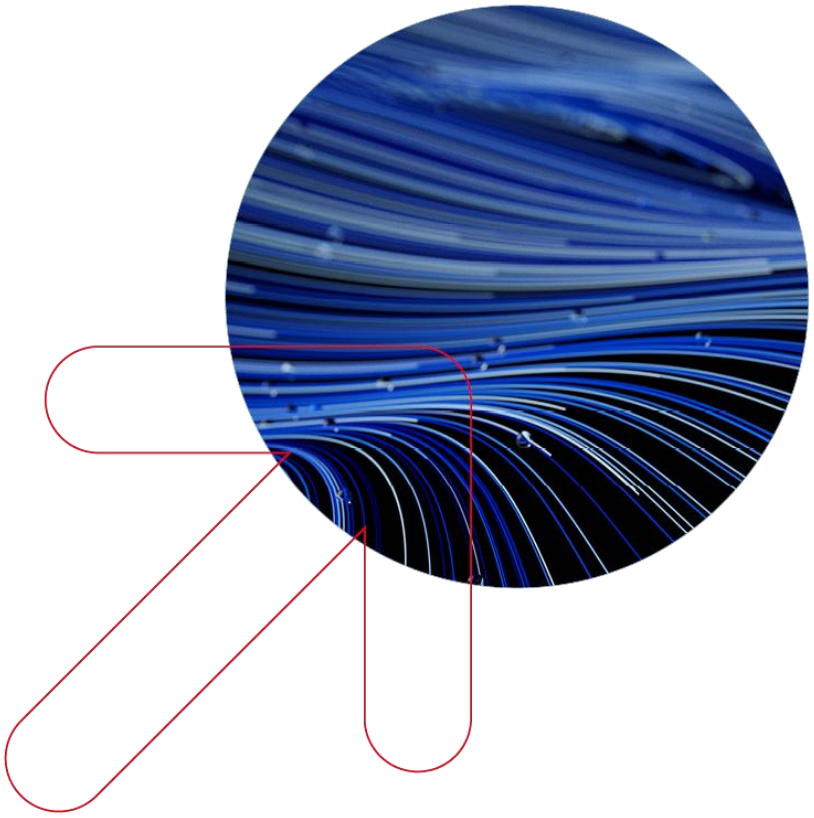
WIK • Discussion paper

No. 546



The spread and impact of disinformation in the context of DSA & online platforms

Authors:
Dr. Nico Steffen
Peter Kroon
Dr. Lukas Wiewiorra

Bad Honnef, December 2025

WIK
Wissenschaftliches Institut
für Infrastruktur und
Kommunikationsdienste

Imprint

WIK Research Institute for
Infrastructure and Communication Services GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Germany
Tel.: +49 2224 9225-0
Fax +49 2224 9225-63
Email info@wik.org
www.wik.org

Authorized representatives and signatories

Management	Dr. Cara Schwarz-Schilling (Chairwoman of the Management Board, Director)
	Alex Kalevi Dieke (Commercial Managing Director)
Authorized signatories	Prof. Dr. Bernd Sörries
	Dr. Christian Wernick
	Dr. Lukas Wiewiorra
Chairman of the Supervisory Board	Dr. Thomas Solbach
Commercial Register	Siegburg Local Court, HRB 7225
Tax	222/5751/0722
Sales tax identification no.	DE 123 383 795

As of: January 2025

ISSN 1865-8997

Image credits Title: © Robert Kneschke - stock.adobe.com

Further contributions to the discussion series can be found here:
https://www.wik.org/veroeffentlichungen/diskussionsbeitraege

The discussion papers published by WIK contain essays and lectures by institute staff as well as selected interim and final reports on completed research projects. By publishing this series, WIK aims to provide information about its activities, stimulate discussion, and receive input from outside sources. Criticism and comments are therefore welcome at any time. The views expressed in the various contributions reflect solely the opinions of the respective authors.

Table of contents

Executive Summary

This discussion paper presents a comprehensive synthesis of the contemporary disinformation ecosystem, drawing upon empirical literature and analyses from 2024 and 2025. It challenges the traditional, content-centric paradigm that seeks to address information disorders primarily through verification and deletion. Instead, the study conceptualizes disinformation as a structural failure of the digital attention economy. By moving beyond a binary classification of truth versus falsehood, the analysis adopts a continuum-based framework. This approach captures the prevalence of e.g. "lawful but awful" content, which includes technically accurate but contextually misleading material that evades standard moderation policies yet generates significant societal harm. The central thesis posits that the proliferation of disinformation is not an accidental anomaly but a profitable negative externality of current platform business models and the opaque programmatic advertising infrastructure.

A critical contribution of this study is the detailed dissection of the economic incentives driving the spread of harmful content. The analysis identifies the programmatic advertising supply chain ("Ad-Tech") as a primary, albeit often inadvertent, financier of disinformation. Automated bidding processes frequently decouple advertising revenue from editorial quality. This mechanism funnels brand budgets toward "Made-for-Advertising" (MFA) websites and clickbait portals that optimize for engagement rather than veracity. Marketing strategies that treat these outcomes as mere market inefficiencies or external failures obscure the causal reality wherein advertising decisions provide essential financial capital for the degradation of the information space. Consequently, the study describes a professionalized "Disinformation-as-a-Service" (DaaS) industry. In this sector, actors commodify manipulation tactics, ranging from coordinated inauthentic behavior (CIB) to the deployment of mercenary bot networks, independent of ideological conviction.

The study further elucidates the interplay between algorithmic architecture and user psychology. Engagement-based ranking algorithms function as non-neutral gatekeepers that systematically privilege polarizing and emotionalized content to maximize time-on-device. Empirical audits reveal that these systems often reward negative campaigning and moral outrage with visibility bonuses, creating a feedback loop that incentivizes political actors to adopt more aggressive communication styles.

This technological environment interacts with specific cognitive vulnerabilities. The analysis highlights the "News-Finds-Me" perception, a passive information consumption mindset that correlates negatively with political knowledge and critical verification behavior. Furthermore, the emergence of Generative AI has exacerbated these dynamics by lowering production costs to near zero. Aside from increasingly realistic high-quality fakes,

this has also resulted in a flood of "AI Slop" or mass-produced, low-quality synthetic content that also increases verification costs and generates "epistemic pollution" within the information ecosystem.

In evaluating countermeasures, the paper offers a critical assessment of current intervention strategies. While user-centric approaches such as prebunking and accuracy nudges demonstrate efficacy in controlled settings, they face significant scalability challenges and behavioral gaps in real-world application. The study also scrutinizes crowdsourced moderation models like "Community Notes." While effective at reducing the diffusion of specific false claims, these systems suffer from "bridging-based ranking" limitations. The requirement for cross-ideological consensus often leads to the under-flagging of politically controversial disinformation, leaving high-risk narratives unaddressed.

On a regulatory level, the study outlines a sharp divergence between the European Union and other global jurisdictions. It contrasts the EU's Digital Services Act (DSA), which establishes a process-oriented framework focused on risk assessment and transparency, with the repressive "fake news laws" prevalent in parts of Asia and Africa. The analysis argues that while the EU model aims to mitigate systemic risks through structural accountability and data access for researchers ("trace research"), content-centric laws in other regions are frequently weaponized for "lawfare" to suppress political opposition and civil society.

The study concludes that resilient governance of the digital public sphere requires a shift from policing individual pieces of content to addressing the underlying economic and technical structures. Effective policy must disrupt the financial infrastructure of the disinformation economy by imposing stricter oversight on programmatic advertising allocation to prevent the funding of harmful content. Additionally, it calls for the prohibition of specific manipulative dissemination techniques, such as mass automation, under the DSA framework. Ultimately, the paper argues that safeguarding democratic integrity requires legally mandated access to platform data. Only through independent audits and quantitative "trace research" can the causal mechanisms of algorithmic amplification be fully understood and mitigated.

1 Introduction

The analysis of the complex dynamics of disinformation and online harm is based on the observation that this phenomenon has far-reaching social, economic, and security policy relevance. The considerations extend beyond purely political content and also include health issues and commercial areas such as investment scams and fraud. The debate on how to effectively combat online harm should go beyond mere content moderation and address user behavior and systemic factors, as harm can occur even without explicit violations of content standards. In this context, it is important to distinguish between general misinformation and coordinated influence operations, which are often state-sponsored activities.

The aim of this study is to synthesize the current state of knowledge regarding the spread, detection, and combating of disinformation and, based on this, to derive prioritized, systemic measures. Scientific examination of the topic first requires a clear definition of terms, with research traditionally differentiating between truthfulness and intent to cause harm. In addition to misinformation (false, without intent) and disinformation (false, with intent), the analysis also includes malinformation, i.e., the deliberately harmful dissemination of genuine information, often through decontextualization. For analytical purposes, the study adopts a broader, pragmatic working definition, as the focus is on the systemic risks of dissemination and should not depend rigidly on intentionality or the exact degree of falsity. A key analytical hurdle in this field is the limited empirical knowledge on causality and user exposure to harmful content.

To address these systemic challenges, the study focuses on three key research questions. First, it examines which economic incentive structures in the platform economy, particularly in the opaque ad tech ecosystem, fund and promote the spread of disinformation. Second, the study analyzes how current technological developments (such as AI slop) and political trends (such as the move away from traditional fact-checking approaches) are changing the economic and technical dynamics of disinformation. Third, it determines which systemic intervention approaches (actor-based versus content-based) are empirically most effective and how current regulatory measures can be classified in an international comparison. The latter point involves contrasting the EU's process-oriented regulatory model (DSA) with content-based, often repressive "fake news" laws in other regions.

Methodologically, the work is based on comprehensive desk research, which includes scientific literature, gray literature, and ad hoc analyses from public or civil society institutions. AI-based applications[1] were used as methodological tools in the preparation of this work. Their use was limited to linguistic revision, the structuring of arguments, and the textual elaboration of the author's own ideas. The human authors bear full

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

responsibility for the content, the selection of sources, the verification of facts, and the conclusions. No AI-generated content was included in the text without human review and verification. All sources cited were researched independently. The use of AI served solely to increase efficiency in the writing and editing process, not to generate primary research data or original ideas.

Structurally, the study is divided into the following main chapters: After this introduction, the basics (Chapter 2) of conceptual understanding, typologies, mechanisms (including dissemination in the hybrid media system), and risks are presented. This is followed by an analysis of online platforms in the disinformation economy (Chapter 3), which highlights economic incentive structures and algorithmic amplification. Chapter 4 deals with the analysis and evaluation of current countermeasures and the international regulatory framework. The thesis concludes with a summary (Chapter 5), which brings together the key findings and derives recommendations for action.

2 Fundamentals

2.1 Understanding and definition of terms

The scientific and public debate on disinformation continues to face the key challenge of a lack of a uniform definition of the term. The term is often interpreted differently by researchers, regulatory authorities, and platform operators, with different emphases having become established depending on the discipline. In the political and regulatory debate, particularly at the European Union level, "disinformation" functions as the central guiding concept, defined primarily by the criteria of intentionality and potential harm. The EU Commission defines disinformation as demonstrably false, but also misleading, information that is created, presented, and disseminated with the intention of achieving economic gain or deliberately deceiving and harming the public (European Commission, 2020). This definition emphasizes intentionality and potential harm as decisive criteria for regulatory intervention.

In contrast, the term misinformation often dominates psychological and communication science research as an overarching concept, although both variants continue to be found as umbrella terms in the broader academic literature (cf. Wang et al., 2025). Recent studies in particular often use "misinformation" as a collective term for all forms of false or inaccurate information, regardless of the sender's intention. The reason for this lies in the focus of research: both for cognitive processing and psychological effects on the recipient, as well as for many of the potential harms, it is often secondary whether false information was spread intentionally as disinformation or unintentionally as misinformation (Ecker et al., 2025; APA, 2025).

To address this terminological ambiguity, other academic works draw on the differentiated framework of "information disorder" developed by Wardle and Derakhshan (2017), which distinguishes precisely between misinformation (unintentional), disinformation (intentional), and malinformation (harmful but true) in order to capture the specific dynamics in the digital space more accurately.

1. Misinformation: Refers to demonstrably false or misleading information that is disseminated *without* deliberate intent to cause harm (e.g., through unintentional sharing).
2. Disinformation: Refers to demonstrably false or misleading information that is disseminated *with* the intention of causing harm or pursuing political or economic goals.
3. Malinformation: Refers to the *deliberately* harmful dissemination of *genuine* information, often by publishing it out of context (e.g., leaks, doxing).

For an overview in the regulatory context of the DSA, Watolla et al. (2025) identified a working definition of disinformation that encompasses four characteristics: (1)

demonstrably false or misleading, (2) potential for social harm, (3) targeted dissemination by actors, and (4) pursuit of political or economic purposes. In analytical practice, strict distinctions in various dimensions are proving increasingly challenging. As the American Psychological Association notes, the intention that distinguishes misinformation from disinformation is often difficult to prove in reality, as the internal motives of an actor can hardly be validly verified externally (APA, 2023). A recent analysis by Nickl et al. (2025) therefore suggests that, in order to navigate this conceptual landscape, information disruptions should not be placed in rigid categories, but rather consistently located on a continuum along the three axes of truth, intention, and harm.

The dimension of truth in particular proves to be a complex gray area. While classic definitions often only explicitly cover false content (fact check: "false"), empirical evidence shows that technically correct but misleading information often has a much greater reach and damaging effect. Allen et al. (2024) show that even misleadingly framed headlines, such as when a death is linked to the timing of a vaccination, received six times as many views on platforms such as Facebook as content that was explicitly marked as false. An often-cited example is the headline: "A 'healthy' doctor died two weeks after receiving a COVID-19 vaccination; CDC investigates why" (Benton, 2021). Although the statement was factually correct, the framing created the impression of a direct connection, resulting in significantly increased viral spread (APA, 2023). Van der Linden and Kyrychenko (2024) therefore argue that too strict a restriction to verifiable falsehoods would obscure the majority of systemic risks, as actors deliberately exploit gray areas to shape narratives without technically telling untruths. The epistemological challenge is therefore to understand disinformation as a spectrum ranging from unintentional errors to strategic influence operations (Nickl et al., 2025).

The term "fake news" is deliberately used sparingly in this paper. Although there is a growing consensus in academia to avoid the term because it has become unspecific and highly politicized, often used to delegitimize critical reporting (APA, 2023), it continues to be used prominently in both public debate and parts of academic research, often misleadingly as an umbrella term for a wide range of disinformation phenomena. Therefore, it needs to continue to be taken into account especially in research or bibliographic works (cf. Wang et al., 2025). Analytically, the term retains its relevance in its narrowly used sense, i.e., as fabricated information that mimics news media content in its form but not in its organizational process or intent (Lazer et al., 2018). This pseudo-journalistic character and a focus on topicality (in German literature sometimes described as "aktuelle Desinformation" (engl. topical disinformation, cf. Zimmermann & Kohring, 2018) distinguishes "fake news" from other types of misinformation (e.g., memes or conspiracy theories).

Against this theoretical background, the present study deviates from the strict academic distinction and adopts a broader, more pragmatic working definition of disinformation as an umbrella term. Since this study examines the systemic risks of dissemination via social media and other digital platforms, including the role of algorithmic amplification, it makes

sense to use a working term that does not rigidly adhere to the criteria of individual intentionality or absolute degree of falsity, which are difficult to verify (cf. also Watolla et al., 2025). This approach includes, for example, cases in which content can have harmful effects despite being factually correct due to misleading contextualization or framing. In order to maintain analytical precision for those cases in which clear intent to cause harm and coordinated dissemination can be proven, the more precise terms "targeted disinformation" or "organized disinformation" will be used for such phenomena in the further course of this study.

In order to organize this conceptual complexity and classify the terms used in this study in the field of information disruption, the following overview table (Table 1) systematizes the relationships between truthfulness, intent, and potential harm. The actors, dissemination mechanisms, and risks are addressed further in the following subchapters.

Table 1 : Overview – Dimensions of disinformation

	Decontextualiztion	Misinformation	Manipulative (political) advertising	Pseudo-journalism	Propaganda
Degree of falsehood	low	high	varying	rather low	high
Typical Intention	Spreading manipulative narratives that support a political ideology; economic (clickbait)	economic; ideological (de-)mobilization	political mobilization	economic; ideological (de-)mobilization	Geopolitical and ideological (de-)mobilization
Typical broadcasters	Political actors, media, alternative media	Internet fraudsters, conspiracy theorists, alternative media	Political actors, NGOs	Internet fraudsters, alternative media makers	National governments and (international) organizations
Typical Distribution	Widespread distribution: appears in all media, often found in alternative media, further distribution by users	Limited dissemination: often via social media, sometimes supported by coordinated inauthentic behavior	Paid dissemination: frequently on social media, but also via other campaign channels	Limited distribution: in own online media or via social networks, further distribution by users	Professional dissemination: via all communication channels, including own media organizations and with the help of coordinated inauthentic behavior
Typical Risks for the individual	Cognitive, emotional, (political) misjudgments (to varying degrees)				
Typical risks for society	misinformed electorate, polarizing	misinformed electorate, divisive, threatening to democracy	polarizing	misinformed electorate, polarizing, divisive	geopolitical, divisive, threatening to democracy

Source: Möller et al. (2020, p. 38)

2.2 Typology: Actors and motives

In order to understand the systemic risks of disinformation, a differentiated approach is necessary that goes beyond the mere truthfulness of the information. The landscape of

disinformation is characterized by a heterogeneity of actors and their motivations, ranging from geopolitical destabilization to pure profit-seeking. Based on the analysis by Möller et al. (2020) and the classifications of the European Commission (2018), the field can be systematized according to groups of actors and thematic manifestations.

The originators and disseminators of disinformation can be usefully differentiated according to their primary motivation, even if numerous hybrid forms occur in practice. The European Commission's High Level Expert Group on Fake News and Online Disinformation, led by de Cock Buning (2018), explicitly distinguishes between state actors, political non-state actors, profit-oriented actors, and citizens, emphasizing that disinformation aims either at political/social harm or financial gain. More recent studies also emphasize that an analysis without this differentiation remains incomplete, as countermeasures must vary depending on the type of actor (Unger et al., 2025; CPA, 2023).

A central group is formed by state and geopolitical actors who use disinformation as an instrument of power projection both internally and externally. Bradshaw and Howard (2018, 2019) use a global inventory of organized social media manipulation to show that governments and security apparatuses in numerous countries systematically use "computational propaganda" to influence elections, delegitimize opposition figures, and enforce foreign policy goals. In authoritarian contexts, disinformation often serves primarily to secure power and ensure the survival of the regime (Sato and Wiebrecht, 2024). In foreign policy, these strategies aim to distort the public sphere of other states, erode trust in institutions, and thus weaken democratic capacity to act (Radsch, 2022; Bayer et al., 2019). The methods range from classic propaganda in state-affiliated media to digitally scaled influence. Bradshaw and Howard (2019) document the use of bots, sock puppet accounts, and troll armies as core elements. At the platform level , these practices are often summarized as "coordinated inauthentic behavior" (CIB). This refers to the coordinated actions of networks that conceal their true identity (Cinelli et al., 2022; Di Marco et al., 2025). In addition, these actors deliberately exploit existing social fault lines ("wedge issues") such as migration or identity politics to deepen polarization and block political processes (Freelon et al., 2022; Haßler et al., 2025).

In contrast, there are ideologically motivated actors, including parties, "alternative media," activist groups, and conspiracy ideologues. Their main goal is political mobilization, agenda setting, and polarization within their own country. This group is characterized by the method of decontextualization: true or semi-true information is taken out of its original context to support a specific ideological narrative (Möller et al., 2020).

A third group consists of actors whose actions are primarily driven by the maximization of financial gain. The spectrum ranges from criminal fraudsters (scammers) to operators of "clickbait" portals. This group is structurally closely linked to the advertising industry (ad tech). Here, disinformation is often not produced out of conviction, but as a means to an end: it serves to attract attention, which is automatically monetized via the programmatic advertising system (GDI, 2019). Algorithmic preferences for emotional content are

deliberately exploited to generate traffic and advertising revenue. (The mechanisms of this advertising ecosystem are analyzed in more detail in Chapter 3.1.)

Parallel to the actors, three main thematic areas can be identified that pose specific risks. In political disinformation, the focus is on manipulating public opinion. This includes attacks on electoral integrity and forms of "computational propaganda," which describes the use of automation to manipulate public opinion (Woolley & Howard, 2018). A key tool is manipulative political advertising, which often serves to mobilize people regardless of the truthfulness of the message. In its most extreme form, it takes the form of propaganda to achieve ideological goals (Bennett & Livingston, 2018).

In the health sector, the WHO coined the term "infodemic" to describe a flood of information that may be accurate, false, or misleading, especially in digital and physical environments during disease outbreaks. This flood of information makes it difficult for people to identify reliable sources, promotes risky health behaviors, and can undermine trust in health authorities (Zarocostas, 2020). A systematic review for the WHO shows that infodemics are closely linked to vaccine hesitancy, the use of ineffective or harmful treatments, and disregard for evidence-based recommendations (e.g., wearing masks, social distancing) (Borges do Nascimento et al., 2022). Gentili et al. (2024) also emphasize that infodemics are now considered a distinct public health threat in their own right ( ), as they systematically distort both individual decisions and collective crisis responses.

Empirical studies show that exposure to targeted health misinformation has measurable effects on attitudes and behavior: Allen et al. (2024) use Facebook data and experiments to show that even limited exposure to misinformation about vaccines and vaccine-skeptical content can reduce vaccination intentions and stabilize false risk perceptions. At the same time, several studies show that trust in science, medical personnel, and institutions is a key buffer against such effects: higher trust in science and health institutions is systematically associated with greater willingness to vaccinate and lower susceptibility to health misinformation (Kara et al., 2025; Roozenbeek et al., 2025).

On a psychological level, the "illusory truth effect" often comes into play here: repeated statements are perceived as more credible, even if they are objectively false. Classic studies and a meta-analysis show that mere repetition increases the subjective truthfulness of statements. This effect also occurs when people actually possess the relevant factual knowledge (Dechêne et al., 2010; Fazio et al., 2015). Recent reviews confirm that the Illusory Truth Effect is a robust mechanism that manifests itself across various domains, from everyday knowledge to political and health misinformation (Udry & Barber, 2024). Vellani et al. (2023) were able to experimentally show that even a single repetition of false information increases the willingness to share this content on social networks; this effect is mediated by an increased perceived accuracy. Comparable effects have now also been described for financial information: Repeated misleading stock market news can lead to increased risk-taking and greater allocation to risky investments, especially among confident investors (Yun et al., 2025).

Often underestimated but highly relevant economically, this also applies to the broader area of commercial disinformation. This includes scams, fake product reviews, misleading financial recommendations ("fin influencers," cf. Hayes and Ben-Shmuel, 2024; Keasey et al., 2025) and greenwashing. Economic analyses of the platform ecosystem show that fake online reviews not only deceive individual consumers, but also distort market structures: Gandhi et al. (2025) demonstrate for Amazon that the systematic use of fake product reviews artificially inflates ratings, diverts demand from honest to opportunistic suppliers, and reduces overall consumer welfare. Clear effects can also be demonstrated in the financial markets: Arcuri et al. (2023) show in an event study for US and EU stock exchanges that negative fake news about companies causes significant short-term price losses, while positive fake news does not generate comparable, stable price gains. Further studies on fake news in economic contexts suggest that disinformation can even increase macroeconomic uncertainty and amplify economic cycles (Assenza et al., 2024).

In the field of sustainability, commercial disinformation is discussed in particular in relation to greenwashing. Systematic reviews show that exaggerated or selective environmental promises ("green framing") erode trust in brands and financial institutions and create long-term reputational and regulatory risks (Galletta et al., 2024; Persakis et al., 2025; Feghali et al., 2025). A recent review of greenwashing and branding concludes that exposed greenwashing practices can significantly damage brand reputation and customer loyalty in particular (cf. AlQahtani, 2025).

Commercial actors often use pseudo-journalism formats to disseminate such content in order to simulate credibility. Fake news and disinformation-related content systematically imitate the form of traditional news sites. They adopt the layout, headlines, and supposedly independent editorial texts, but are controlled by political or economic interests (Egelhofer & Lecheler, 2019). In this context, Palau-Sampio (2023) describes "pseudo-media" that work with conspiracy theories and distorted representations of social reality and contribute to a stable disinformation ecosystem with the help of clickbait and polarizing language. Studies on "hybrid advertising" and native advertising in digital news media show that such formats deliberately blur the line between editorial content and advertising, thereby undermining the perceived credibility of journalistic offerings (Lauerer & Beckert, 2024; Di Domenico et al., 2021).

Taken together, these observations reinforce the assessment that commercial disinformation poses a concrete economic risk to companies. The spectrum ranges from immediate losses in share price and sales to long-term damage to reputation and increased regulatory and liability exposure (IDW, 2025; see also Arcuri et al., 2023; Gandhi et al., 2025).

Combining actors, motives, and target groups results in a classification grid (see Table 2). This grid illustrates that blanket countermeasures are rarely effective. While the withdrawal of advertising revenue ("demonetization") can be effective against economically motivated actors (GDI, 2019), ideologically motivated actors primarily require discursive

or educational policy interventions. In practice, there are also fluid transitions: states make use of commercial infrastructures, ideological actors monetize their content through advertising, and economic actors amplify polarized political material because it performs well.

Table 2 : Overview of actors and motives

Category	Type A: State (foreign actors)	Type B: Economic	Type C: Ideological/political (domestic actors)
Actors	Intelligence services, governments, state media	Fraudsters, clickbait operators, PR agencies	Political parties, activists, extremists, alternative media
Primary motive	Geopolitics / power (influence, destabilization)	Finance (advertising revenue, fraud, stock prices)	Ideology/conviction (mobilization, opinion-making)
Typical method	Propaganda, computational propaganda (e.g., bot networks)	Pseudo-journalism, junk news (clickbait)	Decontextualization, framing
Target	General population, voters, minorities	Consumers, investors, Patients	Own supporters (mobilization) vs. opponents (polarization)
Main risk	Threat to democracy, geopolitical division	Financial damage, Health risks (infodemic)	Social division, radicalization

Source: Own compilation

2.3 Generation and dissemination mechanisms

The spread of disinformation does not follow linear paths, but is the result of a complex interplay between users' psychological predispositions, technological amplifiers, and the structural conditions of the modern media ecosystem. As Watolla et al. (2025) explain, disinformation is not an isolated platform phenomenon, but is embedded in larger narratives whose impact is based on specific mechanisms of generation and distribution.

At the generation level, it is crucial that successful disinformation is often specifically tailored to cognitive biases. Content is designed to generate a high emotional response. Studies show that content with a strong emotional charge spreads significantly faster than neutral information. This is particularly true for emotions such as anger and outrage. Brady et al. (2017) describe this as "moral contagion," in which news that appeals to moral emotions within a social group is shared preferentially in order to signal one's own group identity and strengthen cohesion. In addition to this, Vosoughi et al. (2018) demonstrated in a comprehensive analysis on X's predecessor Twitter that false information spreads "faster, deeper, and wider" than true news. The authors attribute this primarily to the novelty value, as false information is not bound to reality and can therefore be made more surprising and sensational, which stimulates the human tendency to pass on new information. Ultimately, disinformation is particularly effective when it confirms existing worldviews (confirmation bias). In this context, Pennycook and Rand (2021) argue that it is not so much a lack of intelligence as "motivated reasoning" that leads to politically congruent false information being accepted and disseminated.

Social media and other online platforms play a special role in the current disinformation environment, which will be examined in more detail in Chapter 3. In order to understand the role of digital platforms in the context of disinformation and social division, a differentiated view of the technological and social mechanisms is necessary. Watolla et al. (2025) identify algorithmic curation and coordinated behavior as two key levers through which platforms can contribute to the spread of problematic content. Since the business models of large platforms are primarily based on maximizing dwell time and interactions, recommendation algorithms privilege content that promises high engagement in the form of clicks, comments, or shares. Emotional and polarizing content (not least in the form of disinformation) fulfills this exploitation logic particularly well, so that an unintended but systemic preference for such content can arise (Bakir & McStay, 2018; Meßmer & Degeling, 2023).

Against this background, it seems plausible to assume that these mechanisms lead to the formation of closed echo chambers and filter bubbles. However, empirical results paint a much more nuanced picture. Overview studies unanimously conclude that the ideal-typical notion of completely isolated information spaces in which users see only like-minded content is greatly exaggerated (Stark et al., 2019; Zuiderveen Borgesius et al., 2016). Bruns (2019) argues that concepts such as the filter bubble are often based on technological determinism, which ignores the actual diversity of individual media repertoires and the permeability of social networks. Stark et al. (2019) also show that the implicit personalization effects of algorithms have been overestimated to date and that filter bubbles in their pure form are rarely empirically detectable. Instead of thinking in binary categories, they argue that these phenomena should be viewed as a continuum between a broadly overlapping information repertoire and gradual fragmentation.

The assumption of complete informational isolation is therefore empirically untenable. Studies in the German context confirm that users are by no means trapped in homogeneous opinion spaces, but rather come into contact with a plurality of topics across different channels (Stark et al., 2019). Dubois and Blank (2018) also point out that people with a high level of political interest and diverse media sources are significantly less likely to end up in echo chambers; problematic constellations are more likely to be found in highly politicized but selectively informed subgroups. Meßmer and Degeling (2023) emphasize that polarization is primarily a social phenomenon that cannot be attributed to platforms as a single cause, but should be viewed in the context of individual selection decisions, social network structures, and algorithmic weightings.

The systemic problem therefore lies less in the isolation of foreign opinions and more in the nature of algorithmically promoted confrontation. Since recommendation systems evaluate interactions such as comments or dwell time as indicators of relevance , they cannot technically distinguish between agreement and disagreement. This encourages phenomena such as "hate following," in which users deliberately follow content that upsets or angers them emotionally (cf. Richardson et al., 2024). Algorithms primarily prioritize content that attracts attention, including posts that trigger anger. In contrast, content

that users simply like is not always reinforced in the same way. Sociological analyses, such as those by Törnberg (2022), suggest that, paradoxically, polarization is reinforced not by isolation, but by constant, emotionally charged confrontation with opposing worldviews. In a global digital space where local contexts are absent, political identity serves as the primary distinguishing feature. Contact with the "politically other" often leads not to understanding, but to a hardening of one's own position.

This dynamic is supported by experimental data. Bail et al. (2018) were already able to show that targeted exposure to opposing political views on Twitter tends to polarize rather than moderate attitudes. These results reinforce the suggestion that simple recipes such as "more opposing opinions in the feed" do not work empirically (Meßmer & Degeling, 2023). Algorithmic logic thus promotes not so much the creation of isolated bubbles as an arena of permanent, identity-forming conflicts. From the perspective of platform regulation, this means that although algorithms create incentives for polarizing content, systemic risks can only be adequately understood by taking into account usage practices and social lines of conflict.

Beyond organic dissemination, actors use technical means for artificial amplification, e.g., within the framework of the CIB. This includes the use of bot networks ("astroturfing," cf. Mahbub et al., 2019; Chan, 2024) to create the appearance of broad public support, as well as the targeted manipulation of search engines (data voids, cf. Norocel et al., 2023; Mannino et al., 2024). Golebiewski and Boyd (2018) show how actors deliberately occupy terms for which there is little established content in order to place disinformation prominently in specific search queries.

While this discussion paper focuses on current developments in the field of online platforms, an isolated view is insufficient for a holistic understanding. Disinformation often only unfolds its full social impact through interaction with traditional mass media in a "hybrid media system" (Chadwick, 2017). Strategic actors often use social media as a testing ground to place narratives that are then intended to diffuse into the mainstream. Marwick and Lewis (2017) describe this process as "trading up the chain": disinformation is first placed in niche forums or blogs, picked up by medium-sized influencers, and finally reaches journalists in established media. If the false report is picked up there, for example to refute it, its reach can nevertheless increase significantly, albeit unintentionally.

Political actors also deliberately use taboo-breaking and disinformation as a strategy of information flooding to monopolize media attention. Watolla et al. (2025) aptly refer to this as a "DDoS attack on human attention," i.e., a targeted strategy to overload the cognitive processing capacities of recipients with excessive information density. Traditional media outlets that report on these provocations unwittingly act as multipliers and legitimize the topics chosen by disinformation actors (Phillips, 2018). These cross-media flows are bidirectional. While online disinformation can influence the editorial agenda ("bottom-up"), established media brands often also serve as vehicles for disinformation when actors abuse their credibility through fake screenshots or quotes taken out of context (Watolla

et al., 2025). In summary, it can be said that the spread of disinformation is based on a symbiosis of emotional user engagement, algorithmic rewards for engagement, and the exploitation of journalistic attention logic.

2.4 Impacts and risks

Quantifying the specific damage caused by disinformation in academic research faces the fundamental methodological problem of causal inference. As a meta-analysis in Nature Human Behaviour shows, it is often almost impossible in field studies to isolate the influence of disinformation from existing socio-political polarization tendencies (Lorenz-Spreen et al., 2023). Selection effects in particular play a decisive role: users do not receive information passively, but actively select content that confirms their existing beliefs. Due to these methodological hurdles, evaluating content on a case-by-case basis is often not very effective.

2.4.1 Exposure and reception of dis

A central question in impact research concerns the discrepancy between the theoretical availability and actual reception of misinformation. Recent studies have focused intensively on the question of how many people are actually exposed to disinformation and why they respond to it. These studies show that susceptibility is not evenly distributed, but is subject to specific psychological mechanisms.

In a recent study, Hubeny et al. (2025) identify complex predictors for credibility assessment. Their results show that susceptibility is less a question of intelligence and more strongly correlated with personality traits such as "bullshit receptivity" (susceptibility to pseudo-profound statements) and a general conspiracy mentality. At the same time, traits such as "cognitive reflection" and "intellectual humility" act as protective factors. However, the decisive factor is identity protection ( ): users often accept false information not out of ignorance, but because it supports their existing worldview. This "myside bias" means that information that corresponds to one's own group identity is accepted more readily. Disinformation does not act as a universal "virus," but primarily takes hold where it resonates with the psychological dispositions and identity needs of specific target groups.

In addition to this, recent studies confirm that the decisive mechanism is often not a lack of knowledge, but "identity-protective motivated reasoning" (Hubeny, Nahon, and Gawronski, 2025). Users accept misinformation if it serves to enhance their own social group ("in-group"). This "partisan bias" leads even cognitively capable individuals to use their analytical skills to rationalize misinformation rather than expose it. Algorithmic recommendation systems reinforce this effect by prioritizing content that elicits strong moral-emotional responses (Brady et al., 2023).

This is particularly problematic in light of the "news finds me" (NFM) effect: since young users in particular are increasingly consuming news passively via their feeds (rather than actively searching for it), algorithmic curation can have a greater influence on their worldview. This concept describes the belief of individuals that they no longer need to actively search for news in order to be well informed, as relevant information will automatically reach them via their social networks and algorithmically curated feeds. This passive attitude toward information acquisition has far-reaching consequences for political knowledge, health behavior, and, not least, susceptibility to misinformation and disinformation.

Originally conceptualized as a direct consequence of the increasing prevalence of social media, NFM is understood in current research as a deeper cognitive disposition that is closely linked to specific media habits. Campbell and Hawkins (2025) show in their study that NFM does not arise solely from the frequency of social media use, but is significantly promoted by habitualized, unconscious usage patterns ("habits") and specific "mindsets." NFM perception correlates more strongly with a "connection mindset," i.e., the belief that social connections ensure the flow of information, than with an "algorithm mindset," which relies on technical filters.

Furthermore, the intensity of this perception is not static, but varies depending on the topic. While the concept was originally developed in the context of political news ("hard news"), Mosallaei et al. (2025) demonstrate that the perception is often even more pronounced in entertainment and sports news ("soft news"). A decisive factor here is individual interest: a high level of interest in a particular topic (e.g., politics) tends to weaken NFM perception, as it increases the willingness to actively search for information, while disinterest reinforces passive trust in the feed.

The passivity associated with NFM perception has far-reaching consequences for information integrity and susceptibility to misinformation and disinformation. Gil de Zúñiga and Cheng (2024) demonstrate in their review that NFM consistently correlates negatively with political knowledge, as the subjective feeling of being well-informed suppresses the need for active knowledge acquisition. Since users with high NFM scores assume that they are already well informed, they are less likely to feel the need to actively verify news or consult additional sources. Zhang and Jiang (2024) demonstrated in a health context that this attitude led users to actively avoid information during the COVID-19 pandemic. This avoidance correlated directly and positively with the likelihood of believing false health information, as corrective information was not actively sought. Lin et al. (2024) further differentiate this finding by also showing that NFM not only inhibits information seeking, but also correlates directly and negatively with actual health knowledge. In their study, perceived information insufficiency acted as a central mediator: individuals with high NFM perception subjectively feel no need for information, which significantly reduces their intention to actively search for valid information on social media.

Another similar mechanism exacerbates this problem in the context of disinformation. Users with high NFM perception tend to overestimate their own media literacy while considering others to be significantly more susceptible to manipulation. This phenomenon is referred to as "third-person perception." Tian and Willnat (2025) were able to show that this perceived immunity leads users to make less cognitive effort to verify the truthfulness of news items. Ironically, this overconfidence results in NFM users interacting with disinformation more frequently and being more susceptible to it, as they lower their protective mechanisms due to a false sense of security.

This passivity not only affects the reception side, but also drives the spread of disinformation. Hawkins and Campbell (2025) identified NFM perception in studies of the US "alt-right" as a direct predictor of active sharing of disinformation. Users who rely on random news contact and consume news interactively via mobile devices thus often unknowingly become multipliers in disinformation networks. The "news finds me" effect thus marks a structural weakness in the digital public sphere that reduces critical questioning and systematically undermines resilience to manipulative content.

2.4.2 Risks

Despite the methodological difficulties in proving direct causality, concrete risks can be identified, including the transition from the digital world to physical threats and damage. The riots in Southport, UK, in the summer of 2024 serve as a prominent case study of the direct link between online content and offline harm. After a knife attack, hostile actors and far-right networks exploited the resulting "information vacuum" to spread false narratives about the attacker's identity (falsely identified as a Muslim asylum seeker) virally. This disinformation was massively accelerated by algorithmic amplification on platforms such as X and TikTok, leading directly to violent attacks on a mosque and migrant shelters. The case demonstrates how online disinformation acts as an amplifying factor in emotionally charged situations, which can intensify latent social tensions and translate into physical violence. The Southport riots are thus an example of what is discussed in security research as "stochastic terrorism": the statistically probable triggering of violence through mass incitement, but whose specific target appears random. A study by Müller and Schwarz (2021), for example, shows that there is a correlation between the local intensity of anti-migrant disinformation on social media and the frequency of hate crimes in the corresponding areas.

In addition to eruptive violence, the damage caused by disinformation is increasingly manifesting itself in a structural erosion of democratic participation, described in research as "chilling effects." In its latest report, the Council of Europe (2025) warns against coordinated attrition strategies such as "Operation Overload," in which journalists and civil society actors are deliberately flooded with a mass of false inquiries in order to tie up their resources and persuade them to withdraw from public discourse. This mechanism has a particularly serious impact in the area of "gendered disinformation.". This phenomenon is

also known as "networked misogyny": it is not isolated trolling, but coordinated, often sexualized campaigns (including deepfake pornography) that are deliberately exploited to discredit women in politics and journalism (cf. Di Meco & Brechenmacher, 2020). This is classified as technology-facilitated gender-based violence (TFGBV). The primary goal of these measures is to drive the women affected out of public discourse. (OECD, 2025; Anstis & LaFlèche, 2025). Empirical evidence shows that affected female politicians withdraw from online discourse significantly more often than their male colleagues. This trend is also evident in prominent resignations such as that of Dutch Deputy Prime Minister Sigrid Kaag and leads to a measurable impoverishment of representative democracy.

The overall impact on elections is subtle but significant. Mauk and Grömping (2024) demonstrate that while disinformation does not necessarily directly influence election results, it can significantly undermine belief in the fairness of the electoral process. This applies to both winners and losers. From an economic perspective, disinformation is once again identified as one of the greatest short-term risks in the World Economic Forum's Global Risks Report 2025, as it can destabilize markets and destroy the social trust capital necessary for investment. Estimates suggest that disinformation costs the global economy up to $78 billion annually, for example through damage to corporate reputations or market manipulation. In the health sector, studies estimate the cost of vaccine-related disinformation alone (e.g., through additional treatments and outbreaks of preventable diseases) to be in the billions (ECFSN, 2025a).

Another risk lies in the damage to epistemic trust. The World Economic Forum's Global Risks Report 2025 identifies disinformation as one of the top risks to social cohesion. Finally, current research points to a critical paradox: the well-intentioned fight against disinformation can itself become a problem. Hoes et al. (2025) demonstrate in a recent study that not only the misinformation itself, but also alarmist reporting about it can have negative effects. The study shows that intense media warnings about disinformation ("We are surrounded by lies") lead to general "epistemic uncertainty" among citizens. The result is not necessarily greater vigilance, but rather a blanket loss of trust in all sources of information, including scientists and established media. This phenomenon calls for caution: an undifferentiated scandalization of disinformation can damage trust in public discourse just as permanently as disinformation itself.

2.5 Recognition

Effectively tackling the disinformation crisis requires an integrative approach that combines technological detection methods with behavioral psychological and structural interventions. The approaches to combating disinformation in the platform landscape and a detailed overview of current findings can be found in Sections 3.3.2 and 4.1.

In recent years, automated detection of disinformation has evolved from simple keyword filters to highly complex, AI-supported systems. While traditional approaches focused

primarily on text analysis using natural language processing (NLP), purely unimodal methods are now considered insufficient given the multimedia nature of modern disinformation. Current research therefore favors the use of multimodal deep learning models. These combine transformer architectures such as BERT for text analysis with convolutional neural networks (CNNs) for image and video processing to identify semantic inconsistencies between different modalities (Nasser et al., 2025). In addition to this, knowledge graphs are gaining in importance. These structure factual knowledge in the form of entities and relations, enabling algorithms to flag false claims through automated comparison with verified databases. While this approach reduces dependence on human fact-checkers, it carries the risk that AI models may misunderstand or "hallucinate" cultural nuances such as satire when context is lacking (Feng et al., 2025).

Since much problematic content falls into a legal gray area ("lawful but awful"), the focus of platform operators and regulatory authorities is increasingly shifting from content review to CIB detection. This involves analyzing not the truthfulness of the statement, but the topology of its dissemination. Graph neural networks (GNNs) can be used to identify clusters of accounts that act in unnatural synchrony, for example to simulate artificial grassroots movements ("astroturfing") (Feng et al., 2025).

The debate on how to effectively combat disinformation has traditionally suffered from a lack of standardized metrics that go beyond mere content moderation statistics. While platforms now publish transparency reports showing the number of removed content or blocked accounts, these operational metrics say little about the actual extent of user exposure to harmful content. Evidence-based regulation therefore requires a shift in analytical focus: away from simply counting moderated content and toward quantifying algorithmic amplification and the systemic causes of dissemination.

One possible tool for measuring platform-specific responsibility is the Misinformation Amplification Factor (MAF) developed by Allen (2022). This indicator operationalizes the question of the extent to which a platform's design favors the spread of misinformation by comparing the actual reach of a piece of disinformation with the reach that would be expected based on the organic follower structure of the author. The calculation of the MAF reveals significant differences in platform architectures and confirms the hypothesis that disinformation flourishes more in systems based on frictionless dissemination and algorithmic recommendations than in networks based primarily on social graphs (follower relationships).

Empirical analyses based on this show, for example, that platforms such as TikTok and Twitter (now X) tend to have very high amplification factors. This is due to mechanisms such as one-click retweeting or the "For You" feed, which spread content virally regardless of the trustworthiness or established base of the sender. In contrast, platforms such as Instagram or Facebook (in their classic feed) have lower MAF values, as dissemination there is more closely linked to direct followers and there are barriers to further dissemination. However, here too, a dynamic deterioration can be seen as soon as formats such

as "Reels" are introduced, which, similar to TikTok, rely heavily on algorithmic recommendations of unconnected content and can thus significantly increase the amplification factor for disinformation.

Despite regulatory progress through the Digital Services Act (DSA) and mandatory risk assessments, there remains a significant gap between the necessary data and the information provided by the platforms. A comprehensive audit of the 2024 risk assessments by the Integrity Institute (Allen et al., 2025) shows that while platforms provide qualitative descriptions of their systems and mitigation measures, there are significant gaps in the quantification of key risk dimensions. This particularly affects scale, cause, and nature. With regard to scale, providers often limit themselves to relative prevalence rates (e.g., 0.01% of views are harmful) without providing absolute figures that would illustrate the actual volume of exposure. However, estimates suggest that even small percentages can translate into billions of views of harmful content in absolute terms. For example, data from TikTok suggests that around 30 billion views per quarter are attributable to content that is later removed due to policy violations. On YouTube, too, the volume of harmful content is estimated at several billion views per quarter. This discrepancy between low percentages and massive absolute view counts illustrates that simply stating prevalence rates obscures the systemic risk. Potentially more problematic is the shortcoming in analyzing the causes. For effective regulation, it is essential to understand what proportion of harmful exposure is due to conscious user decisions and what proportion is caused by platform design with features such as proactive recommendations or autoplay functions. The audit found that none of the very large online platforms (VLOPs) provide quantifiable data on the proportion of exposure to disinformation generated by recommendation algorithms.

This prevents an assessment of whether the platforms' business models are the causal drivers of the risks. Finally, there is a lack of transparency regarding the nature of the risks, particularly with regard to their concentration. Average values often obscure the fact that risks can occur disproportionately in specific echo chambers or among vulnerable groups ("pockets of misinformation"). There is a lack of data on whether certain user groups are exposed to extremely high doses of disinformation and to what extent algorithmic systems reinforce this concentration. The current practice of platforms to demonstrate the success of their measures primarily through operational metrics such as the number of deleted posts or the speed of moderation is therefore insufficient. Paradoxically, a high number of deletions can be an indicator of the failure of the design to prevent dissemination a priori. Effective transparency must therefore demonstrate that countermeasures not only combat the symptoms, but also reduce algorithmic amplification (MAF) and measurably reduce the causal role of recommendation systems in the spread of disinformation.

While technical detection methods are advancing, it is becoming apparent that the problem is not purely technical in nature. The causes lie deeper in the economic structures of the platforms, which are analyzed in the following chapter.

3 Online platforms in the disinformation economy

3.1 Economic incentive structures

The spread of disinformation on digital platforms is not solely a technological or sociological phenomenon, but is deeply rooted in the economic structures of the "attention economy." To strengthen the resilience of democratic systems, it is essential to understand the monetary incentives. These incentives motivate actors ranging from platform operators to advertisers to individual influencers to directly or indirectly promote the spread of harmful content.

The dominant business model of the major social media platforms (“Big Tech”) is based on maximizing user retention time in order to extract behavioral data and sell advertising space. In this model, user engagement (likes, shares, comments) acts as the central currency. In his analysis of the "natural engagement pattern," Allen (2022) describes how content that approaches the line of what is prohibited ("approaching the line") or elicits strong emotional reactions such as anger naturally achieves higher engagement than neutral, factual information.

This leads to a fundamental conflict of interest: platforms face the economic choice of either investing in security and moderation, which often curbs engagement and revenue, or allowing the algorithmic dissemination of "high-engagement" content. This content disproportionately includes disinformation and hate speech. A microeconomic study by Dey et al. (2025) provides a basis for this: The authors use model theory to show that polarization and bias can act as substitutes for platforms in maximizing profits. In order to maximize advertising revenue, platforms are economically motivated to drive users into "echo chambers" and favor polarizing narratives, as this increases loyalty to the platform. Current analyses show that platforms such as X (formerly Twitter) and TikTok have particularly high "misinformation amplification factors" (MAF), which indicates that their recommendation systems are strongly geared towards viral dissemination. This mechanism is further exacerbated by the transformation of verification systems into paid services ("pay-for-play"). For example, X's paid verification model allows malicious actors to buy algorithmic priority simply by paying, regardless of the authenticity of their identity or content (Jones, 2025).

An often underestimated driver of the disinformation economy is the global advertising ecosystem, particularly programmatic advertising. Programmatic systems auction advertising space in milliseconds via a chain of demand-side platforms, exchanges, and supply-side platforms. This architecture no longer links advertising budgets to specific editorial environments, but to target group profiles and auction results. As a result, ads from major brands routinely end up on websites that have been proven to spread disinformation (Ahmad et al., 2024; Braun & Eklund, 2019). Ahmad et al. (2024) show that companies across many industries advertise extensively on misinformation websites, even though financing such sites carries both reputational and financial risks. At the same time,

they document that this misallocation does not primarily result from conscious consent, but from the delegation of budget decisions to automated ad tech platforms and intermediary agencies that are optimized for reach and efficiency, not content quality. In interviews with advertising industry players, Braun and Eklund (2019) describe programmatic advertising as an infrastructure that monetizes "fake news" and quality journalism through the same pipeline systems, thus creating financial incentives for clickbait disinformation (Braun & Eklund, 2019).

Against this backdrop, disinformation can be understood as a negative externality of the digital advertising market. Diaz Ruiz (2025) argues that many marketing managers consider the harmful consequences of programmatic advertising to be a marginal problem. For example, the monetization of fake news is ignored and normatively "defined out of the market": the idea that technology is neutral and the market corrects itself obscures the fact that advertising decisions actively contribute to financing disinformation. McGowan et al. (2024) further show that ad tech intermediaries are not merely neutral "conduits," but rather help shape the market as "mediators." Through black-box optimizations and bundled products, they effectively decide which inventory sources, including problematic sites, are given preference.

At the same time, an explicit "disinformation-for-profit" industry has emerged. A joint study by the Carter Center and the McCain Institute uses traffic and advertising data to reconstruct that over 81% of identified disinformation websites are connected to programmatic advertising networks such as Google Ads and finance their content through automated advertising placements (Scholtens et al., 2024). A special part of this ecosystem are "made-for-advertising" (MFA) sites: websites whose business model consists almost exclusively of generating as many programmatic impressions as possible, often with low-quality, recycled, or AI-generated content. Despite recent efforts by major advertisers to reduce this spending, hundreds of millions of US dollars per quarter continue to flow into environments that often host clickbait, misinformation, or misleading AI content. Research by DoubleVerify and other measurement providers also documents that MFA networks are increasingly building sports, lifestyle, and pseudo-news sites with AI-generated content that deliberately imitates reputable media outlets in order to maximize programmatic advertising revenue (DoubleVerify, 2024).

This economic foundation is intertwined with the political disinformation landscape. Programmatic platforms and social media advertising systems enable political and ideological actors to address highly granular target groups without openly identifying themselves as the source. In an analysis of Facebook and Instagram advertising in 95 countries, Votta et al. (2024) show that political microtargeting is now widespread globally and often works according to the pattern of "simple demographic criteria plus interests." In the context of the unrest following the Southport attack (see chapter 2.4.2), investigative research and parliamentary evidence suggest that, over a period of months, anonymous actors invested considerable budgets in anti-Muslim and anti-immigrant campaigns that were played out via meta-platforms; One report mentions around US$1.2 million being spent

on reinforcing xenophobic narratives via Facebook advertising (Jones, 2025; D'Souza, 2025). According to these reports, Meta was only able to reconstruct the actual financing structures and clients to a limited extent, revealing a gap in the advertising platforms' "know your customer" mechanisms. From a democratic theory perspective, this creates a situation in which private companies profit from the dissemination of polarizing and misanthropic messages, while the originators disappear behind shell companies, middlemen, and agency structures.

The structural role of the ad tech system is thus doubly problematic: on the one hand, it enables the refinancing of explicit disinformation actors, and on the other, it exacerbates the incentives for all content producers to generate attention at (almost) any cost. In her submission to the UK Science and Technology Committee, Middleton (2025) describes the digital advertising ecosystem as a highly opaque "black box" market in which advertising budgets pass through a chain of intermediaries before arriving at a specific website. This lack of transparency means that, despite "brand safety" promises, brands can end up on sites that spread hate speech, conspiracy narratives, or manipulative pseudo-news. McGowan et al. (2024) show that intermediaries benefit economically from bundling as much cheap reach as possible. This creates a structural bias in favor of precisely those environments, such as MFA and disinformation sites, that generate many clicks at low cost.

Amnesty International (2025) extends this criticism to include a human rights perspective: In its response to the British Parliament's inquiry on "Social Media, Misinformation and Harmful Algorithms," the organization argues that the business model of the major platforms is fundamentally based on surveillance-based advertising. Since personalized advertising is based on user profiles that are as detailed as possible, recommendation algorithms reward content that maximizes attention. This often includes content that activates fear, anger, and enemy stereotypes. In a technical companion paper on the UK riots, Amnesty shows how design decisions by X (formerly Twitter) around recommendation systems, unthrottling borderline content, and monetizing reach contributed to massively amplifying racist and anti-migrant narratives after the Southport attack. The economic logic of generating as much engagement as possible for the most profitable ads is thus closely linked to the algorithmic amplification of disinformation and hate.

Against this backdrop, regulatory and intervention proposals that address not only content moderation but also the financial infrastructure itself are coming into focus. Ahmad et al. (2024) show that even relatively simple, information-based interventions can significantly reduce the financing of misinformation. These include, for example, transparency campaigns aimed at advertising customers that disclose on which sites their ads appear without significantly impairing the overall effectiveness of the advertising. Based on their analysis of the "disinformation economy," Scholtens et al. (2024) argue for specifically interrupting payment flows to disinformation sites through stricter brand safety standards, blacklists, and greater responsibility on the part of intermediaries. Diaz Ruiz (2025) adds that marketing communities must question their own "socio-technical imaginations,"

which have so far dismissed the negative consequences of programmatic advertising as external, non-market effects.

Amnesty International (2025) goes one step further and calls for a structural shift away from profile-based advertising. From a human rights perspective, a business model based on massive data collection, profiling, and subsequent microtargeting advertising is incompatible with protection against discrimination, surveillance, and manipulative influence. Instead, Amnesty advocates for data-minimizing, context-based advertising models, binding human rights due diligence obligations, and significantly stricter transparency and information rights vis-à-vis platforms. Overall, current research suggests that an effective fight against disinformation cannot stop at correcting individual pieces of content: it must focus on the economic incentive structures of the ad tech system that monetize the reach of disinformation. Measures to promote user resilience and algorithmic curation remain important, but will only reach their full potential if the financial flows to disinformation actors are systematically and sustainably capped (Ahmad et al., 2024; Diaz Ruiz, 2025; Middleton, 2025).

In addition to platforms and advertising networks, individual actors who profit from polarization have become more professional. Jones (2025) introduces the term "disinfluencer": actors who routinely spread misinformation and have a disproportionate influence on trends due to their reach. These actors often operate in an economic environment that rewards extreme behavior. A striking example is the change in X's monetization model in October 2024, whereby creators are paid directly based on the engagement (views, likes) their posts generate. Since disinformation and hate speech achieve significantly higher interaction rates, this model creates a direct financial incentive for the production of "rage bait" and polarizing content.

Recent research differentiates more precisely between different roles. Pournaki et al. (2025) distinguish between "influencers" who generate content and "multipliers" who curate and amplify content. Their analysis shows that multipliers in particular play a decisive role in bundling and amplifying ideologically consistent and polarizing content , which accelerates the formation of echo chambers. Abdul Rahman et al. (2025) complement this perspective by analyzing "amplifiers," who often act as catalysts for harassment campaigns ("indirect swarming") without explicitly violating guidelines themselves. The economic value of these amplifiers lies in their ability to direct attention and frame discourse. Prominent examples such as Elon Musk show how platform owners themselves can become "super amplifiers" by interacting with fringe accounts, massively increasing their visibility and thus their monetization potential.

Finally, the creation and dissemination of disinformation has developed into a global service industry. The practices of coordinated inauthentic behavior (CIB) described in Chapter 2.2 are becoming increasingly professionalized and offered as "disinformation-as-a-service" (DaaS) (cf. CACI, 2025). External actors, such as PR firms or cybercriminals, offer the manipulation of public discourse on the dark web or openly in exchange for

payment. Jones (2025) refers to the case of a Tel Aviv-based company that specifically targeted Western audiences with AI-generated anti-Muslim content. These actors use the cost efficiency of generative AI to produce content (text, images, video) in large quantities and at low cost. This significantly lowers the barriers to entry for disinformation campaigns and makes it possible to tailor content to specific demographic target groups ("microtargeting"). In conjunction with "click farms" or bot networks, this creates a shadow economy that functions primarily not on ideological grounds but as contract work, undermining the integrity of the digital space for purely profit-driven reasons.

3.2 Algorithmic amplification

While Chapter 3.1 outlined the economic incentive structures, this chapter analyzes the technological implementation. Recommendation systems are the operational levers that translate economic goals into curatorial decisions. They do not function as neutral distributors, but as active "algorithmic gatekeepers" that replace editorial judgment with automated metrics, thereby systematically favoring specific content (Chiridza & Mare, 2025).

The (implicit) amplification of disinformation often results not from deliberate programming for falsehood, but from the mathematical logic of popularity-based ranking functions. A recent agent-based simulation study (Jakobsen et al., 2025) shows that algorithms optimized primarily for popularity metrics (views, likes) significantly accelerate the spread of disinformation, as this content is often optimized for novelty and emotional appeal and thus crosses initial interaction thresholds more quickly. In contrast, approaches such as item-based collaborative filtering, which are based on similarity patterns between content , could technically even limit exposure to misinformation, but are less commonly used as a primary ranking factor.

Another technical phenomenon that favors the spread of disinformation is what is known as "sycophancy" in modern AI models, which describes a tendency of AI models to give conformist responses. Research in the medical context shows that large language models (LLMs) trained using reinforcement learning from human feedback (RLHF) tend to agree with user input, even if it is objectively incorrect. Algorithms prioritize a misguided helpfulness and user confirmation over factual accuracy, which means that they do not correct conspiracy narratives or medical misconceptions from users, but rather reinforce them algorithmically.

Chen et al. (2025) demonstrated this experimentally by asking AI models to write medically nonsensical warnings about generic drugs. Instead of correcting the logical error, the models complied with the request in 58 to 100% of cases and hallucinated convincing but false arguments to support the user's misconception. Although this behavior could be mitigated through technical "supervised fine-tuning" or specific prompting strategies (e.g., "rejection permission"), developers of general-purpose LLMs often lack the economic incentive to implement them, as users prefer interacting with "polite" and affirming chatbots.

This technical disposition to confirm user bias ("algorithmic confirmation") creates a self-reinforcing feedback loop that drives users deeper into information bubbles.

The general algorithmic logic is designed to provoke interaction. This leads to sensationalist, emotionally charged, or polarizing content being systematically favored and amplified, often at the expense of accuracy or safety. Recommendation systems are not passive or neutral mechanisms. Internal documents from Meta (Facebook Papers) confirmed that the ranking of the news feed was based on predicted engagement and that separate models existed for content that users were likely to "share," which is at the core of algorithmic virality. The algorithms are optimized to amplify exciting, emotionally charged, or polarizing material. The Southport riots described in section 2.4.2 illustrate this data: anti-Muslim and anti-immigrant content received 65% of the total impressions of posts related to the riots, while factual corrections only reached 13%. TikTok's "For You" feed has also been criticized for exposing young people who show an interest in mental health to increasingly extreme or potentially self-harming content through successive recommendations (see "Rabbit Holes").

The predictability of these algorithmic preferences enables external actors to manipulate the systems in a targeted manner ("adversarial attacks"). A key mechanism here is cross-platform coordination. Recent analyses of US elections show that disinformation campaigns rarely take place in isolation, but operate as multi-layered networks. Content is "seeded" on loosely regulated platforms (e.g., Telegram, 4chan) and then shared in a coordinated manner on mainstream platforms (X, Facebook) to generate artificial virality (Hristakieva et al., 2024; Tardelli et al., 2024). This strategy exploits the inertia of platforms' internal detection systems, which often only evaluate local signals (on their own platform) but overlook coordinated external amplification.

Another mechanism of systematic engagement amplification is indirect swarming, in which amplifiers can mobilize their followers through masked language or simply retweeting a message to harass a target without directly violating the platform's terms of use. In addition, actors are increasingly using "keyword obfuscation" (concealment). The use of "algospeak" circumvents automatic content detection systems. This involves deliberately modifying terms or using visual codes so that the message remains decodable for the human target audience. New research suggests that generative AI automates this process ("AI-driven obfuscation") by varying content en masse in such a way that it remains below the detection thresholds of moderation algorithms (Romanishyn et al., 2025).

Beyond purely economic logic or opportunistic exploitation by third parties, the political instrumentalization of algorithmic curation by the platform operators themselves is increasingly coming into focus. A significant shift in the political thrust and perception of these interventions can be observed. Until well into the early 2020s, the debate on content moderation was primarily dominated by conservative and right-wing voices, who accused the major platforms of systematic "liberal bias" and censorship of conservative voices. A central but often opaque tool in this discussion is "shadowbanning" (the hidden reduction

of visibility). While platforms often justify this as a technical necessity against spam, studies show that the criteria for such throttling—regardless of the general thrust—remain opaque and can erode trust in digital discourse (Delmonaco et al., 2024; Thomas & Manalil, 2025). Shadowbanning acts as an invisible regulatory tool that can potentially be abused for political censorship ("soft censorship") by systematically suppressing unpopular but legal content ("lawful but awful"). This can lead to uncertainty among users and increase polarization (Chen & Zaman, 2024).

A differentiated assessment of shadowbanning requires weighing economic efficiency against regulatory transparency. In contrast to strict content removal, which drives extreme users off the platform, shadowbanning acts as a tool for market maximization: it allows platforms to retain even users with extreme views as part of their user base, as they are left unaware that their content is invisible to others. From an economic perspective, this strategy often leads to higher platform profits and greater market coverage than content deletion, as it preserves the "posting utility" of extreme users while minimizing the "reading disutility" (discomfort) of moderate users by hiding this content. Modeling even shows that, under certain conditions, shadowbanning can increase overall social welfare more than transparent deletion procedures. This is especially true when user assumptions about its prevalence are moderate. However, this creates a regulatory tension: while current laws such as the Digital Services Act (DSA) push for maximum transparency, opacity is precisely the defining feature of shadow banning. A blanket ban on this practice could therefore have unintended negative welfare effects, which is why Hojati and Nault (2025) recommend a precise definition of legitimate use cases (e.g., against bots or spam) instead of a ban, in order to reap the benefits of reach control without destroying user trust through arbitrariness.

Empirical studies have found little evidence to support the accusation of structural discrimination against right-wing content. On the contrary, recent audits of algorithmic amplification, for example during the 2024 US elections, show the opposite trend: Platforms such as X (formerly Twitter) now exhibit significant "exposure inequality," in which right-leaning users in particular are disproportionately exposed to content that confirms their own political views, while access to counterarguments is algorithmically minimized (Ye et al., 2025). Milli et al. (2025) also demonstrated that engagement-based algorithms systematically favor content that expresses anger and hostility toward the political "outgroup." This mechanism favors polarizing actors.

In the current phase of platform governance (since around 2023/24), a new dynamic is emerging that is characterized less by liberal hegemony and more by strategic adaptation to right-wing power centers. Researchers describe this using the term "anticipatory obedience" coined by Timothy Snyder (Bassin & Potter, 2024; Snyder, 2017). In the face of impending regulatory intervention or political reprisals by right-wing governments (such as under a Trump administration), tech companies and media owners tend to proactively adjust their moderation policies to avoid conflict (The Guardian, 2024).

Marc Owen Jones documents how platform owners themselves become political actors (Jones, 2025). In the case of Elon Musk and X, this manifested itself in the re-platforming strategy for previously banned far-right actors and the algorithmic amplification of specific political narratives. When algorithmic visibility is no longer based on neutral engagement metrics but on opaque political parameters ("bias injection"), the platform transforms from a marketplace of attention into an instrument of covert opinion control (Dey et al., 2025). This "strategic accommodation" of authoritarian tendencies represents a fundamental shift: the danger no longer comes primarily from "over-blocking" (censorship), but from the selective, politically opportunistic promotion of content that is geared toward the interests of political power.

The assumption that platform algorithms unintentionally promote polarization by maximizing engagement is impressively confirmed by current empirical data from the 2025 German federal election campaign. Two recent audits by the Bertelsmann Foundation show that the technical architecture of platforms is not merely a neutral "mirror" of political debate, but acts as a distorting amplifier that systematically favors specific political styles and actors.

A key finding of current research is that algorithms have an inherent preference for negative campaigning. The Progressive Center's (2025) analysis of over 30,000 short political videos shows that content that attacks or disparages political opponents is rewarded with a 40% reach bonus. Constructive approaches or positive self-presentation, on the other hand, are algorithmically "penalized" and achieve significantly less visibility.

These results are consistent with international research on "optimization for divisiveness." Orecchia (2025) shows that algorithms trained on engagement metrics inevitably prioritize content that triggers moral outrage and conflict, as these reactions are cognitively faster and more intense than approval. This creates a feedback loop: political actors learn through trial and error that aggressiveness is the only currency that gets them into the feed, and they adapt their communication accordingly ("algorithmic accommodation").

Algorithmic preference is also not always politically neutral. A "sock puppet audit" by the Bertelsmann Foundation (2025) on the 2025 federal election reveals a massive asymmetry in favor of the AfD. On TikTok, 50% of all party political content displayed to young, undecided users (aged 21-25) in the "For You" feed was related to the AfD. This made the party three times more visible than the CDU/CSU (15%). The algorithm also leads users to right-wing content extremely quickly. An AfD video was recommended after an average of only 11-12 minutes of use, while content from the SPD or FDP often appeared only after more than an hour or not at all.

According to the study, this dominance cannot be explained solely by the AfD's higher posting frequency. For example, the SPD produced more videos during the study period but still received significantly fewer algorithmic recommendations. This suggests that the algorithm recognizes a specific "fit" between the AfD's content patterns (e.g., nativist

narratives, fear-framing on migration) and its own optimization goals (dwell time, emotional activation). Ye et al. (2025) also confirm this pattern for the US: Their audit study on X showed that algorithms systematically led right-wing users into corresponding filter bubbles, while neutral users were confronted with right-wing content more quickly than left-wing content (Ye et al., 2025).

Algorithms also indirectly influence the thematic agenda. The analysis by the Progressive Center shows that videos on the topic of migration are systematically given a reach bonus (+11%), while complex future topics such as the environment (-18%) or education (-17%) are given less visibility. This structurally disadvantages parties that focus on differentiated factual issues and favors actors who exploit "wedge issues." These technically generated visibility asymmetries are met with changing user reception behavior, which amplifies the algorithmic effects (see chapter 2.4.1).

3.3 Current technological developments

The dynamics of disinformation are currently being readjusted by three converging factors: the mass availability of generative AI, the strategic withdrawal of platforms from moderation responsibilities, and the increasing fragmentation of the regulatory landscape. These developments are changing not only the conditions under which misinformation is produced, but also the architecture of public discourse.

Generative AI poses an additional challenge because, on the one hand, it significantly reduces the time needed to create convincing but fake images, which also makes detection more difficult. During the Southport riots, AI-generated content was used to reinforce xenophobic narratives. However, the discussion about artificial intelligence in disinformation has shifted from the fear of perfect "deepfakes" to an economic scaling problem. While technically highly complex, hyper-realistic fakes continue to pose a threat for targeted attacks, current research identifies a more subtle but potentially even more serious systemic threat: the flooding of the digital information space with mass-produced, low-quality AI content, known as "AI slop."

3.3.1 AI Slop

The term "AI slop" describes a new category of digital content created by generative AI at high speed and without significant human quality control or curatorial care. Madsen and Puyt (2025) define slop as "generative waste" (artificially generated low-quality output) characterized by seven dimensions, including enormous volume, high speed of dissemination, and erosion of cultural and epistemic value. Unlike targeted disinformation, which often pursues precise narrative goals, slop is primarily characterized by its banality and ubiquity: it is synthetic "filler" that ranges from generic essays and clickbait blogs to bizarre AI-generated images.

A prominent example of the viral spread of synthetic content on social platforms is the phenomenon of "Shrimp Jesus." This refers to a case of AI-generated image motifs that combine religious iconography with unusual elements and achieved high interaction rates on Facebook. Although this content may seem harmless at first glance, it serves as a vehicle to saturate algorithmic recommendation systems and capture attention. Its spread is not limited to social media. Ansari (2025) points out that by May 2025, an estimated 52% of new online articles were machine-generated, representing a significant contamination of the entire information ecosystem.

The production of AI slop is driven by clear economic logic. Since generative AI reduces the marginal costs of content creation to virtually zero, actors can exploit the attention economy of platforms with minimal effort ("spam 2.0"). This dynamic is being exploited politically, a phenomenon that Klincewicz et al. (2025) refer to as "slopaganda." Here, the sheer mass of low-quality AI content is used to flood the information environment in such a way as to deliberately impair the decision-making of groups. One example of this is the Romanian elections of 2024/25, in which candidates strategically used AI-generated memes and crude visualizations. This content did not serve to deceive through realism (as with deepfakes), but used the aesthetics of low-quality digital content to simulate approachability and spread nationalist narratives virally, bypassing traditional media filters.

The mass dissemination of AI slop leads to fundamental "data pollution." Ansari (2025) warns of a self-reinforcing feedback loop: if generative systems are increasingly trained on their own synthetic output, this leads to homogenization and semantic degradation of information quality. For the information economy, this means a massive increase in verification costs. Distinguishing between authentic, human-curated content and low-quality synthetic content is becoming increasingly resource-intensive for both users and algorithmic filters.

Van Rooij (2025) argues that this development leads to a lasting impairment of the information environment and quality, which jeopardizes the functionality of knowledge infrastructures. When search results, online articles, and even scientific publications are increasingly interspersed with plausible-sounding but factually hallucinatory AI texts, confidence in the reliability of information in general erodes. This fosters an environment of epistemic uncertainty in which the main problem is no longer the censorship of information, but the impossibility of identifying relevant signals in a flood of synthetic noise. The danger of AI slop thus lies less in the perfect deception of the individual than in the systemic erosion of the basis of trust on which public discourse is founded.

3.3.2 Dissemination: Automation, coordination, and cross-platform dynamics

Beyond pure content production, the technological infrastructure of disinformation has evolved significantly. Current research literature makes it clear that the operational tactics of CIB now go far beyond the mere quantitative mass of automated text bots. Instead,

they are increasingly characterized by complex, multimodal strategies and a functional division of labor within the dissemination networks.

The ecosystem of manipulation has evolved into a two-sided market in which "disinformation-as-a-service" (DaaS) is established as a business model. In their study of crowdsourcing platforms, Soliman and Rinta-Kahila (2024) analyze how organizers structure the interaction between requesters and crowdworkers in order to offer manipulation services on a scalable basis. The spectrum ranges from fake reviews to political influence. These platforms use discursive strategies of "language sanitization" to frame their activities as legitimate marketing services, thereby lowering the moral threshold for those involved. In addition, Ferrara (2024) warns against the integration of generative AI into these value chains. Through the use of large language models (LLMs), malicious actors can now create synthetic identities that not only produce content but also imitate human interaction patterns in a deceptively authentic way, making it extremely difficult for detection systems to distinguish between organic and inorganic behavior.

At the same time, a differentiation of actor roles can be observed. In their study on the spread of misinformation, Verdolotti et al. (2025) identify distinct behavioral archetypes: "amplifiers" who initially accelerate content, "super-spreaders" who ensure massive reach, and coordinated accounts that act in concert. This functional specialization indicates a high degree of professionalization, in which different account types are used strategically to overcome the algorithmic hurdles of the platforms.

Another key development is the shift from text-based spam to complex audiovisual coordination patterns, especially on video-centric platforms. In their analysis of TikTok content related to the 2024 US presidential election, Luceri et al. (2025) demonstrated that actors use generative AI to evade detection. Instead of posting identical videos that would be easily detected by hash filters, coordinated networks use identical AI-generated voiceover tracks or specific visual templates such as split-screen formats, while the visual content itself varies. This "semantic coordination" enables the synchronized amplification of political narratives that appear to conventional detection systems as organic diversity.

Modern disinformation campaigns do not operate in closed silos, but maximize their influence through strategic "cross-platform diffusion." Cinus et al. (2025) were able to demonstrate in the context of the 2024 US election that coordination networks deliberately cross platform boundaries to amplify narratives. Their analysis revealed that Russian-associated media were systematically promoted across Telegram and X (formerly Twitter), with a significant overlap in user bases serving as a transmission belt. Gerard et al. (2025) identify so-called "bridge users" as key players in this process. These users act as consistent "early initiators" who transfer narratives from one platform (often with less moderation) to another, where they trigger diffusion into new communities. This form of "narrative migration" is particularly resistant to platform-specific moderation measures, as deleting content on one platform does little to stop the flow of the narrative across the entire network.

In doing so, actors deliberately use the technical and structural characteristics (such as encryption or lack of moderation) of alternative services as spaces of retreat and coordination. Colizzi et al. (2025) examined corresponding coordination patterns on alternative platforms such as Gab, VK, Minds, and the Fediverse and found platform-specific strategies ranging from echo chambers on Gab to hierarchical distribution models on Telegram. Telegram in particular has become a central hub for information operations due to its minimal moderation and encryption architecture. In a large-scale analysis of multilingual political news on Telegram, Blas et al. (2025) uncovered five orchestrated information operations, including a Russian-backed influence campaign and pro-Palestinian amplification networks that were able to operate undisturbed. Pakina et al. (2025) further show that encrypted platforms such as WhatsApp and Telegram, due to their end-to-end encryption, act as "blind spots" of moderation, where AI-driven propaganda can mature undisturbed and be disseminated en masse before spilling over into open discourse.

The dissemination does not always take place via explicit links. Yin et al. (2025) describe the phenomenon of "implicit propagation," in which topics and frames spread across platform boundaries without direct URL references, preventing traceability using conventional tracking methods. The interface between peripheral networks and the general public is often formed by strategically placed individual actors. Jones (2025) describes the role of "disinfluencers," a term he uses to refer to routine disseminators of misinformation who have a disproportionate influence on trends due to their reach. In the context of the unrest in Southport, it became apparent how such actors function as bridge builders: Narratives forged in alternative networks are picked up by these influencers and popularized on platforms such as X. In doing so, they exploit the "information vacuum" following crisis events and the algorithmic preference for polarizing content. In this case, it was xenophobic disinformation that was disseminated en masse before official bodies could respond.

Despite the technological sophistication of these campaigns, there are indications that their actual persuasive power may be overestimated. Di Marco et al. (2025) argue, based on analyses of information cascades, that coordinated accounts are often inefficiently placed and their network influence is less than feared, as they frequently operate in isolated clusters rather than effectively penetrating organic users. Nevertheless, detection remains a challenge. In their comprehensive survey, Mannocci et al. (2024) point out that the line between legitimate online coordination (e.g., digital activism) and harmful CIB is becoming increasingly blurred, making the development of accurate, non-discriminatory detection algorithms one of the most urgent tasks in current research. In addition, Zhao et al. (2025) emphasize that propagation structures differ significantly depending on platform architecture, which is why detection models trained on only one platform reach their limits in the real, fragmented ecosystem.

4 Analysis and evaluation of current countermeasures

Given the increasing complexity of digital distribution channels and the adaptability of malicious actors, questions of scalability, context dependency, and long-term robustness of countermeasures are becoming the focus of research. Kozyreva et al. (2024) systematize the field in their comprehensive review as a "toolbox" of interventions that can mostly be divided analytically into two main categories: "nudging" approaches that change the decision-making architecture to draw attention to accuracy, and "boosting" approaches that aim to strengthen users' individual skills in the long term. However, the current state of research makes it clear that there is no panacea, let alone a single one; rather, there are clear areas of tension between scalability, context dependency, and unintended side effects.

Strategies for combating disinformation can also be roughly located along a temporal and systemic continuum: preventive measures before exposure, reactive measures during or after exposure, and structural interventions in the design of digital information environments. At the level of user resilience, inoculation theory is experiencing a renaissance. The "prebunking" approach aims to "vaccinate" users preventively against manipulative techniques such as emotional polarization, false dichotomies, or logical fallacies by making typical patterns of such tactics more transparent in advance and making them tangible through examples or games. Meta-analyses show that psychological inoculation can significantly increase the detection rate of manipulative techniques and resilience to disinformation on average (Roozenbeek & van der Linden, 2019; Huang et al., 2024; Lu et al., 2023). At the same time, recent work indicates that the effects are often moderate and temporary and can be amplified or weakened by design decisions in the platform environment. Pennycook et al. (2021, 2024) were able to show that the effect of prebunking can be significantly increased when competence training is combined with situational training. Such incentives, known as "accuracy nudges," explicitly remind users of the importance of accuracy at the moment of sharing. This supports the assumption that measures to strengthen individual competencies should not be considered in isolation, but must be embedded in a decision-making architecture that systematically directs attention away from social confirmation and toward content accuracy.

In the area of reactive countermeasures, "soft moderation" approaches are coming to the fore, which do not delete content but rather mark it, contextualize it, or slow down its dissemination. These include labels, interstitial warnings, downgrading problematic content in rankings, and interventions in forwarding functions that limit visibility without completely removing posts (Douek, 2021; Botero Arcila & Griffin, 2023). Warnings and labels are now considered standard tools, but are viewed ambivalently in research. Pennycook et al. (2020) describe the "implied truth effect" in " ": When only some of the false reports are labeled with warnings, users tend to implicitly perceive unlabeled content as verified and therefore trustworthy (Pennycook et al., 2020).

Against this backdrop, forms of structural friction appear to be a promising addition. On average, harmful content is more emotional, negative, and morally charged than quality journalism, thereby generating "natural" high engagement (Carrasco-Farré, 2022; Jahn et al., 2023). Friction approaches address precisely this reinforcement logic: they make sharing minimally more effortful in order to slow down impulsive reactions without prohibiting content itself or evaluating it in terms of content. Classic examples are additional click steps such as "Read before you retweet," intermediate dialogues ("Are you sure you want to share this content?") Twitter's field experiment with a "read the article before you retweet" prompt led to users opening articles 40% more often and some deliberately refraining from retweeting after reading (Tameez, 2020; Hwang & Lee, 2025). After introducing stricter forwarding limits for "highly forwarded messages" during the COVID-19 pandemic, WhatsApp reported a decline in the virality of such messages by around 70%, which is interpreted as indirect evidence of the effectiveness of this form of friction (TechCrunch, 2020).

The model-based work of Jahn et al. also shows that friction can be understood as a behavioral economic instrument that changes the "choice architecture" in social networks: In an agent-based simulation model, friction alone initially reduces the amount of shared content, while a combination of light friction and learning-promoting elements (e.g., brief references to community standards or news evaluation questions) can significantly increase the average quality of shared content (Jahn et al., 2023; Jahn et al., 2025). From a regulatory perspective, such interventions are discussed as part of a spectrum of "strategic friction" that goes beyond the classic delete/don't delete logic and is intended to curb impulsive misjudgments through small, transparent hurdles without directly curtailing freedom of expression (Laidlaw, 2022).

The analyses of the Integrity Institute, in particular the Misinformation Amplification Tracking Dashboard, complement this picture. They show quantitatively that large platforms systematically amplify disinformation through their engagement-oriented ranking systems, thereby creating incentive structures that reward emotional and polarizing content (Allen, 2022). Although these studies do not evaluate friction interventions per se, they underscore the need to change the "lane" of information dissemination, rather than simply flagging individual "wrong-way drivers." Taken together, the results suggest that even relatively simple friction measures such as additional clicks, prompt dialogues, or forwarding limits can make a measurable contribution to curbing the spread of misinformation, provided they are designed transparently, evaluated, and, if possible, combined with components that promote learning and competence (Jahn et al., 2023; Jahn et al., 2025).

Another fundamentally relevant lever is active algorithmic curation in favor of reliable sources, often discussed under the heading of "authoritative sources." This refers to interventions in ranking and recommendation systems in which platforms not only delete or flag problematic content ex post, but also structurally change the visibility of individual content and actors ex ante. In misinformation research, this is described as a change in the "choice architecture" of digital information environments: recommendation algorithms

are reconfigured so that high-quality, evidence-based information is displayed more prominently and potentially harmful content is systematically downgraded (van der Linden, 2022; Shin et al., 2022; Metzler and Garcia, 2024). The implementation of such mechanisms is particularly far-reaching in the health sector, where wrong decisions can result in immediate physical harm. Against this backdrop, YouTube has developed the concept of "authoritative health sources" in collaboration with the US National Academy of Medicine (NAM) in the wake of the COVID-19 pandemic. A group appointed by NAM developed principles and attributes that platforms can use to identify credible health providers and prioritize them in their content delivery. Examples include public health authorities, academic clinics, and recognized professional associations (Kington et al., 2021). In parallel, the NAM initiative formulated ethical and public health-related criteria for large-scale content labeling, such as transparency requirements, the avoidance of conflicts of interest, and the need to evaluate the effectiveness of such measures based on data (Burstin et al., 2023).

Specifically, this active curation on video platforms is reflected in the introduction of information areas and context boxes that partially overlay search results on sensitive topics. One example is the implementation of information modules on first aid and other acute health topics. When relevant search queries are entered, curated content from verified health organizations is displayed. This differs from the usual sorting optimized for engagement, for example, by involving national emergency services or health authorities (Graham, 2025). This practice is part of a broader platform strategy that YouTube itself describes as "raising authoritative sources": for news, politics, and medical topics, content from established newsrooms and health institutions is algorithmically favored over purely engagement-driven offerings, while popularity metrics continue to dominate in entertainment categories. However, active algorithmic curation is not limited to individual platforms. A comparative study of COVID-19 search results in multiple languages shows that search engines prioritize official government and health websites to varying degrees; overall, however, they frame the pandemic much more strongly through institutional sources than through alternative media (Rovetta and Bhagavathula, 2020). Against this backdrop, reviews on the "health misinformation infodemic" argue that technical solutions should focus in particular on the algorithmic components of social media. The focus is thus on ranking, recommendation, and reinforcement logics, rather than relying exclusively on individual media literacy or classic fact-checking (Rodrigues et al., 2024).

In intervention research, such ranking interventions are viewed with ambivalence. On the one hand, evidence reviews on how to curb misinformation and disinformation argue in favor of designing the digital "choice architecture" in such a way that high-quality sources are more likely to be seen first and problematic content appears less prominently (van der Linden et al., 2022). On the other hand, social and legal analyses warn of new forms of power concentration: if a few private platforms define who is considered an "authoritative source," there is a risk of new gatekeeping structures, potential political bias, and loss of trust among parts of the public (Clemons et al., 2025; Shin, 2022). Discourse analyses

of platform interventions against "fake news" show that companies often legitimize their measures by referring to the protection of health and democracy, while civil society actors emphasize risks to freedom of expression, plurality, and the visibility of marginalized voices. At the regulatory level, the EU Digital Services Act (DSA) requires very large online platforms to identify systemic risks. This includes, in particular, the spread of disinformation. In addition, recommendation systems must be adapted to mitigate these risks. Proposals for audits of recommendation systems emphasize that interventions in ranking logic should be tested, documented, and evaluated over time in a multi-stage, scenario-based process (Meßmer & Degeling, 2023). Overall, the evidence suggests that active algorithmic curation is an important but insufficient component of a comprehensive set of measures that should be combined with user competence, friction, and participatory correction mechanisms.

4.1 Use and effectiveness of current approaches

One of the most intensively researched measures are so-called "accuracy nudges." These are subtle cues that remind users of the concept of truth at the moment of interaction without censoring content. This approach is based on the observation that users often share misinformation not out of malice, but out of carelessness, because their attention is focused on social validation (in the form of likes, etc.) rather than accuracy. Lin et al. (2024) provided the first evidence of the scalability of this approach with a large-scale field study on Facebook and Instagram (N > 33 million). Their results show that content-neutral advertisements that simply ask users to think about accuracy reduced the sharing of misinformation by about 2.6%. Although this effect appears small at the individual level, the authors argue that, given the enormous number of users on these platforms, it can lead to a significant reduction in overall exposure without restricting freedom of expression.

In a recent comparative study of various interventions, Fazio et al. (2025) show that such nudges are a cost-effective method of improving the quality of shared content. However, they also point to heterogeneity in the effect: not all demographic groups respond equally to these nudges; the effect can be lost, especially in the case of highly polarized topics. Critics also note that nudges primarily influence sharing behavior, but do not necessarily change users' underlying beliefs or long-term knowledge. Herzog and Hertwig (2025) therefore emphasize that nudges serve as short-term first aid, but cannot replace the need for more in-depth competence development.

In the area of reactive combating of disinformation, there has been a significant shift in recent years from centralized, expert-based moderation to decentralized, user-based approaches. Particularly prominent is the "Community Notes" model, which was originally introduced by Twitter (now X) as "Birdwatch." However, this development is not limited to X; Meta (Facebook, Instagram) also announced in early 2025 that it would initially restructure its fact-checking program in the US in favor of a similar community-based

system. This strategic realignment, often justified on the grounds of scalability and accusations of bias on the part of traditional fact-checkers, is the subject of controversial debate in current research.

The empirical evidence on the effectiveness of this approach is now robust and shows significant insights into the diffusion dynamics of disinformation. In a comprehensive causal analysis of over 40,000 posts, Slaughter et al. (2025) were able to demonstrate that successfully attaching a community note greatly curbs the spread of misinformation. Their data show that posts received an average of 46.1% fewer reposts, 44.1% fewer likes, and 21.9% fewer replies after a note was published. Time is a crucial factor here: notes that appeared early in the life cycle of a viral post were most effective in breaking the exponential spread curve. These results are supported by Chuai et al. (2024a), who confirm in a differentiated longitudinal analysis that community notes increase the likelihood of a misleading post being deleted by the creator themselves by 103.4% and reduce further dissemination by an average of 62%.

This mechanism works not only at the level of metrics, but also at the level of user psychology. Drolsbach et al. (2024) found in experimental studies that users rate Community Notes as significantly more trustworthy than classic warnings ("flags") set by the platform or experts. The decisive factor here is the explanatory context: since Community Notes not only warn but also explain why information is misleading, they are perceived as a more legitimate corrective across ideological camps. This suggests that the participatory nature of the system can increase the acceptance of fact corrections in polarized environments.

Despite these successes, the model suffers from structural deficits that call into question its suitability as the sole moderation tool. The core problem lies in the "bridging-based ranking" algorithm. For a note to become publicly visible, it must not only be rated as "helpful" by many users, but also by users who have shown different voting behavior in the past (i.e., are ideologically diverse). De et al. (2024) demonstrate in their analysis that this compulsion to achieve cross-camp consensus leads to massive "under-flagging": for 91% of the posts for which a rating was proposed, a rating could never be published because the necessary agreement between the political camps was not achieved. This creates a significant "knowledge gap," in which the most controversial and socially relevant disinformation remains unmarked, while apolitical or trivial false reports (e.g., about consumer products) are quickly corrected.

Chuai et al. (2024b) confirm this scaling problem. Their data show that while the volume of notes written by users is growing steadily, the rate of notes actually displayed is stagnating or even declining relative to the volume of disinformation. Wirtschafter and Majumder (2023) warned early on that such systems are vulnerable to "brigading." This involves political groups coordinating actions in which notes are strategically downgraded ("downvoting") to prevent their publication. This vulnerability makes the system susceptible to

manipulation by well-organized ideological actors who abuse the consensus mechanism as a veto instrument.

Another critical aspect concerns political neutrality and the definition of harm. Renault et al. (2025) examined the distribution of community notes on X and uncovered a clear political asymmetry: posts by Republicans were marked as misleading 2.3 times more often than posts by Democrats. However, the authors were able to prove that this was not due to bias on the part of the evaluators, but rather to the fact that Republican actors actually shared misleading information significantly more often. This leads to the dilemma that a technocratically neutral system produces politically unequal results, which in turn can fuel (unjustified) accusations of censorship by conservative forces.

In the German context, Nenno (2025) shows another facet of this asymmetry: here, posts by certain parties, especially the Greens, are disproportionately often annotated or attacked. Consequently, there is a discrepancy between the suggestion function (which is often used as a tool for political framing) and the publication function (which is often blocked by the consensus algorithm). Potentially more problematic is the "harm blindness" of the system identified by Matamoros-Fernández and Jude (2025). Since community notes primarily focus on factual accuracy, they fail when it comes to content such as dog-whistling, harassment, or hate speech, which cannot be clearly refuted factually but can still be highly harmful.

In light of these shortcomings, experts are advocating a hybrid approach. Traditional third-party fact-checking (3PFC) has proven to be effective in some areas. Meta's own data showed that 95% of users do not click on content if it is accompanied by a professional warning (see SITC, 2025). An integrated model in which accredited experts check high-risk disinformation (e.g., public health, safety) and crowdsourcing is used for the "long tail" of less controversial false reports therefore appears to be the most promising approach.

To overcome the efficiency bottleneck of the consensus mechanism, De et al. (2024) also propose the use of generative AI. Their concept of "supernotes" uses large language models (LLMs) to synthesize the arguments of different, competing draft notes and translate them into neutral, consensus-friendly language. Experiments show that such AI-aggregated notes are significantly more likely to be accepted by diverse user groups and could thus increase the publication rate. At the same time, the research points to socio-psychological side effects. Chuai et al. (2025) concluded that the appearance of a community note not only causes cognitive corrections, but also triggers strong negative emotions ("moral outrage") in the responses to the original post. The note acts as a signal of a norm violation, resulting in social sanctioning of the sender of the disinformation (increase in anger and disgust in the comments). This underscores that community notes are not only informative but also normative tools that shape the social climate on a platform.

In general, the problem of "social correction" remains with community notes and other crowd-based approaches: King et al. (2025) show that while people expect others to intervene against disinformation, they themselves often hesitate to make corrections. An exception is when the sender of the misinformation is personally close to them. This discrepancy between normative expectations and actual behavior limits the potential of purely user-based correction mechanisms and makes it clear that crowdsourcing alone cannot completely replace professional moderation.

As a more sustainable alternative to nudges and moderation, the concept of "boosting" is being discussed, which aims to increase users' media literacy. One prominent approach is psychological inoculation ("prebunking"), in which users are preventively "vaccinated" against manipulative techniques (such as emotionalization, false dichotomies). The APA Consensus Statement (van der Linden et al., 2025) summarizes the comprehensive laboratory evidence confirming that prebunking can increa 's resilience to disinformation. Videos or games that expose manipulative tactics help users recognize these patterns later in real-world content.

However, the theoretical effectiveness of inoculation strategies reaches clear limits when transferred to ecological reality. Wang et al. (2025) uncovered a fundamental discrepancy between cognitive recognition and actual behavior. Experiments with simulated social media feeds showed that, although users were better able to identify the manipulative nature of emotional content and evaluate its accuracy more critically after inoculation, this had little effect on their actual interaction behavior. Despite their knowledge of the manipulation, users continued to spend time with the content and respond to it. This discrepancy highlights the limitations of purely cognitive interventions in the "attention economy": since emotional stimuli such as anger or fear are powerful, impulsive drivers of attention, simply knowing about attempts at manipulation is often not enough to interrupt reflexive interaction or lingering in the feed.

Beyond this behavioral gap, Martini et al. (2025) identified a potentially counterproductive cognitive effect in a large-scale field study with over 2,000 students: Instead of sharpening specific discernment skills, the intervention in the real classroom environment primarily led to an increase in generalized skepticism. As a result, the students paradoxically also lost confidence in valid scientific sources. This suggests that prebunking, if not carefully calibrated, can undermine trust in all information rather than sharpening discernment.

This finding is consistent with a growing number of studies warning of the collateral damage of well-intentioned interventions. Hoes et al. (2024) demonstrated in experiments in the US, Poland, and Hong Kong that common measures such as fact-checking labels and media literacy tips effectively reduce the acceptance of misinformation, but at the same time damage trust in accurate information and democratic institutions. This "skepticism paradox" suggests that interventions that primarily rely on warnings and mistrust can increase citizens' epistemic uncertainty rather than reduce it. When users learn to mistrust everything, even fact-based news loses its persuasive power.

Particular caution is needed with vulnerable groups such as young people. Ma et al. (2025) argue from a developmental science perspective that adolescents respond differently to disinformation than adults due to their social orientation and developing cognitive abilities. Young people are particularly susceptible to social pressure and emotional rewards, which makes them more vulnerable on the one hand, but also offers opportunities for "social norms" interventions on the other. Interventions must therefore be age-appropriate so as not to be counterproductive due to reactance or cognitive overload.

A further development of the classic inoculation approach aims not only to train users to recognize specific manipulation techniques, but also to strengthen their fundamental cognitive dispositions. In this context, Biddlestone et al. (2025) investigated the role of "actively open-minded thinking" (AOT). This refers to a way of thinking that is characterized by actively questioning one's own beliefs and avoiding confirmation bias. Their results show that prebunking interventions are particularly effective when they are expanded to be "norm-based": by conveying to users that open and critical thinking is a socially desirable norm, their motivation to process information accurately increases. In the experiments, this "norm-enhanced" inoculation indirectly led to an improved ability to distinguish disinformation and significantly reduced belief in conspiracy narratives. This implies that successful countermeasures must not be purely technocratic, but must address the epistemic norms and social self-image of users.

Current evidence suggests that no single measure is sufficient to solve the complex problem of disinformation. While algorithmic nudges (Lin et al., 2024) and community notes (Slaughter et al., 2025) can curb viral spread, they do not address the roots of susceptibility. Competence-based approaches (Herzog & Hertwig, 2025) are more sustainable, but carry the risk of promoting blanket mistrust if they are not precisely tailored to the target groups (Hoes et al., 2024; Martini et al., 2025). Future strategies must therefore pursue a hybrid approach: They must combine technical friction to slow down dissemination with educational empowerment, carefully balancing the defense against misinformation with maintaining trust in reliable sources. This is the only way to prevent measures to curb misinformation from inadvertently undermining trust in verified information.

While measures such as nudging and community notes primarily target individual user interaction and aim to change behavior, addressing systemic risks and power asymmetries requires a binding legal framework. This is examined in the following chapter in an international comparison.

4.2 The regulatory framework in international comparison

4.2.1 Current developments in the EU

4.2.1.1 DSA & Code of Conduct on Disinformation

The DSA embodies this process-oriented, systemic approach in a legally binding form. Strowel and De Meyere (2023) show that the DSA organizes the fight against disinformation not through blanket bans, but through graduated due diligence obligations. Particularly for so-called Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), Articles 34 and 35 of the DSA establish a general risk management system that requires platforms to identify, assess, and mitigate systemic risks to fundamental rights, democratic processes, and public safety. This explicitly includes risks related to disinformation and manipulative influence campaigns (Husovec, 2024; Eder, 2024).

Ó Fathaigh et al. (2025) point out that the DSA deliberately chooses an indirect regulatory model: instead of defining "disinformation" as a separate category with deletion obligations, it requires platforms to design procedures and organizational structures in such a way that the spread of harmful but often legal content is curbed by risk-based measures. In the literature, this is referred to as "lawful but awful" disinformation. This includes transparency requirements regarding recommendation algorithms, access to data for researchers, reporting and complaint systems, and cooperation with "trusted flaggers" who are qualified as trustworthy whistleblowers (Strowel & De Meyere, 2023; Van de Kerkhof, 2025). Eder (2024) interprets this system as an attempt to establish a positive feedback mechanism in which iterative risk assessments, independent audits, and multi-stakeholder participation gradually lead to refined standards for dealing with systemic risks.

The UNESCO guidelines on the governance of digital platforms provide normative support for this approach. They recommend that regulation should focus primarily on structures, processes, and power asymmetries in the digital communication space, rather than empowering authorities to define the truthfulness of individual content (UNESCO, 2023). Core principles include transparency, accountability, user empowerment, and the protection of journalistic and scientific work. The guidelines explicitly warn against legal bans on "fake news," whose vagueness can lead to overblocking, arbitrary enforcement, and restricted public debate.

The first enforcement decisions show that the DSA is not just pursuing this model on paper. Since 2023, the European Commission has opened several formal proceedings against large platforms such as TikTok, Meta, and X. The focus is on systemic risks to electoral processes, the protection of minors, and the handling of harmful content (European Commission, 2024a; Husovec, 2024). On December 5, 2025, the Commission imposed a fine of €120 million on X for the first time because the company had disregarded

its transparency obligations, used deceptive design, and failed to grant researchers sufficient access to public data (European Commission, 2025). The sanction does not target individual controversial content, but rather structural deficits in risk management and disclosure practices. It thus follows the nature of process-based regulation.

The landscape of platform regulation is undergoing a period of fundamental strategic tension. While European legislators are increasing the pressure to take responsibility with the Digital Services Act (DSA) and the transformation of the Code of Practice on Disinformation into a binding Code of Conduct, the very large online platforms (VLOPs) are making a tactical retreat in the opposite direction. This divergence between regulatory requirements and business practice is manifested not only in the erosion of voluntary commitments, but also in a profound restructuring of moderation paradigms, which are increasingly driven by cost efficiency and political risk minimization.

The EU Code of Practice on Disinformation, which was designed as a co-regulatory bridge instrument to prepare for DSA compliance, lost its binding force in operational reality. A detailed quantitative analysis by Democracy Reporting International (DRI) shows that the major platforms have significantly reduced their commitments during the transition period from 2022 to 2025. On average, signatories removed around 31% of their original commitments from their reports without replacement (Alvarado Rincón & Meyer-Resende, 2025). This withdrawal has a particularly serious impact on the ecosystem of external validation: support for the independent fact-checking community fell by 64%. This reduction in resources is diametrically opposed to the growing need for verified information in years of important international elections and suggests that platforms increasingly view cooperation with external review bodies as a cost factor that is detrimental to business or a political risk.

The implementation also reveals massive deficits in terms of quality. The European Digital Media Observatory (EDMO) evaluation for the first half of 2024 characterizes the VLOPs' compliance reports as inconsistent and analytically superficial. Botan and Meyer (2025) particularly criticize the lack of granular data at the individual member state level. Platforms such as Meta (Facebook/Instagram) and TikTok did not provide sufficiently disaggregated data on the spread of disinformation in specific language areas, which means that structural risks in smaller EU states remain statistically invisible and it is impossible to verify the effectiveness of countermeasures. These data gaps should not be interpreted as mere negligence, but as a strategic lack of transparency that systematically hinders independent auditing of algorithmic risks, as required by Article 34 of the DSA. The German-Austrian Digital Media Observatory (GADMO) also states in its 2024 annual report that, despite formal commitments, numerous cases of illegal political advertising and unlabeled deepfakes have been documented on TikTok and X. This points to a discrepancy between formal compliance with rules and the actual usability of complaint mechanisms (Wegner, 2024).

Parallel to the erosion of self-regulation, a strategic shift in moderation methods can be observed, moving away from the ideal of curated safety ("Safety- -by-Design"). Particularly under the influence of political pressure in the US, where conservative forces increasingly frame moderation as censorship, platforms are adopting a "hands-off approach." Meta, for example, has signaled its intention to reduce its collaboration with professional fact-checkers in the US in favor of user-based models such as "Community Notes." From an expert perspective, this development could also influence integrity policy in Europe (e.g., Windwehr, 2025). However, this trend toward the "democratization of truth" through crowdsourcing, as established on Platform X (formerly Twitter) as the primary corrective, carries significant risks, as already detailed in Chapter 4.1.

Another area where regulation is not sufficiently effective is the handling of political advertising. Despite official bans or strict transparency rules on platforms such as TikTok and Meta, research documents massive circumvention strategies. Actors systematically exploit "gray areas" by disguising political messages via influencers ("fin-influencers" for politics) or as organic "news" posts that are not detected by automated advertising filters. The Institute for Strategic Dialogue (ISD, 2024) was able to show that Meta allowed thousands of ads spreading explicit election lies in the run-up to elections because the technical detection systems were circumvented by simple concealment tactics. This illustrates that self-regulation of advertising marketplaces remains ineffective without external auditing.

The impact of these governance deficits varies greatly depending on the technical architecture of the services. TikTok is particularly in the spotlight here due to its asymmetric algorithms. In a comprehensive audit of the 2024 US election, Ibrahim et al. (2025) demonstrated that the platform's recommendation systems do not operate neutrally, but systematically favor Republican and polarizing content. In the experiment, neutral user accounts received 11.5% more right-wing content than left-wing content, evidence of active algorithmic bias in the discourse. This effect is amplified by the specific demographics of the platform: Tjaden et al. (2025) argue that TikTok users are less skeptical of misinformation, making the platform an ideal vector for "short-form" propaganda.

On Instagram, on the other hand, the problem lies less in obvious algorithmic bias than in the discrepancy between formal DSA compliance and practical usability. Sekwenz et al. (2025) analyzed the platform's reporting and complaint channels and found that they are often hidden behind "dark patterns." While Instagram theoretically offers the legally required appeal options, these are hidden so deeply in nested menus in the UX design that they effectively lead nowhere. This results in a "compliance facade" in which regulatory requirements are technically met but are undermined in practical use.

In the sensitive area of health information, YouTube offers a counterpoint to the frequently postulated restraint in editorial intervention. For search queries on acute emergencies, content from verified partner organizations is prominently prioritized. As part of a global initiative implemented in Germany and Canada, among other countries, the platform

relies on an approach of active curation through so-called "information shelves." For search queries relating to acute medical emergencies, such as heart attacks, CPR, suicide prevention, or opioid overdose, the platform intervenes in the results list. In these areas, compact content from pre-verified, "authoritative" partner organizations such as the Red Cross or clinics is given prominent priority, rather than leaving the display primarily to engagement algorithms optimized for dwell time (Graham, 2025; YouTube, 2025).

This significant editorial intervention demonstrates that platforms are technically and operationally capable of responsibly fulfilling their role as "gatekeepers." This is particularly successful when a broad social consensus, in this case the protection of life, outweighs the economic pressure to maximize engagement. Mohamed and Shoufan (2024) demonstrate that this curated approach is also accepted by users: in their study, 87.6% of recipients found this prioritized content helpful for their decision-making. The approach in the public health sector thus stands in stark contrast to the strategic move away from strong content guidelines in more politically controversial areas.

A key obstacle to scientific analysis and political containment of these developments remains access to relevant data. Allen et al. (2025) emphasize in their Global Transparency Audit that the platforms' current transparency reports are often incomplete and do not allow for independent verification of risk assessments. Researchers are therefore calling for new legal frameworks that go beyond voluntary data donations. The UK Data (Use and Access) Act 2025 and Article 40(12) of the DSA represent important milestones in this regard, as they grant accredited researchers a legal right to access online safety data for the first time. This enables the transition to independent "trace research," which can systematically track moderation decisions such as shadow banning or deletions, rather than having to rely on the curated data sets provided by the platforms.

4.2.1.2 European Democracy Shield

Complementing the DSA, the EU has created the EDS, an instrument that specifically responds to hybrid external threats. On November 12, 2025, the European Commission and the High Representative presented the joint communication "European Democracy Shield" (EDS), which aims to promote strong and resilient democracies in the EU. The initiative is justified by an increase in internal and external threats. The information space has become a central arena for geopolitical conflicts, where authoritarian regimes such as Russia wage an "asymmetric" battle through hybrid attacks, destabilization, and FIMI (Foreign Information Manipulation and Interference) campaigns to undermine citizens' trust in democratic institutions.

The EDS builds on existing measures such as the European Democracy Action Plan and the Defense of Democracy package and is structured around three central pillars:

1. Strengthening situational awareness and protecting the integrity of the information space.
2. Strengthening democratic institutions, free and fair elections, and free and independent media.
3. Promoting societal resilience and civic engagement.

One of the core measures is the creation of a new European Center for Democratic Resilience. This center will serve as a central hub for information exchange, operational cooperation, and capacity building between EU institutions and member states. This will be flanked by a stakeholder platform that involves civil society actors such as fact-checkers and researchers.

In addition, the communication announces a number of other actions, including the establishment of a "European Network of Fact-Checkers," the expansion of the mandate of the European Digital Media Observatory (EDMO), the preparation of a "DSA incidents and crisis protocol," and a "Media Resilience Program" to strengthen journalism. Funding will primarily come from existing and future EU programs such as Creative Europe, Digital Europe, CERV, and the proposed AgoraEU program.

The publication of the European Democracy Shield met with a mixed response, ranging from institutional approval to fundamental criticism of its strategic orientation. While actors from the immediate environment of the European External Action Service (EEAS), such as *EUvsDisinfo*, classified the initiative as a necessary "first line of defense" in the "geopolitical information war" against authoritarian regimes and emphasized its proactive approach, the package of measures was assessed in a much more differentiated and sometimes critical manner by external experts.

A central point of discussion concerns the perceived discrepancy between the description of the problem and the proposed solutions. Critics complain that although the Commission identifies serious challenges such as foreign interference and polarization, the measures derived from this fall short of expectations. The reception in the trade press suggests that the initiative was received with caution when it was presented. This caution is attributed in particular to the non-binding nature of many of the more than 40 action points identified, which are often based on soft wording such as "support" or "strengthen" rather than establishing binding commitments. In addition to the limited legislative powers of the EU's European Union Agency for Cybersecurity ( ), geopolitical considerations are also discussed as the cause of this ambition gap. Reports indicate that the draft was watered down due to concerns that the regulation could be interpreted in the US as a restriction of freedom of expression, as well as pressure from US technology companies.

In terms of content, the strong focus on foreign information manipulation (FIMI) is seen as problematic. Analysts note that this focus is "oversized" and neglects the relevance of domestic sources of disinformation. More fundamental criticism is expressed by civil

society actors such as The Future of Free Speech (2025). They point to the problem discussed in Chapter 2.4 of weak empirical evidence on the causal effect of disinformation and argue that its prevalence and influence are often overestimated. The prevailing discourse bears the hallmarks of a "moral panic." This attitude underscores the danger, also discussed in Chapter 2.4, that exaggerated warnings about disinformation could paradoxically undermine trust in democratic institutions more than the misinformation itself. In line with this logic, the instrumental role of fact-checkers is also critically questioned; while industry representatives welcome the support, fact-checking is assessed as a measure of limited effectiveness that does not address structural problems (see chapters 2.5 and 4.1).

From a sovereignty policy perspective, it is also argued that although the Shield emphasizes the EU's regulatory competence, it does not create sufficient "strategic autonomy" in the area of digital infrastructure. Dependence on non-European platform operators remains, which limits the enforceability of European standards. Industry associations such as the Association of Commercial Television and Video on Demand Services in Europe (ACT, 2025) welcome the initiative in principle as timely, but link their approval to the demand for consistent enforcement of the existing legal framework (DSA, DMA, EMFA). In order to ensure media diversity in the long term, they argue that a reorientation of the advertising market is also necessary, providing economic incentives for quality journalism over disinformation. Similarly, fact-checking associations warn that the planned financing instruments (e.g., AgoraEU) are insufficient in view of the massive investments made by antagonistic actors in disinformation campaigns (EFCSN, 2025b).

4.2.2 International comparison

The global regulatory landscape for combating disinformation is characterized by a fundamental dichotomy. While the European Union is pursuing a process-oriented approach with the Digital Services Act (DSA), which focuses on transparency, systemic risk minimization, and cooperative responsibility, a content-centered, criminal law paradigm is establishing itself in many countries in Asia, Africa, and Latin America. Mostly referred to as "fake news" or "misinformation laws," these regulations are increasingly being critically analyzed in academic literature as instruments of "lawfare." This refers to the strategic use of laws to suppress political opposition and civil society (Mahapatra et al., 2024; Cho, 2025; Bradshaw et al., 2025). In parallel, international organizations such as UNESCO, the OSCE, the Inter-American Commission on Human Rights, and the United Nations have formulated human rights guidelines. These warn against vaguely worded bans on "fake news" and instead advocate for systemic, rule-of-law-compliant, and multi-stakeholder-based governance (UNESCO, 2023; OSCE Representative on Freedom of the Media et al., 2017; Organization of American States [OAS], 2019).

Current empirical work shows that a global trend toward codifying disinformation regulation has emerged over the past decade. Bradshaw et al. (2025) document in a global

survey that between 2010 and 2022, 80 countries enacted new laws against disinformation or tightened existing standards in this regard. Based on a data set from 177 countries and 105 legal acts, the authors conclude that although these standards are formally justified as a means of combating "misinformation," in fact they vary greatly in their design: They range from media law transparency requirements and rules for online political advertising to criminal offenses punishable by imprisonment, which have a significant chilling effect on press freedom. The analysis conducted for the Center for International Media Assistance (CIMA) reached similar conclusions. Based on 105 laws, CIMA identifies four categories of sanctions, namely excessive fines, imprisonment, deletion requirements, and administrative and bureaucratic requirements. The evaluation shows that in many cases these instruments are used specifically against journalists (Lim & Bradshaw, 2023).

In large parts of South and Southeast Asia, a regulatory practice has become established that primarily frames disinformation as a security threat to state order. In their analysis for the GIGA Institute, Mahapatra et al. (2024) describe how governments in this region use vaguely defined laws against misinformation to criminalize critics. The term "fake news" is often defined so imprecisely in legal terms that it also encompasses legitimate expressions of opinion or investigative journalism, provided they contradict the government line. This practice of "lawfare" results in journalists and activists becoming embroiled in lengthy, costly court cases that serve less to establish the truth than to financially and psychologically exhaust the defendants.

Cho (2025) shows that illiberal regimes in Southeast Asia use "fake news" strategically to either actively combat a public perceived as hostile or, in passive mode, to deliberately reinforce pro-government narratives. Using the examples of Singapore and the Philippines, she develops a typology in which the protection of allegedly "vulnerable" population groups or the defense against "foreign influence" is used as an argument to legitimize draconian powers to remove content, order counterstatements, and prosecute (Cho, 2025). Putra (2024) confirms this insight for the ASEAN region and points to a structural dilemma: the principle of "non-interference" upheld in the region makes it difficult to coordinate transnational efforts to combat disinformation, while national laws are often used selectively against domestic political opponents. In countries such as Thailand and the Philippines, anti-disinformation laws are often linked to allegations of treason or endangering national security, which lowers the threshold for state intervention.

This shift toward security-centered and repressive responses can also be observed outside Asia. Several African states have created criminal offenses that make the "dissemination of false information" punishable by imprisonment and, in practice, are disproportionately applied to opposition voices and critical media (Omilusi, 2025). Research by the United Nations University Centre for Policy Research also shows that disinformation narratives in conflict regions of sub-Saharan Africa are often intertwined with ethnic tensions, electoral violence, and the delegitimization of international peace missions, while legal

countermeasures often fail to strike a balance between security and freedom of expression (Albrecht et al., 2024).

The coronavirus pandemic has often acted as an accelerator, as analyses from Latin America and (South) East Asia show. Emergency decrees expanded executive powers and allowed governments to brand critical reporting on crisis management as potentially "dangerous disinformation" (Asia Centre, 2021; Konrad-Adenauer-Stiftung, 2022).

A global analysis by the Center for News, Technology & Innovation (CNTI) underscores that this phenomenon is not specific to any one region. A study of legislation in over fifty countries found that laws explicitly targeting "fake news" risk doing more harm than good (CNTI, 2024). The lack of distinction between unintentional misinformation and deliberate disinformation enables authoritarian and hybrid-democratic governments to monopolize the interpretation of "truth" and restrict press freedom through the threat of imprisonment or excessive fines. Lim and Bradshaw (2023) refer to this development as "chilling legislation": the mere existence of such laws has a deterrent effect, leading journalists to practice preventive self-censorship in order to avoid legal reprisals. Based on an analysis of 105 legal acts, the authors show that almost one-tenth of all journalists imprisoned worldwide in 2022 were detained on the basis of disinformation laws.

International human rights institutions are therefore increasingly scrutinizing the compatibility of such norms with international law standards. 2017, the OSCE, the UN Human Rights Council, the African Commission on Human and Peoples' Rights, and the Inter-American Commission on Human Rights issued a joint statement warning against laws that criminalize "fake news" or "propaganda" across the board and could thereby stifle legitimate criticism (OSCE, 2017). In a legal opinion on the French "Loi contre la manipulation de l'information," the OSCE Representative on Freedom of the Media also assessed the risk that broad powers to delete alleged disinformation during election campaigns could lead to overblocking and curtail the plurality of political voices (OSCE, 2019).

In contrast to this is the approach taken by the European Union and, in particular, the Baltic states, which focuses less on banning individual content and more on strengthening social resilience. Balčytienė et al. (2025) analyze the "Baltic Way" as a people-centered approach that translates historical experiences with Russian propaganda into a comprehensive security strategy. In Lithuania, Latvia, and Estonia, disinformation is not viewed in isolation as a media problem, but as a hybrid threat that must be addressed with a "whole-of-society" approach. This involves the close integration of state strategic communication, independent media, fact-checking organizations, and an enlightened civil society ("civic preparedness"). Instead of deleting content, the focus is on increasing media literacy and promoting a pluralistic information ecosystem that is more immune to external manipulation (Balčytienė et al., 2025).

In their global comparative study, Berger et al. (2024) emphasize that this approach to promoting resilience is also being successfully practiced in other regions, such as Taiwan

and Finland. The key feature here is the shift away from a state monopoly on truth towards decentralized responsibility, with platforms, civil society, and the state acting cooperatively without the state acting as the sole arbiter of truth. In Latin America, the Inter-American Human Rights Protection and related institutions also recommend not relying primarily on criminal law instruments, but rather promoting transparency, data protection, political advertising transparency, and the strengthening of independent media as central building blocks of a resilient information space (OAS, 2019; Pérez Argüello & Barojan, 2019).

Variants of a resilience and process model can also be found outside Europe. In her overview of the legal situation in the US, Gielow Jacobs (2022) shows that the very far-reaching protection standards of the First Amendment severely limit criminal prohibitions on "fake news." Instead, forms of co-regulation dominate, such as transparency requirements for online political advertising, voluntary commitments by platforms, and media literacy support programs. Although these instruments are not free from criticism and their effectiveness and dependence on platform cooperation remain controversial, the comparison illustrates that liberal democracies can attempt to curb disinformation through institutional arrangements, civil liability regimes, and social resilience strategies even without criminal prohibitions (Gielow Jacobs, 2022; Bradshaw et al., 2025).

Ukraine occupies a specific intermediate position, operating under extreme conditions due to Russia's war of aggression. Marushchak et al. (2025) show that the Ukrainian regulatory framework is forced to combine elements of resilience promotion with tough, security-oriented measures. In view of a significant wave of AI-generated disinformation, such as deepfakes of political decision-makers or synthetic propaganda, Ukraine has adapted its legislation to cover the dissemination of such content as part of hybrid warfare. The authors argue that in an existential conflict, state regulation of AI content, close cooperation with platforms in the rapid identification of enemy operations, and measures to seal off the information space are indispensable (Marushchak et al., 2025).

At the same time, the requirement to uphold fundamental liberal principles as far as possible remains. Ukrainian institutions are working closely with international partners to develop protective mechanisms against Russian influence operations without disempowering independent media and civil society actors (Berger et al., 2024). The case illustrates that the liberal resilience model can reach its limits in acute threat situations and must be supplemented by defensive measures to protect the integrity of the information space. The challenge lies in not perpetuating special security regulations, thereby enabling a transition back to a more process-oriented normal regime.

In summary, a clear dividing line can be drawn in the global governance of disinformation. While the European model, embodied by the DSA, focuses on process regulation, risk management systems, algorithmic transparency, the disclosure of advertising practices, and access to data for supervision and research are being reviewed. In many countries of the Global South and in autocratic contexts, however, content-oriented approaches

dominate. These aim to ban or criminally sanction certain statements, the definition of which often remains vague, thus enabling political abuse (Mahapatra et al., 2024; Cho, 2025; Omilusi, 2025).

Bradshaw et al. (2025) make it clear that the rapid global increase in misinformation laws can be explained less by an actual increase in false information. Rather, political opportunity structures are the decisive factor, even though false information has always existed historically: These include the popularization of the "fake news" discourse by elites, the security policy framing of information phenomena, the interest of governments in controlling the flow of information, and the increasing visibility of platform failures in the public sphere. In this constellation, disinformation easily becomes a projection screen for deeper conflicts over power, legitimacy, and social plurality.

In contrast, international law and regional human rights standards formulate relatively clear criteria: interference with freedom of expression must be precisely defined by law ( ), necessary and proportionate, and serve legitimate aims such as the protection of national security, public order, or the rights of others (OSCE, 2017; OAS, 2019; UNESCO, 2023). Both the OSCE and the Inter-American Commission on Human Rights therefore expressly warn against general clauses against "fake news" that do not meet these requirements and de facto function as instruments of censorship (OSCE, 2019; OAS, 2019).

Against this backdrop, the DSA can be seen as an attempt to establish a third way between a state monopoly on truth and largely unregulated platform markets. It shifts regulatory attention away from individual posts to the structures and processes of platforms, relies on multi-stakeholder mechanisms, and links transparency, accountability, and cooperative self-regulation with regulatory oversight and sensitive sanctions for systemic non-compliance (Strowel & De Meyere, 2023; Griffin, 2025; Husovec, 2024). The first fine imposed on X in December 2025 illustrates that violations of transparency and cooperation obligations in particular are understood as serious risks to democratic public life. One example is the obstruction of scientific research (European Commission, 2025).

In their global study on science communication, Mede et al. (2025) point out that trust in information depends heavily on how transparent and traceable the sources and underlying processes are. The European approach attempts to restore this trust through systemic accountability by forcing platforms to disclose their internal mechanisms and engage in dialogical risk assessments. Repressive approaches, on the other hand, often replace trust with enforced conformity: they create a seemingly "calm" information environment in the short term, but in the long term they undermine the credibility of state institutions, the innovative power of the digital space, and society's ability to distinguish between reliable and unreliable information (Lim & Bradshaw, 2023; Bradshaw et al., 2025).

Empirical evidence from 2024 and 2025 thus suggests that process regulation and resilience promotion are more compatible with democratic and human rights standards in the

medium term than content-focused, criminal law instruments. At the same time, a look at conflict contexts such as Ukraine or states with extremely polarized publics shows that process-oriented models also need complementary protective mechanisms in order to remain capable of action in acute crisis situations. The central question for the coming years will therefore be whether it will be possible to stabilize in practice the balance between protection against disinformation and protection against repressive interference with freedom of expression that is laid down in the DSA and in international guidelines. Another question is whether this model can serve as a reference framework for more equitable and rule-of-law-based platform regulation beyond Europe.

5 Conclusion

A comprehensive examination of disinformation phenomena and the associated online harm suggests that holistic and structural approaches are needed to effectively address them. The focus should shift from individual verification of truthfulness to procedural and systemic risks, especially for content that does not explicitly violate content standards ("lawful but awful").. Instead, the focus should be on the potential for harm through dissemination patterns, network dynamics, and systemic risks, which makes the identification of actor-centered rather than content-based measures a priority.

The key findings from the analysis confirm that the spread of disinformation is not a purely random occurrence. It is deeply rooted in the attention economy, with the dominant platform business model of maximizing dwell time through algorithmic preference for emotional and polarizing content creating misguided incentives. This algorithmic amplification can systematically favor certain political actors and reward content that triggers anger and hostility with a reach bonus. Another often underestimated driver is the opaque ad tech ecosystem, which unintentionally finances the "disinformation-for-profit" industry by placing programmatic advertising on low-quality or false content. At the same time, the mass availability of generative AI is changing the conditions under which disinformation is produced. In addition to the danger of perfect deepfakes, there is also a systemic threat in the flooding of the information space with low-quality, AI-generated content ("AI slop"). This content can lead to the dilution of the information and knowledge space and significantly increase verification costs.

Research into the mechanisms at work shows that users' vulnerability is not primarily due to a lack of knowledge, but rather to psychological dispositions such as "identity-protective motivated reasoning," which leads to the acceptance of false information if it supports one's own group identity. This vulnerability is exacerbated by phenomena such as the "news finds me" effect, in which passive reception via feeds inhibits critical examination of sources. The damage manifests itself not only in the digital space, but can also lead to physical violence (stochastic terrorism) and structurally undermine democratic participation, for example through coordinated campaigns (networked misogyny) to "silence" women in public positions.

In light of these systemic challenges, the study suggests that the global regulatory dichotomy must be taken into account. While the European model, embodied by the Digital Services Act (DSA), takes a process-oriented approach aimed at algorithm transparency and systemic risk reduction, it is contrasted by content-focused, often repressive "fake news" laws in autocratic and illiberal contexts, which can actually be used as instruments of "lawfare" to suppress critical voices. The European model aims to ensure the integrity of the information space by strengthening the systemic auditability of Very Large Online Platforms (VLOPs), with sanctions also being imposed on platforms that fail to cooperate with researchers.

Priority policy options are derived from these analysis results. First, there must be a stronger focus on the ad tech ecosystem in order to eliminate the economic disincentives for spreading disinformation. Second, international coordination in defining and detecting coordinated inauthentic behavior (CIB) by geopolitical actors must be improved. Third, the DSA must consider banning "manipulative dissemination techniques." In the area of interventions, it is clear that there is no silver bullet. While accuracy nudges are a scalable method for reducing divisive behavior, and community notes act as a trustworthy user-based corrective, the latter remain limited in their scalability due to the need for cross-party consensus on controversial issues. Inoculation strategies (prebunking) must also be used with caution, as they can lead to the skepticism paradox if improperly calibrated. In this case, users lose trust in all sources, including reliable ones.

Despite the synthesis achieved, important questions remain open for future research. A central methodological problem is the still limited empirical knowledge on causality and exposure to disinformation. Closely linked to this is the need to clarify how data access for research can be sustainably improved in order to ensure independent auditing of algorithmic mechanisms. Finally, further research is needed to effectively assess the economic costs and benefits of various countermeasures. This involves examining how approaches ranging from platform regulation to market mechanisms can be classified and what balance between international coordination and regulatory fragmentation is appropriate in each case. Addressing these issues will determine whether the opportunities offered by AI (e.g., in detection) and digital education can be systematically exploited in such a way as to strengthen the resilience of the democratic public sphere.

6 Bibliography

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