in einer globalen Erhebung, dass zwischen 2010 und 2022 in 80 Staaten neue Gesetze gegen Desinformation erlassen oder bestehende Normen in diesem Sinne verschärft wurden. Auf Grundlage eines Datensatzes aus 177 Ländern und 105 Rechtsakten arbeiten die Autorinnen heraus, dass diese Normen zwar formal mit der Bekämpfung von „Misinformation“ begründet werden, faktisch aber sehr unterschiedlich ausgestaltet sind: Sie reichen von medienrechtlichen Transparenzpflichten und Regeln für politische Online-Werbung bis hin zu strafrechtlichen Tatbeständen, die mit Haftstrafen sanktioniert werden und erhebliche chilling effects auf die Pressefreiheit ausüben. Zu vergleichbaren Schlussfolgerungen gelangte die Analyse für das Center for International Media Assistance (CIMA). Anhand von 105 Gesetzen identifiziert CIMA vier Sanktionskategorien, nämlich exzessive Geldstrafen, Haft, Löschpflichten und administrativ-bürokratische Auflagen. Die Auswertung zeigt, dass diese Instrumente in zahlreichen Fällen gezielt gegen Journalistinnen und Journalisten eingesetzt werden (Lim & Bradshaw, 2023).

In weiten Teilen Süd- und Südostasiens hat sich eine Regulierungspraxis etabliert, die Desinformation primär als Sicherheitsbedrohung für die staatliche Ordnung rahmt. Mahapatra et al. (2024) beschreiben in ihrer Analyse für das GIGA-Institut, wie Regierungen in dieser Region vage definierte Gesetze gegen Falschinformationen nutzen, um Kritikerinnen und Kritiker zu kriminalisieren. Der Begriff „Fake News“ wird dabei juristisch oft so unpräzise gefasst, dass er auch legitime Meinungsäußerungen oder investigativen Journalismus umfasst, sofern diese der Regierungslinie widersprechen. Diese Praxis des „Lawfare“ führt dazu, dass Journalistinnen und Journalisten sowie Aktivistinnen und Aktivisten in langwierige, kostenintensive Gerichtsprozesse verwickelt werden, die weniger der Wahrheitsfindung als vielmehr der finanziellen und psychischen Erschöpfung der Beklagten dienen.

Cho (2025) zeigt für Südostasien, dass illiberale Regime „Fake News“ strategisch nutzen, um entweder eine als feindlich wahrgenommene Öffentlichkeit aktiv zu bekämpfen oder, im passiven Modus, pro-regierungsnahe Narrative gezielt zu verstärken. Am Beispiel Singapurs und der Philippinen entwickelt sie eine Typologie, in der der Schutz angeblich „verwundbarer“ Bevölkerungsgruppen oder die Abwehr „ausländischer Einflussnahme“

als Argument herangezogen wird, um drakonische Befugnisse zur Entfernung von Inhalten, zur Anordnung von Gegendarstellungen und zur strafrechtlichen Verfolgung zu legitimieren (Cho, 2025). Putra (2024) bestätigt diesen Einblick für den ASEAN-Raum und weist auf ein strukturelles Dilemma hin: Das in der Region hochgehaltene Prinzip der „Nichteinmischung“ erschwert eine koordinierte transnationale Bekämpfung von Desinformation, während nationale Gesetze oft selektiv gegen innenpolitische Gegner eingesetzt werden. In Ländern wie Thailand oder den Philippinen werden Anti-Desinformations-Gesetze häufig mit Vorwürfen des Verrats oder der Gefährdung der nationalen Sicherheit verknüpft, was die Hemmschwelle für staatliche Eingriffe senkt.

Auch außerhalb Asiens lässt sich diese Verschiebung hin zu sicherheitszentrierten und repressiven Antworten beobachten. In mehreren afrikanischen Staaten wurden strafrechtliche Tatbestände geschaffen, die die „Verbreitung falscher Informationen“ unter Haftandrohung stellen und in der Praxis gegenüber oppositionellen Stimmen und kritischen Medien überproportional zur Anwendung kommen (Omilusi, 2025). Forschungen des United Nations University Centre for Policy Research verdeutlichen zugleich, dass Desinformationsnarrative in Konfliktregionen Subsahara-Afrikas vielfach mit ethnischen Spannungen, Wahlgewalt und der Delegitimierung internationaler Friedensmissionen verwoben sind, während rechtliche Gegenmaßnahmen die Balance zwischen Sicherheit und Meinungsfreiheit häufig nicht erreichen (Albrecht et al., 2024).

Die Corona-Pandemie wirkte häufig als Beschleuniger, wie Analysen aus Lateinamerika und (Süd-)Ostasien zeigen. Notstandsverordnungen erweiterten exekutive Kompetenzen und erlaubten es Regierungen, kritische Berichterstattung über das Krisenmanagement als potenziell „gefährliche Desinformation“ zu brandmarken (Asia Centre, 2021; Konrad-Adenauer-Stiftung, 2022).

Eine globale Analyse des Center for News, Technology & Innovation (CNTI) unterstreicht, dass dieses Phänomen kein regionales Spezifikum ist. In einer Untersuchung von Gesetzgebungen in über fünfzig Ländern wurde festgestellt, dass Gesetze, die explizit auf „Fake News“ abzielen, das Risiko bergen, mehr Schaden als Nutzen anzurichten (CNTI, 2024). Die mangelnde Trennschärfe zwischen unbeabsichtigter Fehlinformation und gezielter Desinformation ermöglicht es autoritären wie auch hybrid-demokratischen Regierungen, die Deutungshoheit über die „Wahrheit“ zu monopolisieren und die Pressefreiheit durch drohende Haftstrafen oder exzessive Bußgelder einzuschränken. Lim und Bradshaw (2023) bezeichnen diese Entwicklung als „Chilling Legislation“: Die bloße Existenz solcher Gesetze erzeugt eine abschreckende Wirkung, die dazu führt, dass Journalistinnen und Journalisten präventiv Selbstzensur üben, um juristische Repressalien zu vermeiden. Die Autorinnen zeigen auf Basis von 105 analysierten Rechtsakten, dass fast ein Zehntel aller weltweit inhaftierten Journalistinnen und Journalisten im Jahr 2022 auf Grundlage von Desinformationsgesetzen in Haft saß.

Der kritische Blick internationaler Menschenrechtsinstitutionen richtet sich daher zunehmend auf die Vereinbarkeit solcher Normen mit völkerrechtlichen Standards. Bereits

2017 warnten die OSZE, der UN-Menschenrechtsrat, die Afrikanische und die Interamerikanische Menschenrechtskommission in einer gemeinsamen Erklärung vor Gesetzen, die „Fake News“ oder „Propaganda“ pauschal unter Strafe stellen und dadurch legitime Kritik unterbinden könnten (OSCE, 2017). Die OSZE-Vertreterin für Medienfreiheit bewertete zudem in einem Rechtsgutachten zur französischen „Loi contre la manipulation de l’information“ die Gefahr, dass weit gefasste Eingriffsbefugnisse zum Löschen vermeintlicher Desinformation im Wahlkampf zu Überblockierung und zur Beschneidung der Pluralität politischer Stimmen führen (OSCE, 2019).

Im Kontrast dazu steht der Ansatz der Europäischen Union und insbesondere der baltischen Staaten, der weniger auf das Verbot einzelner Inhalte als auf die Stärkung der gesellschaftlichen Widerstandskraft abzielt. Balčytienė et al. (2025) analysieren den „baltischen Weg“ als einen menschenzentrierten Ansatz, der historische Erfahrungen mit russischer Propaganda in eine umfassende Sicherheitsstrategie übersetzt. In Litauen, Lettland und Estland wird Desinformation nicht isoliert als mediales Problem betrachtet, sondern als hybride Bedrohung, der mit einem „Whole-of-Society“-Ansatz begegnet werden muss. Dies beinhaltet die enge Verzahnung von staatlicher strategischer Kommunikation, unabhängigen Medien, Fact-Checking-Organisationen und einer aufgeklärten Zivilgesellschaft („Civic Preparedness“). Anstatt Inhalte zu löschen, liegt der Fokus auf der Erhöhung der Medienkompetenz und der Förderung eines pluralistischen Informationsökosystems, das immuner gegen externe Manipulation ist (Balčytienė et al., 2025).

Berger et al. (2024) heben in ihrer globalen Vergleichsstudie hervor, dass dieser Ansatz der Resilienzförderung auch in anderen Regionen, etwa in Taiwan oder Finnland, erfolgreich praktiziert wird. Das entscheidende Merkmal ist hierbei die Abkehr vom staatlichen Wahrheitsmonopol hin zu einer dezentralen Verantwortung, bei der Plattformen, Zivilgesellschaft und Staat kooperativ agieren, ohne dass der Staat als alleiniger Schiedsrichter über die Wahrheit auftritt. In Lateinamerika empfehlen der Interamerikanische Menschenrechtsschutz und die damit verbundenen Institutionen ebenfalls, nicht primär auf strafrechtliche Instrumente zu setzen, sondern Transparenz, Datenschutz, politische Werbetransparenz und die Stärkung unabhängiger Medien als zentrale Bausteine eines resilienten Informationsraums zu fördern (OAS, 2019; Pérez Argüello & Barojan, 2019).

Auch außerhalb Europas finden sich Varianten eines Resilienz- und Prozessmodells. Gielow Jacobs (2022) zeigt in ihrem Überblick zur US-Rechtslage, dass die sehr weitgehenden Schutzstandards des First Amendment strafrechtliche Verbote von „Fake News“ stark begrenzen. Stattdessen dominieren Formen der Co-Regulierung, etwa Transparenz-Anforderungen für politische Online-Werbung, freiwillige Selbstverpflichtungen der Plattformen und Förderprogramme zur Medienkompetenz. Obwohl diese Instrumente nicht frei von Kritik sind und ihre Wirksamkeit sowie die Abhängigkeit von Plattformkooperation umstritten bleiben, verdeutlicht der Vergleich, dass liberale Demokratien auch ohne strafrechtliche Verbote versuchen können, Desinformation durch institutionelle Arrangements, zivilrechtliche Haftungsregime und gesellschaftliche Resilienzstrategien einzudämmen (Gielow Jacobs, 2022; Bradshaw et al., 2025).

Eine spezifische Zwischenstellung nimmt die Ukraine ein, die aufgrund des russischen Angriffskrieges unter extremen Bedingungen agiert. Marushchak et al. (2025) zeigen, dass der ukrainische Regulierungsrahmen gezwungen ist, Elemente der Resilienzförderung mit harten, sicherheitsorientierten Maßnahmen zu kombinieren. Angesichts einer erheblichen Welle KI-generierter Desinformation, etwa Deepfakes politischer Entscheidungsträger oder synthetische Propaganda, hat die Ukraine ihre Gesetzgebung angepasst, um die Verbreitung solcher Inhalte als Teil der hybriden Kriegsführung zu erfassen. Die Autorinnen argumentieren, dass in einem existenziellen Konflikt die staatliche Regulierung von KI-Inhalten, die enge Kooperation mit Plattformen bei der schnellen Identifikation feindlicher Operationen sowie Maßnahmen zur Abschottung des Informationsraums unverzichtbar sind (Marushchak et al., 2025).

Zugleich bleibt der Anspruch bestehen, grundlegende freiheitliche Prinzipien so weit wie möglich zu wahren. Ukrainische Institutionen arbeiten eng mit internationalen Partnern zusammen, um Schutzmechanismen gegen russische Einflussoperationen zu entwickeln, ohne unabhängige Medien und zivilgesellschaftliche Akteure zu entmündigen (Berger et al., 2024). Der Fall verdeutlicht, dass das liberale Resilienz-Modell in akuten Bedrohungslagen an Grenzen stoßen kann und durch defensive Maßnahmen ergänzt werden muss, um die Integrität des Informationsraums zu schützen. Die Herausforderung liegt darin, sicherheitsrechtliche Sonderregeln nicht dauerhaft zu verstetigen und so den Übergang zurück zu einem stärker prozessorientierten Normalregime zu ermöglichen.

Zusammenfassend lässt sich eine deutliche Trennlinie in der globalen Governance von Desinformation ziehen. Während das europäische Modell, verkörpert durch den DSA, auf Prozessregulierung setzt, werden Risikomanagement-Systeme, algorithmische Transparenz, die Offenlegung von Werbepraktiken sowie der Zugang zu Daten für Aufsicht und Forschung überprüft. In vielen Staaten des Globalen Südens und in autokratischen Kontexten dominieren hingegen inhaltsorientierte Ansätze. Diese zielen auf das Verbot bzw. die strafrechtliche Sanktionierung bestimmter Aussagen, deren Definition häufig vage bleibt und so politischen Missbrauch ermöglicht (Mahapatra et al., 2024; Cho, 2025; Omilusi, 2025).

Bradshaw et al. (2025) machen deutlich, dass der rasche weltweite Anstieg von Misinformation-Gesetzen weniger durch einen tatsächlichen Anstieg falscher Informationen erklärt werden kann. Ausschlaggebend sind vielmehr politische Opportunitätsstrukturen, obwohl es falsche Informationen historisch immer gegeben hat: Dazu zählen die Popularisierung des „Fake-News“-Diskurses durch Eliten, die sicherheitspolitische Rahmung von Informationsphänomenen, das Interesse von Regierungen an der Kontrolle des Informationsflusses und die zunehmende Sichtbarkeit von Plattformversagen in der Öffentlichkeit. In dieser Konstellation wird Desinformation leicht zur Projektionsfläche für tiefer liegende Konflikte um Macht, Legitimität und gesellschaftliche Pluralität.

Die völkerrechtlichen und regionalen Menschenrechtsstandards formulieren demgegenüber relativ klare Kriterien: Eingriffe in die Meinungsfreiheit müssen gesetzlich präzise

gefasst, notwendig und verhältnismäßig sein sowie legitimen Zielen wie dem Schutz der nationalen Sicherheit, der öffentlichen Ordnung oder der Rechte anderer dienen (OSCE, 2017; OAS, 2019; UNESCO, 2023). Sowohl die OSZE als auch die Interamerikanische Menschenrechtskommission warnen daher ausdrücklich vor Generalklauseln gegen „Fake News“, die diesen Anforderungen nicht genügen und de facto als Zensurinstrument fungieren (OSCE, 2019; OAS, 2019).

Vor diesem Hintergrund kann der DSA als Versuch gelesen werden, einen dritten Weg zwischen staatlichem Wahrheitsmonopol und weitgehend unregulierten Plattformmärkten zu etablieren. Er verschiebt die regulatorische Aufmerksamkeit weg von einzelnen Beiträgen hin zu den Strukturen und Prozessen der Plattformen, setzt auf Multi-Stakeholder-Mechanismen und verknüpft Transparenz, Rechenschaftspflicht und kooperative Selbstregulierung mit behördlicher Aufsicht und empfindlichen Sanktionen bei systemischer Non-Compliance (Strowel & De Meyere, 2023; Griffin, 2025; Husovec, 2024). Die erste Geldbuße gegen X im Dezember 2025 illustriert, dass insbesondere Verstöße gegen Transparenz- und Kooperationspflichten als gravierende Risiken für die demokratische Öffentlichkeit verstanden werden. Ein Beispiel ist die Behinderung wissenschaftlicher Forschung (Europäische Kommission, 2025).

Mede et al. (2025) weisen in ihrer globalen Studie zur Wissenschaftskommunikation darauf hin, dass Vertrauen in Informationen stark davon abhängt, wie transparent und nachvollziehbar die Quellen und die dahinter stehenden Prozesse sind. Der europäische Weg versucht, dieses Vertrauen durch systemische Rechenschaftspflicht wiederherzustellen, indem er Plattformen zwingt, ihre internen Mechanismen offenzulegen und in dialogische Risikobewertungen einzutreten. Repressive Ansätze dagegen ersetzen Vertrauen häufig durch erzwungene Konformität: Sie erzeugen kurzfristig ein scheinbar „ruhiges“ Informationsumfeld, unterminieren aber langfristig die Glaubwürdigkeit staatlicher Institutionen, die Innovationskraft des digitalen Raums und die Fähigkeit der Gesellschaft, zwischen verlässlichen und unzuverlässigen Informationen zu unterscheiden (Lim & Bradshaw, 2023; Bradshaw et al., 2025).

Die empirische Evidenz der Jahre 2024 und 2025 legt somit nahe, dass Prozessregulierung und Resilienzförderung mittelfristig besser mit demokratischen und menschenrechtlichen Standards vereinbar sind als inhaltszentrierte, strafrechtliche Instrumente. Gleichzeitig zeigt der Blick auf Konfliktkontexte wie die Ukraine oder auf Staaten mit extrem polarisierten Öffentlichkeiten, dass auch prozessorientierte Modelle ergänzende Schutzmechanismen benötigen, um in akuten Krisenlagen handlungsfähig zu bleiben. Die zentrale Frage für die kommenden Jahre wird daher sein, ob es gelingt, die im DSA und in internationalen Leitlinien angelegte Balance zwischen Schutz vor Desinformation und Schutz vor repressiven Eingriffen in die Meinungsfreiheit in der Praxis zu stabilisieren. Zudem stellt sich die Frage, ob dieses Modell über Europa hinaus als Referenzrahmen für eine global gerechtere und rechtsstaatlich eingebettete Plattformregulierung dienen kann.

5 Fazit

Die umfassende Untersuchung der Desinformationsphänomene und der damit verbundenen Online-Schäden legt die Schlussfolgerung nahe, dass zu ihrer effektiven Bewältigung ganzheitliche und strukturelle Ansätze erforderlich sind. Der Fokus sollte sich von der Einzelprüfung des Wahrheitsgehalts weiter hin zu prozessualen und systemischen Risiken verlagern, insbesondere bei Inhalten, die keine expliziten Inhaltsstandards verletzen ("lawful but awful"). Stattdessen sind das Schadenspotenzial durch Verbreitungsmuster, Netzwerkdynamiken und Systemrisiken in den Mittelpunkt zu stellen, was die Identifikation von akteurszentrierten statt inhaltsbasierten Maßnahmen prioritär macht.

Die wesentlichen Erkenntnisse aus der Analyse bestätigen, dass die Verbreitung von Desinformation keine rein zufällige Angelegenheit ist. Sie ist tief in der Ökonomie der Aufmerksamkeit verwurzelt, wobei das dominante Plattform-Geschäftsmodell der Maximierung der Verweildauer durch die algorithmische Bevorzugung von emotionalen und polarisierenden Inhalten Fehlanreize schafft. Diese algorithmische Verstärkung kann systematisch bestimmte politische Akteure begünstigen und Inhalte, die Wut und Feindseligkeit auslösen, mit einem Reichweiten-Bonus belohnen. Ein oft unterschätzter Treiber ist zudem das intransparente Ad-Tech-Ökosystem, das die "Disinformation-for-Profit"-Industrie unbeabsichtigt finanziert, indem programmatische Werbung auf minderwertigen oder falschen Inhalten platziert wird. Gleichzeitig verändert die massenhafte Verfügbarkeit Generativer KI die Produktionsbedingungen der Desinformation. Neben der Gefahr perfekter Deepfakes besteht eine systemische Bedrohung auch in der Flutung des Informationsraums mit qualitativ minderwertigen, KI-generierten Inhalten („AI Slop“). Diese Inhalte können zur Verwässerung des Informations- und Erkenntnisraums führen und die Verifikationskosten erheblich erhöhen.

Die Untersuchung der Wirkmechanismen zeigt, dass die Anfälligkeit der Nutzer nicht primär auf mangelndem Wissen, sondern auf psychologischen Dispositionen wie dem "Identity-Protective Motivated Reasoning" beruht, was dazu führt, dass Falschinformationen akzeptiert werden, wenn sie die eigene Gruppenidentität stützen. Diese Anfälligkeit wird durch Phänomene wie den "News-Finds-Me"-Effekt verstärkt, bei dem passive Rezeption über Feeds die kritische Prüfung von Quellen hemmt. Die Schäden manifestieren sich dabei nicht nur im digitalen Raum, sondern können zu physischer Gewalt führen (stochastischer Terrorismus) und die demokratische Teilhabe strukturell zersetzen, etwa durch koordinierte Kampagnen (Networked Misogyny) zum "Silencing" von Frauen in öffentlichen Positionen.

Angesichts dieser systemischen Herausforderungen legt die Studie die Schlussfolgerung nahe, dass die globale Regulierungsdichotomie beachtet werden muss. Während das europäische Modell, verkörpert durch den Digital Services Act (DSA), einen prozessorientierten Ansatz verfolgt, der auf die Transparenz der Algorithmen und die systemische Risikominderung abzielt, stehen dem inhaltszentrierte, oft repressive „Fake-News“-Gesetze in autokratischen und illiberalen Kontexten gegenüber, die hier tatsächlich als

Instrumente des „Lawfare“ zur Unterdrückung kritischer Stimmen genutzt werden können. Das europäische Modell zielt darauf ab, die Integrität des Informationsraums durch die Stärkung der systemischen Auditierbarkeit der Very Large Online Platforms (VLOPs) zu sichern, wobei auch die fehlende Kooperation der Plattformen mit Forschenden sanktioniert wird.

Auf Grundlage dieser Analyseergebnisse leiten sich prioritäre Policy-Optionen ab. Erstens muss der stärkere Fokus auf das Ad-Tech-Ökosystem erfolgen, um die ökonomischen Fehlanreize der Desinformationsverbreitung zu unterbinden. Zweitens ist die internationale Koordination bei der Definition und Erkennung von Koordiniertem Inauthentischem Verhalten (CIB) von geopolitischen Akteuren zu verbessern. Drittens muss im Rahmen des DSA eine Prüfung eines Verbots von „manipulativen Verbreitungstechniken“ erfolgen. Im Bereich der Interventionen zeigt sich, dass es kein Allheilmittel gibt. Während Accuracy Nudges eine skalierbare Methode zur Reduzierung des Teilungsverhaltens darstellen, und Community Notes als vertrauenswürdiges nutzerbasiertes Korrektiv wirken, bleiben letztere aufgrund des notwendigen parteiübergreifenden Konsenses bei kontroversen Themen in ihrer Skalierbarkeit gehemmt. Auch Inokulationsstrategien (Prebunking) müssen mit Vorsicht angewandt werden, da sie bei unsachgemäßer Kalibrierung zum Skeptizismus-Paradoxon führen können. In diesem Fall verlieren Nutzer das Vertrauen in alle Quellen, einschließlich verlässlicher.

Trotz der erzielten Synthese bleiben wichtige offene Fragen für die zukünftige Forschung bestehen. Ein zentrales methodisches Problem ist der weiterhin begrenzte empirische Kenntnisstand zu Kausalität und Exposition von Desinformation. Eng damit verknüpft ist die Notwendigkeit zu klären, wie der Datenzugang für die Forschung nachhaltig verbessert werden kann, um eine unabhängige Auditierung der algorithmischen Mechanismen zu gewährleisten. Schließlich bedarf es weiterer Forschung, um die ökonomischen Kosten und den Nutzen verschiedener Gegenmaßnahmen effektiv bewerten zu können. Dabei ist zu untersuchen, wie sich Ansätze von der Plattformregulierung bis hin zu Marktmechanismen einordnen lassen und welche Balance zwischen internationaler Koordination und regulatorischer Fragmentierung jeweils angemessen ist. Die Auseinandersetzung mit diesen Fragen wird darüber entscheiden, ob es gelingt, die Chancen für KI (z. B. in der Detektion) und die digitale Aufklärung systemisch so zu nutzen, dass die Resilienz der demokratischen Öffentlichkeit gestärkt wird.

6 Literaturverzeichnis

Abdul Rahman, E., Campaioli, G., Rea, S., Di Bartolomeo, S., Keim, B., and Wörner, L., and Ochsner, B., “Supercharging Online Harassment: Amplifiers and Indirect Swarming and Their Potential Threat to Democracies”, forthcoming.

Ahmad, W., Sen, A., Eesley, C., & Brynjolfsson, E. (2024). Companies inadvertently fund online misinformation despite consumer backlash. Nature, 630(8015), 123-131.

Albrecht, E., Fournier-Tombs, E., & Brubaker, R. (2024). Disinformation and peacebuilding in Sub-Saharan Africa: Security implications of AI-altered information environments. New York, NY: United Nations University and Interpeace.

Allen J., Gurley S., Bonilla S., Shen N., Global Transparency Audit, Integrity Institute, 2025, https://drive.google.com/file/d/1MJHx4cx24XV4UZUfW1iIoMKPRkqIyLQa/view

Allen, J. (2022). Misinformation amplification analysis and tracking dashboard. Integrity Institute, October, 13.

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. Science, 384(6699). https://doi.org/10.1126/science.adk3451

AlQahtani, F. A. (2025). Trust or Trickery? A Systematic Review of Greenwashing and Branding. International Review of Management and Marketing, 15(6), 424–432. https://doi.org/10.32479/irmm.20758

Alvarado Rincón D and Meyer-Resende M (2025) Big tech is backing out of commitments counter-ing disinformation—What’s next for the EU’s code of practice? | Democracy ReportingInternational. 7 February. https://democracy-reporting.org/en/office/EU/publications/big-tech-is-backing-out-of-commitments-countering-disinformation-whats-next-for-the-eus-code-of-practice

Amnesty International (2025). Written evidence. House of Commons, Science, Innovation and Technology Committee inquiry into Social Media, Misinformation, and Harmful Algorithms. https://committees.parliament.uk/work/8641/social-media-misinformation-and-harmful-algorithms/publications/written-evidence/

Ansari, M. S. (2025). AI Slop and Data Pollution in the Age of Generative AI: Strategic Risks, Economic Consequences, and Governance Pathways for Business, Management, and the Creative Industries. https://doi.org/10.2139/ssrn.5649410

APA / Van Der Linden, S., Albarracín, D., Fazio, L. K., Deen Freelon, Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2023). Using Psychological Science to Understand and Fight Health Misinformation AN APA CONSENSUS STATEMENT. November 2023. https://doi.org/10.13140/RG.2.2.18193.13929

Arcuri, M. C., Gandolfi, G., & Russo, I. (2023). Does fake news impact stock returns? Evidence from US and EU stock markets. Journal of Economics and Business, 125, 106130.

Asia Centre. (2021). Defending Freedom of Expression: Fake News Laws in East and Southeast Asia. https://asiacentre.org/defending-freedom-of-expression-fake-news-laws-in-east-and-southeast-asia/

Assenza, T, F Collard, P Fève and S J Huber (2024), “From Buzz to Bust: How Fake News Shapes the Business Cycle”, CEPR Working Paper 18912.

Association of Commercial Television and Video on Demand Services in Europe (ACT). (2025, November 12). ACT welcomes European Democracy Shield and calls for actions to support media sustainability. https://www.acte.be/publication/act-welcomes-european-democracy-shield-and-calls-for-actionsto-support-media-sustainability/

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences, 115(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. Digital journalism, 6(2), 154-175.

Balčytienė, A., Dāvidsone, A., & Siibak, A. (2025). What a Human-Centred Approach Reveals About Disinformation Policies: The Baltic Case. Media and Communication, 13.

Bassin, I., & Potter, M. (2024, October 8). On anticipatory obedience and the media. Columbia Journalism Review. https://www.cjr.org/analysis/anticipatory-obedience-bassin-potter-scheppele-orban-trump-hungary-media-punish.php

Bayer, J., Bitiukova, N., Bard, P., Szakács, J., Alemanno, A., & Uszkiewicz, E. (2019). Disinformation and propaganda–impact on the functioning of the rule of law in the EU and its Member States. European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. European journal of communication, 33(2), 122-139.

Benton, J. (2021, August 24). Facebook sent a ton of traffic to a Chicago Tribune story. So why is everyone mad at them? Nieman Lab. https://www.niemanlab.org/2021/08/facebook-sent-a-ton-of-traffic-to-a-chicago-tribune-story-so-why-is-everyone-mad-at-them/

Berger, C., Freihse, C., & Meyer zu Schwabedissen, O. (2024). Effectively countering disinformation - Perspectives from every continent. https://doi.org/10.11586/2024076

Biddlestone, M., Ziemer, C. T., Maertens, R., Roozenbeek, J., & van der Linden, S. (2025). Norm-enhanced prebunking for actively open-minded thinking indirectly improves misinformation discernment and reduces conspiracy beliefs. Journal of Experimental Social Psychology, 121, 104818.

Blas, L., Saraf, D., Salkar, T., Adadurova, N., Luceri, L., & Ferrara, E. (2025). Large-scale detection of multilingual coordinated activity on Telegram. npj Complexity, 2(1), 33.

Borges do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: a systematic review of reviews. Bulletin of the World Health Organization, 100(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Botan, M., & Meyer, T. (2025). Implementing the EU Code of Practice on Disinformation: An Evaluation of VLOPSE Compliance and Effectiveness (Jan–Jun 2024). EDMO: European Digital Media Observatory. https://edmo.eu/publications/implementing-the-eu-code-of-practice-on-disinformation-an-evaluation-of-vlopse-compliance-and-effectiveness-jan-jun-2024/

Botero Arcila, B., & Griffin, R. (2023). Social media platforms and challenges for democracy, rule of law and fundamental rights. Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies, PE.

Bradshaw, S., & Howard, P. N. (2018). The global organization of social media disinformation campaigns. Journal of International Affairs, 71(1.5), 23-32.

Bradshaw, S., & Howard, P. N. (2019). The global disinformation order: 2019 global inventory of organised social media manipulation.

Bradshaw, S., Lim, G., & Haque, M. (2025). True Costs of Misinformation| The Global Spread of Misinformation Laws. International Journal of Communication, 19, 23.

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. Trends in Cognitive Sciences, 27(10), 947-960.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313-7318.

Braun, J. A., & Eklund, J. L. (2019). Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism. Digital Journalism, 7(1), 1–21. https://doi.org/10.1080/21670811.2018.1556314

Brooks, P., & Duetz, J. (2025). Conspiracy accusations. Inquiry, 68(8), 2798-2819.

Bruns, A. (2019). Are filter bubbles real?. John Wiley & Sons.

Burstin, H., Curry, S., Ranney, M. L., Arora, V., Wachler, B. B., Chou, W. Y. S., ... & Wallace, K. (2023). Identifying Credible Sources of Health Information in Social Media: Phase 2—Considerations for Non-Accredited Nonprofit Organizations, For-Profit Entities, and Individual Sources. NAM perspectives, 2023, 10-31478.

CACI. (2025, February 7). Disinformation-as-a-service: The cybercrime epidemic destabilizing the world. DarkBlue | CACI. https://www.caci.com/darkblue/blog/disinformation-as-a-service

Campbell, S. W., & Hawkins, I. (2025). Social (media) psychology of the “news-finds-me” perception: habits, mindsets, and beliefs. Journal of Computer-Mediated Communication, 30(5), zmaf010.

Carrasco-Farré, C. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. Humanit Soc Sci Commun 9, 162 (2022). https://doi.org/10.1057/s41599-022-01174-9

Chadwick, A. (2017). The hybrid media system: Politics and power. Oxford University Press.

Chan, J. (2024). Online astroturfing: A problem beyond disinformation. Philosophy & Social Criticism, 50(3), 507-528.

Chen, S., Gao, M., Sasse, K. et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. npj Digit. Med. 8, 605 (2025). https://doi.org/10.1038/s41746-025-02008-z

Chen, Y. S., & Zaman, T. (2024). Shaping opinions in social networks with shadow banning. Plos one, 19(3), e0299977.

Cho, C. S. M. (2025). Illiberal responses to “fake news” in Southeast Asia. Democratization, 32(5), 1091–1111. https://doi.org/10.1080/13510347.2024.2442395

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2024). Community-based fact-checking reduces the spread of misleading posts on social media (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2409.08781

Chuai, Y., Sergeeva, A., Lenzini, G., & Pröllochs, N. (2025). Community Fact-Checks Trigger Moral Outrage in Replies to Misleading Posts on Social Media. CHI Conference on Human Factors in Computing Systems (CHI ’25), April 26-May 1, 2025, Yokohama, Japan, 1(1). https://doi.org/10.1145/3706598.3713909

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1–52. https://doi.org/10.1145/3686967

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. Decision Support Systems, 160, 113819.

Cinus, F., Minici, M., Luceri, L., & Ferrara, E. (2025, April). Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 us election. In Proceedings of the ACM on Web Conference 2025 (pp. 541-559).

Clemons, E.K., Schreieck, M. & Waran, R.V. (2025) Managing disinformation on social media platforms. Electron Markets 35, 52. https://doi.org/10.1007/s12525-025-00796-6

CNTI (Center for News, Technology & Innovation). (2024). Most 'Fake News' Legislation Risks Doing More Harm than Good amid a Record Number of Elections in 2024. 3 September. https://innovating.news/article/most-fake-news-legislation-risks-doing-more-harm-than-good-amid-a-record-number-of-elections-in-2024

Colizzi, C., Sala, A. A. D., Fenza, G., & Gajewski, L. (2025). Investigating Coordinated Inauthentic Behavior on Alternative Platforms During the 2024 U.S. Election. ICWSM. https://doi.org/10.36190/2025.19

Commonwealth Parliamentary Association (CPA). (2023). Parliamentary Handbook on Disinformation, AI and Synthetic Media.

Council of Europe (2025). Platform to Promote the Protection of Journalism and Safety of Journalists. (2025, March). Europe Press Freedom Report 2024: Confronting political pressure, disinformation, and the erosion of media independence [Report]. Council of Europe. https://rm.coe.int/prems-013425-gbr-2519-annual-report-2025-correction-cartooning/1680b48f7b

Das Progressive Zentrum, & Bertelsmann Stiftung. (2025). How to Sell Democracy Online (Fast). Zenodo. https://doi.org/10.5281/zenodo.17098386Verwiebe, R., Philipp, A., Bobzien, L., Wolfgram, J., Weißmann, S., Kohler, U., & Tjaden, J. (2025). Digitalisiert, politisiert, polarisiert? Bertelsmann Stiftung. https://doi.org/10.11586/2025070

de Cock Buning, M. (2018). A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation. Publications Office of the European Union.

De, S., Bakker, M. A., Baxter, J., & Saveski, M. (2024). Supernotes: Driving Consensus in Crowd-Sourced Fact-Checking. http://arxiv.org/abs/2411.06116

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. Personality and Social Psychology Review, 14(2), 238–257. https://doi.org/10.1177/1088868309352251

Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., & Haimson, O. L. (2024). " What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and" Prove" Shadowbanning. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1-39.Thomas, S., & Manalil, P. (2025). Digital silence: the psychological impact of being shadow banned on mental health and self-perception. Frontiers in Psychology, 16, 1659272.

Dey, D., Lahiri, A., & Mukherjee, R. (2025). Polarization or Bias: Take Your Click on Social Media. Journal of the Association for Information Systems, 26(3), 850-878.

Di Domenico, G., Sit, J., Ishizaka, A., & Nunan, D. (2021). Fake news, social media and marketing: A systematic review. Journal of business research, 124, 329-341.

Di Marco, N., Brunetti, S., Cinelli, M., & Quattrociocchi, W. (2025). Post-hoc evaluation of nodes influence in information cascades: The case of coordinated accounts. ACM Transactions on the Web, 19(2), 1-19.

Di Meco, L., and Brechenmacher, S., (2020). 'Tackling Online Abuse and Disinformation Targeting Women in Politics.' Carnegie Endowment for International Peace. https://carnegieendowment.org/2020/11/30/tackling-online-abuse-and-disinformation-targeting-women-in-politics-pub-83331

Diaz Ruiz, C. A. (2025). Disinformation and fake news as externalities of digital advertising: a close reading of sociotechnical imaginaries in programmatic advertising. Journal of Marketing Management, 41(9–10), 807–829. https://doi.org/10.1080/0267257X.2024.2421860

DoubleVerify. (2024). 2024 Global Insights Report. https://doubleverify.com/hubfs/46126064/content/DV_Report_GlobalInsights_2024_Global.pdf

Douek, E. Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability'(2021). Columbia Law Review, 121, 759.

Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. PNAS nexus, 3(7), pgae217. https://doi.org/10.1093/pnasnexus/pgae217

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. Information, Communication & Society, 21(5), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

D'Souza, S. (2025). How platform design amplified misinformation in the Southport attack aftermath. https://lexiekirkconnellkawana.substack.com/p/how-platform-design-amplified-misinformation

Ecker, U. K. (2025). Misinformation: Current directions and new insights. Journal of Applied Research in Memory and Cognition, 14(2), 149.

Eder, N. (2024). Making systemic risk assessments work: how the DSA creates a virtuous loop to address the societal harms of content moderation. German Law Journal, 25(7), 1197-1218.

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. Annals of the international communication association, 43(2), 97-116.

Emeric, A., Victor, C. (2024). Interpretable Cross-Platform Coordination Detection on Social Networks. In: Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M. (eds) Complex Networks & Their Applications XII. COMPLEX NETWORKS 2023. Studies in Computational Intelligence, vol 1144. Springer, Cham. https://doi.org/10.1007/978-3-031-53503-1_12

Europäische Kommission. (2020, December 3). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan (COM/2020/790 final). EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

Europäische Kommission. (2025b, December 4). Commission fines X €120 million under the Digital Services Act [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934

European Fact-Checking Standards Network. (2025a, October 1). The real cost of health misinformation and how fact-checkers work to address it. https://efcsn.com/news/2025-10-01_the-real-cost-of-health-misinformation-and-how-fact-checkers-work-to-address-it/

European Fact-Checking Standards Network (EFCSN). (2025b, November 12). EFCSN statement on the European Democracy Shield: How to effectively safeguard European information spaces. https://efcsn.com/news/2025-11-12_efcsn-statement-on-the-european-democracy-shield-how-to-effectively-safeguard-european-information-spaces/

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. Journal of experimental psychology: general, 144(5), 993.

Fazio, L., Rand, D. G., Lewandowsky, S., Susmann, M., Berinsky, A. J., Guess, A. M., … Swire-Thompson, B. (2025). Combating misinformation: A megastudy of nine interventions

designed to reduce the sharing of and belief in false and misleading headlines. https://doi.org/10.31234/osf.io/uyjha

Feghali, K., Najem, R., & Metcalfe, B. D. (2025). Greenwashing in the era of sustainability: A systematic literature review. Corporate Governance and Sustainability Review, 9(1), 18–31. https://doi.org/10.22495/cgsrv9i1p2

Feng, X., Luo, J., Yang, Y., El Baz, D., & Shi, L. (2025). Health Misinformation Detection: Approaches, Challenges and Opportunities. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 62, 00469580251384784.

Ferrara, E. (2024). Charting the Landscape of Nefarious Uses of Generative Artificial Intelligence for Online Election Interference. arXiv. https://doi.org/10.48550/ARXIV.2406.01862

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2022). Black trolls matter: Racial and ideological asymmetries in social media disinformation. Social Science Computer Review, 40(3), 560-578.

The Future of Free Speech. (2025, May 23). Public Consultation Feedback. European Democracy Shield. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14587-European-Democracy-Shield/F3555167_en

Galletta, S., Mazzù, S., Naciti, V., & Paltrinieri, A. (2024). A PRISMA systematic review of greenwashing in the banking industry: A call for action. Research in International Business and Finance, 69, 102262.

Gandhi, A., Hollenbeck, B., & Li, Z. (2025). Misinformation and Mistrust: The Equilibrium Effects of Fake Reviews on Amazon.com. National Bureau of Economic Research. https://doi.org/10.3386/w34161

Gentili, A., Villani, L., Osti, T., Corona, V. F., Gris, A. V., Zaino, A., ... & Cascini, F. (2024). Strategies and bottlenecks to tackle infodemic in public health: a scoping review. Frontiers in Public Health, 12, 1438981.

Gerard, P., Hanley, H. W. A., Luceri, L., & Ferrara, E. (2025). Bridging the Narrative Divide: Cross-Platform Discourse Networks in Fragmented Ecosystems (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.21729

Gielow Jacobs, L. (2022). Freedom of Speech and Regulation of Fake News. The American Journal of Comparative Law, 70(Supplement_1), i278–i311. https://doi.org/10.1093/ajcl/avac010

Gil de Zúñiga, H., & Cheng, Z. (2024). Origin and evolution of the News Finds Me perception: Review of theory and effects. Media Influence on Opinion Change and Democracy: How Private, Public and Social Media Organizations Shape Public Opinion, 151-179.

Global Disinformation Index (GDI). (2019). The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech? https://propaganda-then-and-now.net/wp-content/uploads/2019/09/gdi_ad-tech_report_screen_aw16.pdf

Golebiewski, M., & Boyd, D. (2019). Data voids. https://datasociety.net/library/data-voids/.

Graham, G. (2025, October 23). Elevating first aid information in Canada on YouTube search. Google Blog. https://blog.google/intl/en-ca/products/inside-youtube/elevating-first-aid-information-in-canada-on-youtube-search/

Griffin, R. (2025). The Politics of Risk in the Digital Services Act: A Stakeholder Mapping and Research Agenda. Weizenbaum Journal of the Digital Society, 5(2). https://doi.org/10.34669/wi.wjds/5.2.6

Haßler, J., Magin, M., Russmann, U., Wurst, A.-K., Balaban, D. C., Baranowski, P., Jensen, J. L., Kruschinski, S., Lappas, G., Machado, S., Novotná, M., Marcos-García, S., Petridis, I., Rožukalne, A., Sebestyén, A., & Von Nostitz, F. (2025). Weaponizing Wedge Issues:

Strategies of Populism and Illiberalism in European Election Campaigning on Facebook. Media and Communication, 13. https://doi.org/10.17645/mac.10718

Hastuti, H., Maulana, H. F., Lawelai, H., & Suherman, A. (2025). Algorithmic influence and media legitimacy: a systematic review of social media's impact on news production. Frontiers in Communication, 10, 1667471.

Hawkins, I., & Campbell, S. W. (2025). (Fake) news-finds-me: Interactive social and mobile media uses and incidental news reliance as antecedents of fake news-sharing. Computers in Human Behavior, 168, 108658.

Hayes, A. S., & Ben-Shmuel, A. T. (2024). Under the finfluence: Financial influencers, economic meaning-making and the financialization of digital life. Economy and Society, 53(3), 478–503. https://doi.org/10.1080/03085147.2024.2381980

Herzog, S. M., & Hertwig, R. (2025). Boosting: Empowering citizens with behavioral science. Annual Review of Psychology, 76.

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. Nature Human Behaviour, 8(8), 1545-1553.

Hojati, A., & Nault, B. R. (2025). Content Moderation with Shadowbanning. Information Systems Research. https://pubsonline.informs.org/doi/10.1287/isre.2024.1140

House of Commons Science, Innovation and Technology Committee (SITC) (2025). Social media, misinformation and harmful algorithms (Second Report of Session 2024–25, HC 441). UK Parliament. https://committees.parliament.uk/publications/48745/documents/258221/default/

Huang, G., Jia, W., & Yu, W. (2024). Media literacy interventions improve resilience to misinformation: a meta-analytic investigation of overall effect and moderating factors. Communication Research, 00936502241288103.

Hubeny, T. J., Nahon, L. S., & Gawronski, B. (in press). Understanding partisan bias in judgments of misinformation: Identity protection versus differential knowledge Psychological Science.

Hubeny, T. J., Nahon, L. S., Ng, N. L., & Gawronski, B. (2025). Who Falls for Misinformation and Why?. Personality and Social Psychology Bulletin, 01461672251328800.

Husovec, M. (2024). The Digital Services Act's red line: what the Commission can and cannot do about disinformation. Journal of Media Law, 16(1), 47–56. https://doi.org/10.1080/17577632.2024.2362483

Hwang, E. H., & Lee, S. (2025). A nudge to credible information as a countermeasure to misinformation: Evidence from twitter. Information Systems Research, 36(1), 621-636.

Ibrahim, H., Jang, H. D., Aldahoul, N., Kaufman, A. R., Rahwan, T., & Zaki, Y. (2025). TikTok's recommendations skewed towards Republican content during the 2024 US presidential race. arXiv preprint arXiv:2501.17831.

Institut der Wirtschaftsprüfer in Deutschland e. V. (IDW). (2025). Fake News – Risiken und Handlungsbedarf für Gesellschaft, Unternehmen und Wirtschaftsprüfer (IDW-Positionspapier).

Institute for Strategic Dialogue (ISD). (2024). Social media platforms fall short on enforcing ads policies. https://www.isdglobal.org/digital_dispatches/social-media-platforms-fall-short-on-enforcing-ads-policies/

Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2023). Friction Interventions to Curb the Spread of Misinformation on Social Media (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.11498

Jahn, L., Rendsvig, R.K., Flammini, A. et al. A perspective on friction interventions to curb the spread of misinformation. npj Complex 2, 31 (2025). https://doi.org/10.1038/s44260-025-00051-1

Jakobsen, L., Holden, A. J., Gürcan, Ö., & Özgöbek, Ö. (2025). Agent-Based Exploration of Recommendation Systems in Misinformation Propagation. arXiv preprint arXiv:2507.21724.

Jones, M. O. (2025, March 12). Written evidence submitted by Marc Owen Jones (PhD) (SMH0071) [Written evidence]. UK Parliament. https://committees.parliament.uk/written-evidence/138332/pdf/

Kara, S., Hatipoğlu, S. S., Arslanoğlu, N. Z., & Erdoğan, Z. (2025). The Impact of Trust in Science on COVID-19 Vaccine Attitudes: Parallel Mediation Through Conspiracy Beliefs and General Vaccine Hesitancy. The Eurasian Journal of Medicine, 1. https://doi.org/10.5152/eurasianjmed.2025.251024

Keasey, K., Lambrinoudakis, C., Mascia, D. V., & Zhang, Z. (2025). The impact of social media influencers on the financial market performance of firms. European Financial Management, 31(2), 745-785.

King, C., Phillips, S.C. & Carley, K.M. A path forward on online misinformation mitigation based on current user behavior. Sci Rep 15, 9475 (2025). https://doi.org/10.1038/s41598-025-93100-7

Kington, R. S., Arnesen, S., Chou, W. Y. S., Curry, S. J., Lazer, D., & Villarruel, A. M. (2021). Identifying credible sources of health information in social media: principles and attributes. NAM perspectives, 2021, 10-31478.

Klincewicz, M., Alfano, M., & Fard, A. E. (2025). Slopaganda: The interaction between propaganda and generative AI (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2503.01560

Konrad-Adenauer-Stiftung. (2022). Freedom of expression and press in Latin America. https://www.kas.de/en/web/europaeische-und-internationale-zusammenarbeit/freedom-of-expression-and-press-in-latin-america

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., Panizza, F., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. Nature human behaviour, 8(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Laidlaw, E. B. (2022). Mis- Dis- and Mal-Information and the convoy: An examination of the roles and responsibilities of social media. Public Order Emergency Commission.

Lauerer, C., & Beckert, J. (2024). Pushing boundaries—hybrid advertising in digital news media: a content analysis of media kits. Digital Journalism, 1-20.

Lim, G., & Bradshaw, S. (2023). Chilling legislation: Tracking the impact of “fake news” laws on press freedom internationally. Center for International Media Assistance, 19.

Lin, H., Garro, H., Wernerfelt, N., Shore, J. C., Hughes, A., Deisenroth, D., … Rand, D. G. (2024, February 7). Reducing misinformation sharing at scale using digital accuracy prompt ads. https://doi.org/10.31234/osf.io/u8anb

Lin, Y., Chen, M., Lee, S. Y., Yi, S. H., Chen, Y., Tandoc, E. C., … Salmon, C. T. (2024). Understanding the Effects of News-Finds-Me Perception on Health Knowledge and Information Seeking During Public Health Crises. Health Communication, 39(2), 352–362. https://doi.org/10.1080/10410236.2023.2165750

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nature human behaviour, 7(1), 74-101.

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. D. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. Journal of Medical Internet Research, 25, e49255.

Luceri, L., Salkar, T. V., Balasubramanian, A., Pinto, G., Sun, C., & Ferrara, E. (2025). Coordinated Inauthentic Behavior on TikTok: Challenges and Opportunities for Detection in a Video-First Ecosystem (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2505.10867

Ma, I., Sultan, M., Kozyreva, A., & van den Bos, W. (2025). Understanding the impact of misinformation on adolescents. Nature Human Behaviour. https://doi.org/10.1038/s41562-025-02338-8

Madsen, D. Ø., & Puyt, R. W. (2025). The 7Vs of AI Slop: A Typology of Generative Waste. Available at SSRN 5558018.

Mahapatra, S., Sombatpoonsiri, J., & Ufen, A. (2024). Repression by Legal Means: Governments' Anti-Fake News Lawfare. GIGA Focus Global, 1. https://doi.org/10.57671/GFGL-24012

Mahbub, S., Pardede, E., Kayes, A. S. M., & Rahayu, W. (2019). Controlling astroturfing on the internet: a survey on detection techniques and research challenges. International journal of web and grid services, 15(2), 139-158.

Mannino, M., Garcia, J., Hazim, R., Abouzied, A., & Papotti, P. (2024). Data Void Exploits: Tracking & Mitigation Strategies. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 1627–1637). CIKM '24: The 33rd ACM International Conference on Information and Knowledge Management. ACM. https://doi.org/10.1145/3627673.3679781

Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., & Cresci, S. (2024). Detection and Characterization of Coordinated Online Behavior: A Survey (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2408.01257

Martini, C., Floris, M., Ronzani, P. et al. The impact of interventions against science disinformation in high school students. Sci Rep 15, 34278 (2025). https://doi.org/10.1038/s41598-025-16565-6

Marushchak, A., Petrov, S., & Khoperiya, A. (2025). Countering AI-powered disinformation through national regulation: learning from the case of Ukraine. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1474034

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. New York: Data & Society Research Institute, 359, 1146-1151.

Matamoros-Fernández, A., & Jude, N. (2025). The importance of centering harm in data infrastructures for 'soft moderation': X's Community Notes as a case study. New Media & Society, 27(4), 1986–2011. https://doi.org/10.1177/14614448251314399

Mauk, M., & Grömping, M. (2024). Online disinformation predicts inaccurate beliefs about election fairness among both winners and losers. Comparative Political Studies, 57(6), 965-998.

McGowan, A., MacKenzie, D., & Caliskan, K. (2024). Intermediaries, mediators and digital advertising's tensions. Journal of Cultural Economy, 17(5), 513–531. https://doi.org/10.1080/17530350.2024.2360919

Mede, N. G., Cologna, V., Berger, S., C. Besley, J., Brick, C., Joubert, M., W. Maibach, E., Mihelj, S., Oreskes, N., S. Schäfer, M., van der Linden, S., Abdul Aziz, N. I., Abdulsalam, S., Abu Shamsi, N., Aczel, B., Adinugroho, I., Alabrese, E., Aldoh, A., … Alfano, M. (2025). Public Communication about Science in 68 Countries: Global Evidence on How People Encounter and Engage with Information about Science. Science Communication, 0(0). https://doi.org/10.1177/10755470251376615

Meßmer, A.-K., & Degeling, M. (2023). Auditing Recommender Systems -- Putting the DSA into practice with a risk-scenario-based approach (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2302.04556

Metzler, H., & Garcia, D. (2024). Social Drivers and Algorithmic Mechanisms on Digital Media. Perspectives on Psychological Science, 19(5), 735-748. https://doi.org/10.1177/17456916231185057

Middleton, K.(2025). The Hidden Forces and Harms of the Digital Advertising Ecosystem: Briefing Paper for UK Parliamentary Select Committee Inquiry, Supplementary written evidence by Dr Karen Middleton (SMH0077). Conscious Advertising Network https://committees.parliament.uk/writtenevidence/139719/html/

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. PNAS nexus, 4(3), pgaf062.

Mohamed, F., & Shoufan, A. (2024). Users' experience with health-related content on YouTube: an exploratory study. BMC Public Health, 24(1), 86.

Mosallaei, A., Wang, L., & Ognyanova, K. (2025). From Politics to Entertainment: Exploring "News Finds Me" Perceptions Across News Topics. Social Media+ Society, 11(4), 20563051251382442.

Möller J, Hameleers M, Ferreau F. Typen von Desinformation und Misinformation: Verschiedene Formen von Desinformation und ihre Verbreitung aus kommunikationswissenschaftlicher und rechtswissenschaftlicher Perspektive; 2020. Verfügbar unter: www.die-medienanstal-ten.de/fileadmin/user_upload/die_medienanstalten/Publikationen/Weitere_Veroeffentlichungen/GVK_Gutachten_final_WEB_bf.pdf

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. Journal of the European Economic Association, 19(4), 2131-2167.

Nasser, M., Arshad, N. I., Ali, A., Alhussian, H., Saeed, F., Da'u, A., & Nafea, I. (2025). A systematic review of multimodal fake news detection on social media using deep learning models. Results in Engineering, 26, 104752.

Nenno, S. (2025). Do Community Notes have a party preference?. In Digital Society Blog. Zenodo. https://doi.org/10.5281/zenodo.14899291

Nickl, P. L., Sultan, M., Stinson, C., Stock, F., Hertwig, R., & Kozyreva, A. (2025, August 13). Global Crisis or Overblown Problem? Three Tools to Clarify Contentious Issues in Misinformation Research. https://doi.org/10.31235/osf.io/4vhwq_v1

Norocel, O. C., & Lewandowski, D. (2023). Google, data voids, and the dynamics of the politics of exclusion. Big Data & Society, 10(1), 20539517221149099.

Organization of American States (OAS). (2019, October). Guide to guarantee freedom of expression regarding deliberate disinformation in electoral contexts [Report]. https://www.oas.org/en/iachr/expression/publications/Guia_Desinformacion_VF%20ENG.pdf

OECD (2025), "Mapping policy responses to technology-facilitated gender-based violence in the G7 countries", OECD Public Governance Policy Papers, No. 75, OECD Publishing, Paris, https://doi.org/10.1787/b0887189-en.Anstis, S., & LaFlèche, É. (2025). Gender-based digital transnational repression as a global authoritarian practice. Globalizations, 22(4), 671-688.

Ó Fathaigh, R., Buijs, D., & Hoboken, J. V. (2025). The Regulation of Disinformation Under the Digital Services Act. Media and Communication, 13.

Omilusi, M. (2025). Fake news, election-related disinformation laws, and citizens' rights in African political ecology. Journal of African Elections, 24(1), 1–25. https://doi.org/10.20940/jae/2025/v24i1a1

Orecchia, M. (2025). Engaged, enraged, amplified : the algorithmic logic behind political amplification. European University Institute. https://doi.org/10.2870/9527549

OSCE (2017). United Nations Special Rapporteur on Freedom of Opinion and Expression, Organization for Security and Co-operation in Europe Representative on Freedom of the Media, Organization of American States Special Rapporteur on Freedom of Expression, & African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information. (2017, March 3). Joint declaration on freedom of expression and "fake news", disinformation and propaganda [Joint declaration]. OSCE Representative on Freedom of the Media. https://rfom.osce.org/fom/302796

OSCE Representative on Freedom of the Media. (2019, January 11). OSCE Media Freedom Representative publishes legal review of French laws against manipulation of information [Press release]. https://rfom.osce.org/representative-on-freedom-of-media/408926

Pakina, A. K., Sharma, A., & Kejriwal, D. (2025). AI-Driven Disinformation Campaigns: Detecting Synthetic Propaganda in Encrypted Messaging via Graph Neural Networks. International Journal Science and Technology, 4(1), 12-24.

Palau-Sampio, D. (2023). Pseudo-Media Disinformation Patterns: Polarised Discourse, Clickbait and Twisted Journalistic Mimicry. Journalism Practice, 17(10), 2140–2158. https://doi.org/10.1080/17512786.2022.2126992

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in cognitive sciences, 25(5), 388-402.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management science, 66(11), 4944-4957.

Pennycook, G., Berinsky, A. J., Bhargava, P., Lin, H., Cole, R., Goldberg, B., ... & Rand, D. G. (2024). Inoculation and accuracy prompting increase accuracy discernment in combination but not alone. Nature Human Behaviour, 8(12), 2330-2341.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590-595.

Persakis, A., Nikolopoulos, T., Negkakis, I.C. et al. Greenwashing in marketing: a systematic literature review and bibliometric analysis. Int Rev Public Nonprofit Mark 22, 957–992 (2025). https://doi.org/10.1007/s12208-025-00452-x

Phillips, W. (2018). The oxygen of amplification: Better practices for reporting on extremists, antagonists, and manipulators online. Data et Society Research Institute.

Pournaki, A., Gaisbauer, F., & Olbrich, E. (2025). How Influencers and Multipliers Drive Polarization and Issue Alignment on Twitter/X. Proceedings of the International AAAI Conference on Web and Social Media, 19(1), 1599-1615. https://doi.org/10.1609/icwsm.v19i1.35890

Putra, B. A. (2024). Fake news and disinformation in Southeast Asia: how should ASEAN respond?. Frontiers in Communication, 9, 1380944.

Pérez Argüello, M. F., & Barojan, D. (2019). Mexico. In L. Bandeira, D. Barojan, R. Braga, J. L. Peñarredonda, & M. F. Pérez Argüello (Eds.), Disinformation in democracies: Strengthening digital resilience in Latin America (pp. 20–29). Atlantic Council. https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/

Radsch, Courtney, AI and Disinformation: State-Aligned Information Operations and the Distortion of the Public Sphere (July 12, 2022). OSCE Representative on Freedom of the Media, Organization for Security and Co-operation in Europe, July 2022, Available at SSRN: https://ssrn.com/abstract=4192038

Renault, T., Mosleh, M., & Rand, D. G. (2025). Republicans are flagged more often than Democrats for sharing misinformation on X's Community Notes. Proceedings of the National Academy of Sciences, 122(25), e2502053122.

Richardson, J. E., Giraud, E. H., Poole, E., & de Quincey, E. (2024). 'Hypocrite!'Affective and argumentative engagement on Twitter, following the Christchurch terrorist attack. Media, Culture & Society, 46(6), 1105-1123.

Rodrigues, F., Newell, R., Babu, G. R., Chatterjee, T., Sandhu, N. K., & Gupta, L. (2024). The social media Infodemic of health-related misinformation and technical solutions. Health Policy and Technology, 13(2), 100846.

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. Frontiers in Artificial Intelligence, 8, 1569115.

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. Journal of risk research, 22(5), 570-580.

Roozenbeek, T., van den Berg, C., Lambooij, M.S. et al. Trust in institutions and misinformation susceptibility both independently explain vaccine skepticism. Sci Rep 15, 37655 (2025). https://doi.org/10.1038/s41598-025-21452-1

Rovetta, A., & Bhagavathula, A. S. (2020). Global infodemiology of COVID-19: analysis of Google web searches and Instagram hashtags. Journal of medical Internet research, 22(8), e20673.

Sato, Y., & Wiebrecht, F. (2024). Disinformation and Regime Survival. Political Research Quarterly, 77(3), 1010-1025. https://doi.org/10.1177/10659129241252811

Scholtens, M., Pizano, P., Karpawich, M., & Kuckes, G. (2024). The disinformation economy. The Carter Center & McCain Institute for International Leadership. https://www.cartercenter.org/wp-content/uploads/2024/05/the-disinformation-economy-mccain-may-2024.pdf

Sekwenz, M.-T., Wagner, B., & De Bruijn, H. (2025). From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.01787

Shin, D., Hameleers, M., Park, Y. J., Kim, J. N., Trielli, D., Diakopoulos, N., Helberger, N., Lewis, S. C., Westlund, O., & Baumann, S. (2022). Countering Algorithmic Bias and Disinformation and Effectively Harnessing the Power of AI in Media. Journalism & Mass Communication Quarterly, 99(4), 887-907. https://doi.org/10.1177/10776990221129245

Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). Community notes reduce engagement with and diffusion of false information online. Proceedings of the National Academy of Sciences, 122(38), e2503413122.

Snyder, T. (2017). On tyranny: Twenty lessons from the twentieth century. New York: Tim Duggan Books.

Soliman, W., & Rinta-Kahila, T. (2024). Unethical but not illegal! A critical look at two-sided disinformation platforms: Justifications, critique, and a way forward. Journal of Information Technology, 39(3), 441-476.

Stark, B., Magin, M., & Jürgens, P. (2021). Maßlos überschätzt. Ein Überblick über theoretische Annahmen und empirische Befunde zu Filterblasen und Echokammern. Digitaler Strukturwandel der Öffentlichkeit: Historische Verortung, Modelle und Konsequenzen, 303-321.

Strowel, A., & De Meyere, J. (2023). The Digital Services Act: transparency as an efficient tool to curb the spread of disinformation on online platforms?. J. Intell. Prop. Info. Tech. & Elec. Com. L., 14, 66.

Tameez, H. (2020, September 25). Following successful experiments, Twitter will prompt all users to read the articles they're about to retweet. Nieman Journalism Lab.

Tardelli, S., Nizzoli, L., Avvenuti, M., Cresci, S., & Tesconi, M. (2024). Multifaceted online coordinated behavior in the 2020 US presidential election. EPJ Data Science, 13(1), 33.

TechCrunch. (2020, April 27). WhatsApp's new limit cuts virality of "highly forwarded" messages by 70%. TechCrunch.

The Guardian. (2024, 26. Oktober). 'Anticipatory obedience': newspapers' refusal to endorse shines light on billionaire owners' motives. https://www.theguardian.com/us-news/2024/oct/26/anticipatory-obedience-newspapers-endorsement-refusal

Tian, Y., & Willnat, L. (2025). From news disengagement to fake news engagement: Examining the role of news-finds-me perceptions in vulnerability to fake news through third-person perception. Computers in Human Behavior, 162, 108431.

Tjaden, J., Wolfgram, J., Philipp, A., Weißmann, S., Bobzien, L., Kohler, U., & Verwiebe, R. (2025). Does the TikTok feed lean right? Exposure to Political Party Content among non-partisan users during regional and federal elections in Germany. Center for Open Science. https://doi.org/10.31235/osf.io/7vdex_v1

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. Proceedings of the National Academy of Sciences, 119(42). https://doi.org/10.1073/pnas.2207159119

Udry, J., & Barber, S. J. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. Current Opinion in Psychology, 56, Article 101736. https://doi.org/10.1016/j.copsyc.2023.101736

UNESCO. (2023). Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multi-stakeholder approach.

Unger, S., Klapproth, J., Boberg, S., Bösch, M., Vief, N., Stöcker, C., & Quandt, T. (2025). Features of disinformation: an expert interview study on the perception of disinformation among political, governmental, media and business elites in Germany. Journal of Elections, Public Opinion and Parties, 35(3), 472–494. https://doi.org/10.1080/17457289.2025.2514199

van de Kerkhof, J. (2025). Article 22 Digital Services Act: Building trust with trusted flaggers. Internet Policy Review, 14(1). https://doi.org/10.14763/2025.1.1828

van der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. Nat Med 28, 460–467 (2022). https://doi.org/10.1038/s41591-022-01713-6

van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2025). Using psychological science to understand and fight health misinformation: An APA consensus statement. American Psychologist. https://doi.org/10.1037/amp0001598

van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2025). Using psychological science to understand and fight health misinformation: An APA consensus statement.American Psychologist. Advance online publication. https://doi.org/10.1037/amp0001598

van Rooij, I. (2025) AI slop and the destruction of knowledge. https://doi.org/10.5281/zenodo.16905559

Vellani, V., Zheng, S., Ercelik, D., & Sharot, T. (2023). The illusory truth effect leads to the spread of misinformation. Cognition, 236, 105421.

Venkataramakrishnan, S. (2025). From Disinformation to Violence. Counter Terrorist Trends and Analyses, 17(2), 8-15.

Verdolotti, E., Luceri, L., & Giordano, S. (2025). Predicting, evaluating, and explaining top misinformation spreaders via archetypal user behavior. Online Social Networks and Media, 50, 100336.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Votta, F., Kruschinski, S., Hove, M., Helberger, N., Dobber, T., & de Vreese, C. (2024). Who Does(n't) Target You? Mapping the Worldwide Usage of Online Political Microtargeting. Journal of Quantitative Description: Digital Media, 4. https://doi.org/10.51685/jqd.2024.010

Wang, J., Zhai, Y., & Shahzad, F. (2025). Mapping the terrain of social media misinformation: A scientometric exploration of global research. Acta Psychologica, 252, 104691. https://doi.org/10.1016/j.actpsy.2025.104691

Wang, S. Y. N., Phillips, S. C., Carley, K. M., Lin, H., & Pennycook, G. (2025). Limited effectiveness of psychological inoculation against misinformation in a social media feed. PNAS nexus, 4(6), pgaf172. https://doi.org/10.1093/pnasnexus/pgaf172

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Watolla, A., Zerrer, P., Rau, J., Merten, L., Kettemann, M.C., & Puschmann, C. (2025). Gesellschaftliche Auswirkungen systemischer Risiken. Demokratische Prozesse im Kontext von Desinformationen. Bundesnetzagentur. https://www.dsc.bund.de/DSC/DE/Aktuelles/studien/Auswirkungen%20Systemischer%20Risiken.pdf?__blob=publicationFile&v=3

Wegner, Susanne. (2024). Angstmache & Feindbilder: Wie Desinformation den Wahlkampf 2024 prägt und was die Plattformen dagegen tun. German-Austrian Digital Media Observatory [GADMO].

Windwehr, S. (2025, February 18). Trump vs. Europe: The role of the Digital Services Act. Heinrich Böll Stiftung. https://eu.boell.org/en/2025/02/18/trump-vs-europe-role-digital-services-act

Wirtschafter, V., & Majumder, S. (2023). Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter's Community Notes. Journal of Online Trust and Safety, 2(1), 1–11. https://doi.org/10.54501/jots.v2i1.139

World Economic Forum. (2025, January 15). The Global Risks Report 2025: 20th edition. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf

Woolley, S. C., & Howard, P. N. (Eds.). (2018). Computational propaganda: Political parties, politicians, and political manipulation on social media. Oxford University Press.

Ye, J., Luceri, L., & Ferrara, E. (2025, June). Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X During the 2024 US Presidential Election. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2349-2362).

Yin, J., Jia, H., Zhou, B., Tang, T., Ying, L., Ye, S., Peng, T.-Q., & Wu, Y. (2025). Blowing Seeds Across Gardens: Visualizing Implicit Propagation of Cross-Platform Social Media Posts. IEEE Transactions on Visualization and Computer Graphics, 31(1), 185–195. https://doi.org/10.1109/tvcg.2024.3456181

YouTube. (2025). Gesundheitsinformationen auf YouTube. YouTube Help. https://support.google.com/youtube/answer/9795167?hl=de

Yun, J. H., An, J., & Platt, M. L. (2025). The Impact of Repeated Financial Misinformation on Investments. Elsevier BV. https://doi.org/10.2139/ssrn.5187289

Zarocostas, J. (2020). How to fight an infodemic. The Lancet, 395(10225), 676.

Zhang, L., & Jiang, S. (2024). “I Know News Will Find Me”: Examining the Relationship Between the “News-Finds-Me” Perception and COVID-19 Misperceptions. Health Communication, 39(13), 3032-3043.

Zhao, C., Wei, L., Qin, Z., Zhou, W., Song, Y., & Hu, S. (2025). MPPFND: A Dataset and Analysis of Detecting Fake News with Multi-Platform Propagation (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.15834

Zimmermann, F., & Kohring, M. (2018). „Fake News “als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs. M&K Medien & Kommunikationswissenschaft, 66(4), 526-541.sowie u.a. https://hateaid.org/fake-news/

Zuiderveen Borgesius, F.J., Trilling, D., Möller, J., Bodó, B., de Vreese, C.H., & Helberger, N. (2016). Should we worry about filter bubbles? Internet Policy Review, 5(1). https://doi.org/10.14763/2016.1.401

ISSN 1865-8997