Media and Communication
2026 • Volume 14 • Article 12499
https://doi.org/10.17645/mac.12499

cogitatio

ARTICLE

Open Access Journal

Threat and Remedy: AI's Technological and Normative Role in Democratic Discourse and Counter Speech

Diana Rieger and Mario Haim

Department of Media and Communication, LMU Munich, Germany

Correspondence: Diana Rieger (diana.rieger@ifkw.lmu.de)

Submitted: 31 March 2026 **Accepted:** 8 July 2026 **Published:** 3 September 2026

Issue: This article is part of the issue "The Role of AI for Counter Speech: Detection, Intervention, and Risks" edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at https://doi.org/10.17645/mac.i500

Abstract

Artificial intelligence (AI) has fundamentally transformed online discourse, serving simultaneously as a source of and a potential solution to threats to democracy. This article conceptually examines AI's multifaceted role in digital public spheres, beginning with an analysis of how AI shapes democratic processes online, ranging from information curation and political participation to dialogue facilitation and the construction of epistemic infrastructures. We identify several key risks: AI amplifies harmful content through automated content generation, coordinated manipulation, and the algorithmic mainstreaming of borderline content that often evades traditional detection systems. Within this broader democratic context, we position counter speech as a critical intervention strategy and examine both reactive measures (e.g., automated detection and content moderation) and proactive approaches (e.g., prebunking, algorithmic downranking, friction design, and AI-mediated dialogue). We argue that effective counter speech increasingly relies on human–AI collaboration, where AI supports human counter speakers with factual resources, emotional scaffolding, and scalability, while human actors preserve the authenticity and agency that make counter speech normatively meaningful. However, realizing the potential of such collaboration requires moving beyond technological fixes toward democratically legitimated governance structures. Drawing on examples from Wikipedia and decentralized platforms, we demonstrate how transparent and participatory institutional arrangements can foster more resilient discourse environments. We conclude that societies can harness AI's democratic potential while mitigating its risks only by proactively establishing legitimate deliberative processes grounded in broadly shared discourse norms.

Keywords

artificial intelligence; content moderation; counter speech; democratic discourse; online hate speech; online radicalization

1. Introduction

Scholarly engagement with the internet has produced a persistent dilemma. On one side are findings emphasizing broader access to information, more multimodal and thus more inclusive forms of content presentation, and the relative ease of participation (e.g., Papacharissi, 2004). From this perspective, the internet holds the potential to facilitate constructive and wide-ranging public exchange. On the other side are findings showing that participation and attention remain concentrated among resource-rich actors (e.g., Hindman, 2008). Anonymity and the frequent absence of effective law enforcement have, in turn, been implicated in the development of a "polluted information ecosystem" (Wardle, 2018, p. 951). From this perspective, the internet may serve not as a catalyst for democracy but as a force that stagnates or even undermines it. In response to these concerns, the European Union (EU) emerged as one of the earliest actors to pursue comprehensive regulatory approaches to internet governance (e.g., Klausa, 2023).

More recently, however, this assessment has been complicated by a series of structural reversals. Major intermediaries have increasingly moved away from actively supporting law enforcement efforts online. Instead, they have invoked libertarian principles, reasserting the position that, as non-authors, they bear no responsibility for the content circulating on their platforms. This shift has been accompanied by changes in content governance practices, including the replacement of professional fact-checking mechanisms with so-called community notes. This trend is perhaps most visible in the transformation of X (formerly Twitter) under Elon Musk's ownership, which has been characterized as fostering "platform illiberalism" and concentrating disproportionate influence over public opinion formation in the hands of a single owner (Neuberger, 2026).

Wardle (2018) conceptualizes the online information environment along two axes: veracity and intent to harm. Whereas misinformation refers to false content shared without intent to deceive, and malinformation involves accurate content deployed to cause harm, disinformation, the guiding concept in this article, combines falsity with deliberate intent. As the veracity of online content is increasingly challenged by the widespread adoption of generative artificial intelligence (AI), the dimension of harmful intent becomes particularly relevant in relation to hate speech, polarization, and radicalization (Rieger et al., 2024). Following Wardle (2018), we use the term "harmful content" as an overarching category encompassing disinformation, hate speech, and conspiracy content. Although these phenomena differ in their mechanisms, social consequences, and legal implications, they share an important feature: All are common targets of counter speech interventions. It is this shared characteristic that justifies considering them together in the present analysis.

Harmful content likely conflicts with several EU regulatory frameworks (Farrand, 2025). More important for the purposes of this article, however, is its impact on the quality and climate of online debate (Haim & Neuberger, 2022). Contemporary communication research increasingly characterizes the internet as a fragmented and cacophonous arena of opinion exchange marked by declining discourse quality. This assessment is supported by empirical evidence showing that in the absence of regulatory intervention or active content moderation, levels of hate speech on online platforms are substantially higher (Hickey et al., 2025; Rieger et al., 2021).

It is within this deteriorating discursive environment that AI has emerged as a double-edged force that not only accelerates the production and dissemination of harmful content but also offers tools capable of supporting counter speech at an unprecedented scale. Counter speech has gained considerable scholarly and practical traction in recent years as a promising speech-based alternative to purely regulatory responses to online harm (Benesch et al., 2016; Obermaier et al., 2025). Rather than removing harmful content, counter speech seeks to address harmful expression through additional speech, thereby preserving freedom of expression while mitigating the negative effects of hate speech and disinformation (Benesch et al., 2016). Like many other communicative practices, however, counter speech is increasingly being reshaped through AI-driven automation (Wu et al., 2025). AI applications are therefore transforming the conditions under which democratic discourse takes place.

Throughout this article, AI is understood as a socio-technical construct oriented toward the machine-based completion of tasks that would ordinarily require human intelligence (cf. Russell & Norvig, 2020, p. 19). More specifically, this article attends to AI as it manifests across diverse applications that shape digital debate and democratic discourse. Democratic discourse is understood here as the exchange of politically relevant information that contributes to individual opinion formation. This occurs directly through discussions on social networking platforms, comment sections of traditional mass media websites, AI systems such as ChatGPT, and messaging apps such as WhatsApp or Telegram. It also occurs indirectly through AI-supported discoverability of political information, such as through search engines, content aggregators, and algorithmic recommendation and filtering systems.

The communicative transformation at stake lies in the fact that AI-supported applications have evolved from predominantly curatorial tools into actors that increasingly function as recipients and producers of communication. AI is consequently positioned simultaneously as author, putative solution, and ostensibly neutral intermediary within these dynamics. This makes public deliberation about AI's appropriate role particularly contentious (Friemel & Neuberger, 2023; Jungherr & Schroeder, 2023). In this context, AI and human actors increasingly constitute complex, interdependent socio-technical systems whose collective outcomes cannot be reduced to either human or machine behavior alone. As a result, the governance of digital public spheres differs fundamentally from traditional forms of media regulation (Tsvetkova et al., 2024).

This article therefore pursues the core goal of discussing AI's role in democratic discourse in light of its capacity both to facilitate and to counter harmful content. Particular attention is given to counter speech as one of the most promising responses to AI-amplified threats. The article proceeds in four steps. First, we discuss and contextualize the threat potentials that AI poses to democratic online discourse. Second, we examine available response options by distinguishing between direct and indirect forms of intervention. Third, we focus on counter speech, understood here as deliberate communicative responses to harmful content, and evaluate AI's role as both obstacle and enabler of effective counter speech. Fourth, we consider the normative and governance conditions necessary for the democratic legitimation of AI-assisted counter speech. Methodologically, the article adopts a structured conceptual synthesis, integrating empirical findings from communication science, human–computer interaction, and AI studies into a coherent analytical framework. Our aim is theoretical integration and normative clarification rather than systematic meta-analysis.

2. Threat Potentials

Given the proliferation of AI-supported applications, AI's role in discursive processes has become increasingly multifaceted. Whereas earlier developments primarily concerned the curation of information (e.g., through search engines), more recent applications of generative AI have increasingly assumed a productive function, directly generating discursive content. This is evident in the AI-supported manipulation of electoral processes and, more broadly, in the large-scale production and dissemination of false, misleading, or intolerant content (e.g., Pamment & Tsurtsumia, 2025). AI may function as both author (e.g., through deepfakes) and amplifier (e.g., through social bots or astroturfing) of such content (García-Orosa, 2022; Mantello et al., 2023). Generative AI has also been deployed to accentuate visual characteristics associated with extremist ideologies (e.g., Aryan phenotypes; Hiller & de las Casas, 2025). Content generated or amplified in these ways often aligns closely with the filtering and recommendation architectures of social media platforms and search engines, which privilege emotion, surprise, and personalization. As a result, such content can achieve disproportionate visibility and diffusion, not only through AI actors but also through subsequent amplification by human users and institutional actors, such as news organizations. For instance, YouTube's recommendation algorithm amplifies conspiracy and extremist content even when adjacent source material consists of civil-society awareness campaigns (Schmitt et al., 2018; Zieringer & Rieger, 2023; for TikTok, Matlach et al., 2025).

Closely related to this dynamic is another threat potential of generative AI: mainstreaming. Mainstreaming refers to the process through which radical or extremist actors render their content accessible to broader audiences while minimizing its conspicuousness, thereby evading moderation while maximizing reach (Rothut et al., 2024; Shin, 2024, Chapter 3). AI can facilitate this process by generating problematic content, concealing its intent, or embedding it in the form of dog whistles—that is, coded references recognizable primarily to in-group audiences. Applications range from the linguistic camouflage of extremist content through paraphrasing or coded language to multimodal concealment, such as embedding text within images, using insider emojis, or placing messages within audio tracks accompanied by innocuous visual material. AI can also support moderation evasion through subtle machine-generated alterations that mislead search filters or large language model (LLM) prompts that create seemingly neutral contextual framing. These practices can be further combined with automation and audience-building strategies, including social bots, deepfakes, and the dissemination of seemingly innocuous "teaser" content that redirects users to less moderated spaces. One likely consequence is the proliferation of so-called borderline content (Macdonald & Vaughan, 2024): material that does not explicitly promote hatred, only indirectly signals disinformation, or operates near the threshold of glorifying violence, thereby occupying a legal and regulatory gray area.

The shift in AI's role from curation to content generation also means that AI-amplified threats do not affect all users equally. Marginalized groups—including ethnic and religious minorities, LGBTQ+ individuals, women, journalists, and activists—are disproportionately targeted by hate speech and coordinated harassment online (Davis, 2025; Obermaier et al., 2018; Van Houtven et al., 2024). The consequences extend beyond individual harm. Exposure to such content has been shown to reduce willingness to participate in online public discourse, creating a chilling effect on precisely those voices that democratic deliberation requires (Rieger et al., 2024; Rothut et al., 2023). These same groups are often among those most likely to engage in, or benefit from, counter speech understood as deliberate communicative responses to harmful content. Yet sustained engagement in counter speech can impose substantial emotional and cognitive burdens

(Baider, 2023; Ping et al., 2025), and AI-amplified hate can quickly overwhelm individual capacities to respond. AI has also increasingly assumed the role of a discursive recipient, processing and interpreting ongoing debates. AI enables harmful content to be tailored more precisely to individual users and specific conversational contexts. Recent experimental evidence documents attitudinal effects on specific worldviews when AI engages individually with users' arguments in persuasive interactions (e.g., Costello et al., 2024; Salvi et al., 2025).

Several of these threat potentials were already apparent prior to the current wave of generative AI. Examples include microtargeting during the 2016 US presidential election and the algorithmically curated isolation of discourse in so-called echo chambers, which has frequently been identified as a driver of social polarization, particularly in binary political contests, such as the Brexit referendum. Similarly, debates surrounding so-called "fake news"—that is, content designed to imitate trustworthy news media—predated generative AI but gained new impetus through the AI-enabled Doppelgänger campaign (e.g., Pamment & Tsurtsumia, 2025). Beginning in 2022, dozens of websites appeared in Western countries around national elections that visually mimicked credible news outlets while embedding disinformation among copied news stories. These sites appeared to have been largely created and operated using AI and were subsequently linked to the Russian Social Design Agency.

Beyond the immediate manipulation of public opinion, a further and less visible threat potential lies in the systematic corruption of AI training data. Modern AI systems, particularly LLMs, are trained on vast quantities of internet-derived and ostensibly credible sources that may contain distortions such as stereotypes, normative biases, or unequal forms of representation (cf. Gallegos et al., 2024). The concentration of training pipelines around a relatively small number of overlapping corpora further compounds this risk, as diversification demands both technical effort and commercial willingness to depart from established practices. A particularly concerning manifestation of this threat is the emergence of so-called AI swarms: networks of coordinated AI-driven personas that can systematically flood the internet with fabricated content, contaminating training corpora and calcifying false narratives in model weights over successive model-training cycles (Schroeder et al., 2026). As a result, distinguishing reliable information from increasingly sophisticated disinformation has become progressively more difficult. Once incorporated into training datasets, such distortions may be reproduced and disseminated by AI systems themselves, potentially contributing to the marginalization of already vulnerable groups. The limitations of existing deliberative processes—and their tendency to respond reactively rather than proactively to technological change (Frischlich et al., 2017)—are, in part, a consequence of this complexity.

3. Response Options

Mapping available response options requires, as a first step, a systematic delineation of the scope for intervention. To identify the countermeasures that AI may support, we draw on the broader literature while avoiding technologically deterministic assumptions (see, e.g., Battista & Mangone, 2025; Jungherr, 2023; Katzenbach, 2021; Paltieli, 2023). Rather than focusing on specific technical implementations, we examine the structural logic of available responses. Broadly speaking, these fall into two categories. The first comprises direct interventions capable of exerting direct influence on the threat potentials AI poses to democratic discourse. The second consists of more indirect and longer-term measures that address the structural conditions under which those threats emerge. Both forms of response are, in our view,

not merely desirable but structurally necessary for managing the multidimensional threat potentials outlined above.

We begin with direct interventions. A substantial body of research addresses the automated detection of content meeting specified harm-related criteria—for example, AI-generated posts that may constitute criminal offenses such as incitement to violence, or AI-orchestrated campaigns that simultaneously disseminate coordinated disinformation narratives. Such approaches are inherently adversarial: Harmful actors' technological capabilities evolve alongside, and often ahead of, those of moderators, raising persistent questions about institutional capacity and resource allocation.

Equally important is the normative framework within which such interventions operate (see also Barnoy et al., 2025). At issue is the question of exactly what should be prevented. The preservation of diversity is already regarded as a guiding democratic principle in the context of platform governance (Schneiders et al., 2024). If the objective were limited to the enforcement of existing legal standards, automated detection of incitement to violence would appear self-evidently necessary. No other approach is plausibly scalable given the volume of online content. Yet, the costs and limitations of delegating offense detection to algorithmic systems should not be underestimated. Biases and distortions in training data may lead to inconsistent enforcement, while AI-assisted actors can circumvent detection through many of the strategies discussed above. For this reason, the EU has increasingly placed responsibility on platform operators themselves, requiring not only detection but also transparency and reporting under the Digital Services Act. Nevertheless, the effectiveness of current implementation remains open to question, as reporting requirements focus primarily on the number and speed of detected cases rather than on broader societal outcomes.

A separate challenge concerns AI-orchestrated opinion-making. This is best understood as a freedom-of-expression issue rather than a technical one: No democratic legal framework can legitimately prohibit opinions as such (unlike incitement), and legislation attempting to do so would in significant part constitute censorship, carrying severe risks under conditions of institutional bad faith. The legal middle ground is therefore narrow (e.g., Wachter et al., 2021). At the same time, numerous studies have demonstrated that even legally permissible forms of expression, including hate speech, can reduce openness to participation in discourse, thereby reducing plurality within public debate (Rieger et al., 2024; Rothut et al., 2023; Schulze et al., 2024).

The normative challenges are further complicated by variation in public attitudes toward moderation. Acceptance of algorithmic moderation is far from universal, and opposition to censorship significantly shapes users' willingness to support even well-intentioned interventions (Wiesner, 2026). In many cases, the empirical evidence already points toward viable policy responses. The primary obstacle is therefore not epistemic but political: a lack of normative consensus and, ultimately, of political will. More targeted interventions presuppose greater normative consensus and the precondition for such consensus is, in turn, more sustained public deliberation.

A recurring concern in this context is the limited capacity of AI detection systems to identify forms of harmful content beyond explicit or "classical" hate speech (Macdonald & Vaughan, 2024). Implicit hate speech (Rieger et al., 2021), fear speech (Greipl et al., 2024), and other forms of borderline content remain considerably more difficult to detect. The multimodal character of online hate, together with its subtler

manifestations (e.g., negative stereotyping), remains comparatively resistant to automated identification. At the same time, there is reason to suspect that borderline content carries significant normative consequences because its gradual accumulation may contribute to the normalization of harmful discourse (Rothut et al., 2024; Saha et al., 2023). Against this backdrop, the primary deployment context for AI in this domain has been content moderation. This can be implemented through chat assistants designed to maintain political diversity in discourse (Argyle et al., 2023), to argue persuasively against certain views (e.g., conspiracy beliefs; Costello et al., 2024), or to function as a mediator in groups holding divergent perspectives (Tessler et al., 2024).

Indirect response options, by contrast, operate over a longer temporal horizon, seeking to reshape the structural conditions within which threat potentials arise rather than addressing their immediate manifestations. Legal regulation constitutes one such structural mechanism, one which has assumed increasingly specialized forms in Germany and the EU. Key instruments include the General Data Protection Regulation (governing data use in content moderation); the Network Enforcement Act (mandating rapid removal of clearly illegal content); the Digital Services Act (requiring risk assessments and transparency measures from large platforms); and the AI Act (establishing risk-based obligations for AI systems deployed in public communication contexts). Crucially, the effectiveness of such legislative frameworks depends not merely on their formal enactment but also on the co-development of enforcement infrastructures. In the cases cited above, this condition appears to be at least partially satisfied.

4. Counter Speech

This section draws on and extends a growing body of interdisciplinary research on AI and counter speech, most comprehensively synthesized in a recent report produced by an international group of scholars and practitioners (Obermaier et al., 2026).

4.1. Conceptualizing and Assessing AI-Assisted Counter Speech

Counter speech refers to deliberate communicative efforts by individuals, organizations, or AI systems to directly challenge, rebut, or contextualize harmful online content. The concept, however, is far from uniform across the literature. It has variously been conceptualized as a reaction to hate speech aimed at reducing harm and supporting victims (Bahador, 2021; Schieb & Preuss, 2018; Ziegele et al., 2020); as an argumentative, fact-based corrective strategy (Richards & Calvert, 2000); as a bottom-up, decentralized practice driven by grassroots user engagement rather than as top-down moderation (Jia & Schumann, 2025); as a visible expression of public opposition and social pushback (Friess et al., 2021; Porten-Cheé et al., 2020); and as a multi-strategic, platform-based practice encompassing direct replies, dislikes, and flagging mechanisms (Kalch & Naab, 2017). We adopt a broad definition that encompasses all of these dimensions while focusing primarily on counter speech as a direct, content-level response, as this is where AI assistance has been most extensively studied.

For analytical purposes, it is useful to distinguish among three qualitatively distinct modalities of AI-supported counter speech, each of which raises distinct questions regarding effectiveness, authenticity, and governance: (a) AI systems that assist human counter speakers (e.g., by providing fact-based arguments, drafting responses, or offering emotional scaffolding); (b) AI systems that autonomously generate counter speech at scale; and

(c) AI systems that facilitate structured dialogue between users holding opposing viewpoints (Argyle et al., 2023; Tessler et al., 2024). The normative and empirical stakes differ substantially across these modalities, and conflating them risks overstating AI's potential while underestimating the challenges associated with each.

Empirically and normatively, the most promising modality appears to be AI-assisted human counter speech. Collaborative human–AI systems such as CounterQuill, which structures counter speech into learning, brainstorming, and co-writing stages, increase counter speakers' confidence, sense of ownership, and willingness to post, compared with general-purpose chatbots (Ding et al., 2025). These findings suggest that AI deployed in educational and scaffolding roles may be more normatively acceptable and practically effective than fully automated replies.

Counter speakers themselves express nuanced concerns regarding AI involvement. Drawing on in-depths Interviews with experienced counterspeakers and a large-scale survey, Mun et al. (2024) found that users value AI as a source of information, training, and emotional support, while simultaneously expressing concerns about authenticity (the perceived quality of counter speech as genuine, personal, and non-automated), agency (the counter speaker's sense of ownership over and responsibility for their communicative act), and misrepresentation when AI communicates directly on their behalf. Participants also value tools that lower barriers to engagement, including time demands, emotional strain, and limited argumentative skill—without diminishing the personal significance of counter speech (Mun et al., 2024). Moreover, willingness to adopt AI-assisted tools varies according to their design characteristics, particularly where AI risks depleting rather than conserving the scarce emotional and cognitive resources required for sustained counter speech engagement (Frischlich et al., 2024). However, AI support need not end with the drafting of a counter speech message: It can also assist counter speakers in handling the aftermath of their intervention, for instance by helping process hostile replies or by providing emotional support following exposure to backlash (Obermaier et al., 2026). Taken together, these findings point toward what we term the augmentation model, in which AI functions as partner and coach rather than autonomous actor, preserving the authenticity and agency that make counter speech normatively valuable within a socio-technical system.

Fully automated AI-informed counter speech presents a more ambivalent picture. A large field experiment (pre-print) on Twitter/X (N = 2,664) found that generic, human-authored "warning-of-consequences" messages reduced subsequent hate speech, whereas contextualized LLM-generated counter speech was ineffective and occasionally counterproductive, increasing hateful behavior in the short term (Bär et al., 2024). These findings suggest that greater personalization does not automatically translate into greater effectiveness and that carefully crafted human-authored templates might outperform LLM-generated responses in real-world settings.

Beyond effectiveness, a further underexplored risk concerns the biases embedded in generative AI systems used to produce counter speech. LLMs trained on skewed or historically biased corpora may reproduce stereotypes or systematically fail to recognize or adequately counter hate speech targeting marginalized communities (Gallegos et al., 2024). This raises the normative concern that AI-generated counter speech, rather than being a neutral corrective, may replicate the asymmetries it is designed to address. These biases are especially pronounced for the detection of subtle or implicit forms of hate speech, including contexts requiring cultural sensitivity, further compounding the risks for marginalized groups (Obermaier et al., 2026).

Complementary evidence presents a somewhat more optimistic picture. In a series of experiments, Wu et al. (2025; $N = 809$) found that AI-generated counter speech can increase users' willingness to engage when counter speech is present, with empathy-based messages eliciting stronger engagement than fact-based alternatives. At the same time, users were more willing to engage when the AI origin of the message was not disclosed, and perceived trust mediated the relationship between generator type, identity disclosure, and engagement intentions (Wu et al., 2025). Limitations have been identified in the context of conspiracy theories, where GPT-4o, Llama 3, and Mistral, evaluated as debunking agents, frequently produce generic, repetitive, or hallucinated responses, rendering naive prompt-based deployment problematic in practice (Lisker et al., 2025). Although fine-tuning on domain-specific datasets can improve contextual relevance and adequacy, automatic evaluation metrics correlate only weakly with human judgments, underscoring the continuing need for human-in-the-loop evaluation (Cima et al., 2024). Overall, these findings suggest that the promise of scalable automated AI counter speech currently exceeds the available evidence for its effectiveness.

A related but conceptually distinct strategy is prebunking, which seeks to build resilience through the preemptive dissemination of warnings about emerging disinformation (Lewandowsky & van der Linden, 2021). By exposing audiences to attenuated forms of conspiratorial or disinformative content, or to the structural patterns through which such content operates, this preventive measure aims to inoculate audiences who have not yet encountered the full-strength version (Roozenbeek et al., 2021). The analogous function of a "vaccine" (inoculation; see Compton, 2025; Compton & Braddock, 2025) is fulfilled here by AI. One early example is Google's Jigsaw Redirect campaign, which provides counter-messaging before users access problematic YouTube content. AI enables such interventions to be personalized at a scale previously unattainable through traditional campaigns (Biddlestone et al., 2025). However, the quality of AI-generated prebunking content remains uneven, and the tendency of LLMs tasked with debunking conspiracy theories to frequently produce generic or hallucinated content once again highlights the importance of human editorial oversight (Lisker et al., 2025).

While prebunking operates pre-emptively, thus strengthening resistance before harmful content is encountered, counter speech responds reactively to such content already in circulation. These strategies are complementary, and both benefit from AI support. Yet, both also face the same structural challenge: AI's capacity to scale communicative interventions does not automatically translate into effectiveness or normative legitimacy. AI thus occupies a dual role in this domain, serving both as a driver of the harmful content that necessitates counter speech and prebunking and as a tool for addressing it. However, the consequences depend less on the technology itself than on the conditions under which it is deployed.

4.2. Structural and Governance Conditions for Legitimate Counter Speech

Counter speech and prebunking address harmful content at the level of individual communicative acts. Sustaining these efforts at scale, however, requires structural conditions—both technical and normative—that neither automated systems nor individual counter speakers can establish on their own. We identify three complementary structural mechanisms: algorithmic interventions, platform design, and governance frameworks (both legal and deliberative). More fundamentally, the threat potentials associated with AI make it, in our assessment, necessary for democratic societies to establish legitimate procedures through which the norms governing digital communication can be debated, revised, and justified.

Such deliberative processes might be anchored in competing normative visions of what a well-functioning democratic public sphere ought to achieve. Deliberative models, for example, emphasize consensus-oriented and reasoned discourse in which journalism serves a mediating function, whereas liberal models view the public sphere primarily as a forum for competing perspectives to which journalism provides orientation (Donges & Jarren, 2022; Ferree et al., 2002). Building on these normative foundations, prevention strategies can be developed to foster more inclusive, egalitarian, or diverse forms of discourse.

One such strategy involves modifying platform design through what has been termed *friction* (Naderer et al., in press). Friction elements are purposively introduced design features intended to decelerate users' online behavior, thereby fostering conditions for reflection. Examples include multi-step cancellation of subscriptions, repeated "really delete?" confirmations, and daily caps on certain actions (e.g., swipes on Tinder). They increase users' control over the content they aim to encounter, reduce potentially impulsive behavior, offer learning opportunities, and raise awareness of harmful forms of content (Ingegno, 2023). Friction elements are particularly relevant in the context of counter speech because they can slow the production and amplification of harmful content, thereby creating the temporal space within which counter speech interventions can be effective.

A related strategy involves algorithmic interventions by platform operators aimed at reducing the visibility and reach of problematic content, often referred to as downranking or algorithmic demotion (Macdonald & Vaughan, 2024). Unlike outright removal, which may raise concerns regarding freedom of expression (Kozyreva et al., 2023), downranking reduces the algorithmic amplification of particular posts (e.g., with combinations of keywords associated with polarized reactions) or accounts, thereby limiting their dissemination without deleting them. Shadow banning could also serve as an AI-controlled tool for sanctioning content that conflicts with deliberative norms, covertly reducing a user's visibility without notification, whereas downranking lowers a post's distribution priority while leaving it technically accessible (Gillespie, 2022).

More fundamentally, however, the contemporary internet lacks a framework of legitimate governance processes through which its epistemic norms can be collectively negotiated (see also Haim & Neuberger, 2022). Public-serving infrastructure requires democratically legitimated participation as a precondition for the accumulation of social trust and acceptance. Such legitimacy is commonly derived from appointment and consent (*input legitimation*, e.g., the elected European Parliament), from the effectiveness of institutional output (*output legitimation*, e.g., the European Central Bank), or through the framework of an already legitimated process (*throughput legitimation*, e.g., referendums, European Ombudsman, European Citizens' Initiative). Although the internet as a whole fulfills none of these criteria in a comprehensive manner (Deuze & McQuail, 2020), several instructive examples exist.

Wikipedia, for instance, has developed a trusted epistemic process grounded in openness, participation, and transparency. When disputes arise (see Frost-Arnold, 2019), institutionalized meta-discursive mechanisms—including discussion pages and editorial forums—are used to negotiate procedures rather than to adjudicate substantive content directly. The resulting procedural legitimacy has contributed substantially to Wikipedia's credibility as a knowledge resource. Other online examples that involve legitimated processes include decentralized platform infrastructures such as the Fediverse (e.g., Mastodon), open-source communities such as Mozilla, and participatory democracy platforms such as online petition systems.

Building on these institutional logics, recent proposals advocate the creation of public-interest digital infrastructure. One example is a European digital media platform that integrates dissemination, discussion, and archiving as interconnected journalistic functions and is explicitly designed to counter the illiberal tendencies associated with commercially owned platforms (Neuberger, 2026). What unites these examples is not their technical architecture but their institutional logic: legitimacy derived from process rather than outcome, transparency in decision-making, and the availability of structured meta-discursive spaces for norm negotiation. Applied to AI-assisted counter speech, this implies that platforms and civil society actors should establish explicit and publicly accountable frameworks governing when, how, and by whom AI-assisted counter speech may be deployed. The legitimacy of such frameworks should rest not on technical efficiency alone but on inclusive and transparent deliberation.

We acknowledge that substantive consensus regarding desirable discourse norms may be unattainable in pluralistic democracies. A more pragmatic alternative is a graduated approach (see Helberger, 2019). At a minimum, AI-assisted counter speech should be governed by liberal principles of transparency and non-discrimination, including disclosure of AI authorship and equal treatment across social groups. More demanding deliberative requirements, such as the participatory legitimation of platform governance norms, can be pursued where institutional capacity permits. Such an incremental approach does not require prior consensus on all normative questions; rather, it requires only that each level of governance be democratically accountable and institutionally enforceable. This position aligns with recent policy-oriented scholarship advocating stakeholder-specific governance arrangements rather than the one-size-fits-all regulation of AI-assisted counter speech (Obermaier et al., 2026).

Taken together, these considerations highlight the need for democratically legitimate content-governance processes. Such legitimacy depends in part on conceptual clarity regarding what constitutes harmful or borderline content and under what conditions particular actors are authorized to restrict its visibility or distribution. These questions become especially challenging in the context of AI-generated synthetic content, which blurs conventional distinctions between authorship, authenticity, and detectability in ways that existing moderation frameworks were not designed to address (Fisher et al., 2024). The finding that users engage more readily with AI-generated counter speech when its AI origin is not disclosed (Wu et al., 2025) further underscores the importance of transparency and accountability. Macdonald and Vaughan (2024), accordingly, advocate greater transparency in content moderation practices.

These concerns apply equally to AI-assisted counter speech. Who determines what constitutes a legitimate counter speech intervention? What forms of oversight are required when AI generates counter speech at scale but influences public discourse in ways that are no more transparent than the harmful content it seeks to counter? The EU AI Act's risk-based classification framework provides one possible starting point, as systems intended to shape public communication may warrant heightened scrutiny. More broadly, the human–AI collaboration model—in which AI supports rather than replaces human counter speakers—offers a promising means of preserving authenticity and human agency while still enabling scale (Mun et al., 2024; Obermaier et al., 2023). Yet, even such hybrid arrangements require governance. Norms governing disclosure, accountability, and the permissible scope of AI involvement in democratic communication remain underdeveloped and will require continued public deliberation.

5. Conclusion

The contemporary internet has become a fragmented and increasingly discordant arena of opinion exchange, characterized by declining discourse quality and shaped by AI-driven transformations at multiple structural levels. Deployed by actors with widely differing intentions, AI functions simultaneously as author, amplifier, and infrastructure for hate speech, disinformation, conspiracy narratives, and forms of harmful content that increasingly resist clear categorization. Counter speech has emerged as one of the most normatively attractive responses to these developments. Rather than suppressing harmful speech, it seeks to address it through further communication, thereby preserving the deliberative values upon which democratic public spheres depend.

The evidence reviewed in this article, however, cautions against overly optimistic assumptions regarding AI's role in scaling counter speech. Fully automated forms of AI-generated counter speech frequently underperform and may even prove counterproductive. Human–AI collaboration models appear considerably more promising, but they require careful design, oversight, and institutional support, as is typical of socio-technical systems more broadly (Tsvetkova et al., 2024). Although a wide range of countermeasures has attracted substantial scholarly attention, their practical implementation has become increasingly challenging. This is attributable, in part, to the retreat of powerful platform operators (under conditions of shifting political incentives) from active cooperation with content-governance efforts. Policymakers, meanwhile, face not only the technological transformations discussed throughout this article but also a more fundamental challenge: the absence of normative consensus concerning what a democratically desirable digital public sphere should actually entail.

The organization of democratic public communication has always involved balancing competing normative principles. Questions concerning the purpose and structure of the public sphere cannot be resolved through legal regulation alone, nor can they be settled solely through empirical research. Equally, they should not be left to the discretionary preferences of powerful private actors. At present, the dominant framing of digital public spheres reflects commercial interests, an arrangement broadly consistent with the institutional logic of certain media systems.

For the EU, however, this commercial framing is only partially applicable, given its stronger traditions of public-interest media governance. Taken together, the four steps of our analysis point toward a clear conclusion: Addressing AI's threats to democratic discourse requires moving beyond technological fixes toward democratically legitimated governance. First, there is a considerable need for democratically legitimated arenas within which the normative framework conditions for digital public communication can be openly debated and negotiated. To create such arenas, modern democratic participation mechanisms—such as digital citizens' assemblies (e.g., Dienel et al., 2024)—appear particularly well suited. Second, European intermediaries of public significance should be strengthened in order to facilitate sustained exchange among stakeholders from law, academia, platform operators/technology companies, politics, and civil society. Third, the creation of a publicly funded social media network, an idea that has already been proposed by Zuckerman (2020), warrants serious consideration. Finally, greater interdisciplinary collaboration is needed to foster regular exchange among computer scientists, communication scholars, psychologists, legal scholars, and other social scientists. This interdisciplinary agenda should explicitly include counter speech research, as scaling AI-assisted counter speech in ways that preserve human agency,

authenticity, and democratic accountability represents one of the most pressing challenges at the intersection of AI, communication, and democracy.

AI functions simultaneously as a source of democratic risk and as a potential means of addressing that risk. Which role ultimately predominates will depend less on the technology itself than on the governance arrangements, institutional structures, normative frameworks, and democratic deliberation surrounding its deployment. Europe is structurally well positioned to contribute to the development of such frameworks, provided it can translate its regulatory ambitions into durable and publicly legitimate institutions. Ultimately, the future of digital public spheres is not primarily a technological question; it is a political and normative one.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between LMU Munich and Cogitatio Press.

Conflict of Interests

The authors declare no conflict of interests.

LLMs Disclosure

Initially, a shorter text (word count approximately 4,000 words) on the role of AI for democratic discourse online was translated from German to English using Claude Sonnet 4.6. This translation was then restructured, partly rewritten, and extended with a specific focus on counter speech to meet the requirements of this thematic issue.

References

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, *120*(41), Article e2311627120. https://doi.org/10.1073/pnas.2311627120

Bahador, B. (2021). Countering hate speech. In H. Tumber & S. Waisbord (Eds.), *The Routledge companion to media disinformation and populism* (pp. 507-518). Routledge. https://doi.org/10.4324/9781003004431-53

Baider, F. (2023). Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, *11*(2), 249–260. https://doi.org/10.17645/pag.v11i2.6465

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). *Generative AI may backfire for counterspeech*. arXiv. https://doi.org/10.48550/arXiv.2411.14986

Barnoy, A., Bakbani-Elkayam, S., Miller, B., & Keren, A. (2025). The role of epistemic norms in mitigating the spread of misinformation. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251385917

Battista, D., & Mangone, E. (2025). Technological culture and politics: Artificial intelligence as the new frontier of political communication. *Societies*, *15*(4), Article 75. https://doi.org/10.3390/soc15040075

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). *Considerations for successful counterspeech*. Dangerous Speech Project. https://dangerousspeech.org/considerations-for-successful-counterspeech

Biddlestone, M., Roozenbeek, J., Suiter, J., Culloty, E., & van der Linden, S. (2025). Tune in to the prebunking network! Development and validation of six inoculation videos that prebunk manipulation tactics and logical fallacies in misinformation. *Political Psychology*, *46*(6), 1858–1886. https://doi.org/10.1111/pops.70015

Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2024). Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *WWW '25: Proceedings of the ACM on Web Conference 2025* (pp. 5022–5033). Association for Computing Machinery. https://doi.org/10.1145/3696410.3714507

Compton, J. (2025). Inoculation theory. *Review of Communication*, *25*(1), 1–13. https://doi.org/10.1080/15358593.2024.2370373

Compton, J., & Braddock, K. (2025). Inoculation theory and conspiracy, radicalization, and violent extremism. In S. A. Samoilenko & S. Simmons (Eds.), *The handbook of social and political conflict* (pp. 405–413). Wiley. https://doi.org/10.1002/9781119895534.ch36

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Davis, J. (2025). Disinformation in the era of generative AI: Challenges, detection strategies, and countermeasures. In R. Luttrell & A. A. Wallace (Eds.), *Public relations and the rise of AI* (pp. 242–269). Routledge.

Deuze, M., & McQuail, D. (2020). *Media & mass communication theory* (7th ed.). Sage.

Dienel, H.-L., von Blanckenburg, C., & Bach, N. (2024). Mini publics online. Erfahrungen mit Online-Bürgerräten und Online-Planungszellen. In N. Kersting, J. Radtke, & S. Baringhorst (Eds.), *Handbuch Digitalisierung und politische Beteiligung*. Springer VS. https://doi.org/10.1007/978-3-658-31480-4_37-1

Ding, X., Ping, K., Gunturi, U., Çarik, B., Stil, S., Wilhelm, L., Daryanto, T., Hawdon, J., Lee, S., & Rho, E. (2025). Designing human-AI collaboration to support learning in counterspeech writing. In *2025 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 400–412). https://doi.org/10.1109/vl-hcc65237.2025.00052

Donges, P., & Jarren, O. (2022). *Politische Kommunikation in der Mediengesellschaft* (5th ed.). Springer VS.

Farrand, B. (2025). Online platforms, intermediary responsibility, and human rights: Digital copyright as a site of multiple contestations in the EU. In B. Wagner, M. C. Kettemann, K. Vieth-Ditlmann, & S. Montgomery (Eds.), *Research handbook on human rights and digital technology* (pp. 54–68). Edward Elgar Publishing. https://doi.org/10.4337/9781035308514.00010

Ferree, M. M., Gamson, W. A., Gerhards, J., & Rucht, D. (2002). Four models of the public sphere in modern democracies. *Theory and Society*, *31*(3), 289–324. https://doi.org/10.1023/A:1016284431021

Fisher, S. A., Howard, J. W., & Kira, B. (2024). Moderating synthetic content: The challenge of generative AI. *Philosophy & Technology*, *37*, Article 133. https://doi.org/10.1007/s13347-024-00818-9

Friemel, T. N., & Neuberger, C. (2023). The public sphere as a dynamic network. *Communication Theory*, *33*(2/3), 92–101. https://doi.org/10.1093/ct/qtad003

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, *38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Frischlich, L., Klapproth, J., Frank, S., Heckmann, M., Kunze, S. E., & Murgas, T. (2024). Fighting fakes on WhatsApp—Audience perspectives on fact bots as countermeasures. *Digital Journalism*, *12*(5), 700–720. https://doi.org/10.1080/21670811.2024.2341299

Frischlich, L., Rieger, D., Morten, A., & Bente, G. (2017). *Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand*. Bundeskriminalamt.

Frost-Arnold, K. (2019). Wikipedia. In D. Coady & J. Chase (Eds.), *The Routledge handbook of applied epistemology* (pp. 28–40). Routledge.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

García-Orosa, B. (2022). Digital political communication: Hybrid intelligence, algorithms, automation and disinformation in the fourth wave. In B. García-Orosa (Ed.), *Digital political communication strategies* (pp. 3–23). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-81568-4_1

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221117552

Greipl, S., Hohner, J., Schulze, H., Schwabl, P., & Rieger, D. (2024). "You are doomed!" Crisis-specific and dynamic use of fear speech in protest and extremist radical social movements. *Journal of Quantitative Description: Digital Media*, *4*. https://doi.org/10.51685/jqd.2024.icwsm.8

Haim, M., & Neuberger, C. (2022). The paradox of knowing more and less: Audience metrics and the erosion of epistemic standards on the internet. *Studies in Communication and Media*, *11*(4), 566–589. https://doi.org/10.5771/2192-4007-2022-4-566

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, *7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Hickey, D., Fessler, D. M., Lerman, K., & Burghardt, K. (2025). X under Musk's leadership: Substantial hate and no reduction in inauthentic activity. *PLoS ONE*, *20*(2), Article e0313293. https://doi.org/10.1371/journal.pone.0313293

Hiller, A., & de las Casas, P. M. (2025). *Generative AI and the German far right: Narratives, tactics and digital strategies*. Institute for Strategic Dialogue. https://www.isdglobal.org/wp-content/uploads/2025/02/The-use-of-generative-AI-by-the-German-Far-Right.pdf

Hindman, M. (2008). *The myth of digital democracy*. Princeton University Press.

Ingegno, M. (2023, September 29). Friction in design: The good, the bad, and the.. dark. *Design With the Mind in Mind*. https://www.linkedin.com/pulse/friction-design-good-bad-dark-make-it-toolkit

Jia, Y., & Schumann, S. (2025). Tackling hate speech online: The effect of counter-speech on subsequent bystander behavioral intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *19*(1), Article 4. https://doi.org/10.5817/CP2025-1-4

Jungherr, A. (2023). Artificial intelligence and democracy: A conceptual framework. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186353

Jungherr, A., & Schroeder, R. (2023). Artificial intelligence and the public arena. *Communication Theory*, *33*(2/3), 164–173. https://doi.org/10.1093/ct/qtad006

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication and Media*, *6*(4), 395–419. https://doi.org/10.5771/2192-4007-2017-4-395

Katzenbach, C. (2021). "AI will fix this"—The technical, discursive, and political turn to AI in governing communication. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211046182

Klausa, T. (2023). Graduating from 'new-school'—Germany's procedural approach to regulating online discourse. *Information, Communication & Society*, *26*(1), 54–69. https://doi.org/10.1080/1369118X.2021.2020321

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation.

Proceedings of the National Academy of Sciences, *120*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lisker, M., Gottschalk, C., & Mihaljević, H. (2025). Debunking with dialogue? Exploring AI-generated counterspeech to challenge conspiracy theories. In A. Calabrese, C. de Kock, D. Nozza, F. M. Plaza-de-Arco, Z. Talat, & F. Vargas (Eds.), *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)* (pp. 163–178). Association for Computational Linguistics.

Macdonald, S., & Vaughan, K. (2024). Moderating borderline content while respecting fundamental values. *Policy & Internet*, *16*(2), 347–361. https://doi.org/10.1002/poi3.376

Mantello, P., Ho, T. M., & Podoletz, L. (2023). Automating extremism: Mapping the affective roles of artificial agents in online radicalization. In E. Pashentsev (Ed.), *The Palgrave handbook of malicious use of AI and psychological security* (pp. 81–103). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-22552-9_4

Matlach, P.-C. Y., Castillo, A., Drath, C., & Hevesi, E. F. (2025). *Recommending hate: How TikTok's search engine algorithms reproduce societal bias*. Institute for Strategic Dialogue. https://www.isdglobal.org/publication/recommending-hate-how-tiktoks-search-engine-algorithms-reproduce-societal-bias

Mun, J., Buerger, C., Liang, J. T., Garland, J., & Sap, M. (2024). Counterspeakers' perspectives: Unveiling barriers and AI needs in the fight against online hate. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 742). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642025

Naderer, B., Rothut, S., & Rieger, D. (in press). Preventing and countering online radicalization in the digital age: Actors, strategies and challenges. In I. Awan & C. Pelham (Eds.), *International handbook on counter-radicalization*. De Gruyter.

Neuberger, C. (2026). Building a digital media platform for the European public sphere: Plans, projects, pitfalls. *M/C Journal*, *29*(2). https://doi.org/10.5204/mcj.3256

Obermaier, M., Buerger, C., Frischlich, L., Bund, L., Carathanassis, F., Clausen, A., Doseva, S., Eissfeldt, J., Garland, J., Grimme, C., Ghazi-Zahedi, K., Haim, M., Herderich, A., Li, Y., Oetting, H., Puschmann, C., Rho, E., Stoll, A., Wenninger, A., & Ziegele, M. (2026). *The role of AI in counterspeech: Assessing risks and potential* (bidt Analyses and Studies No. 21). Bavarian Research Institute for Digital Transformation. https://doi.org/10.35067/xypq-kn79

Obermaier, M., Hofbauer, M., & Reinemann, C. (2018). Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it. *Studies in Communication and Media*, *7*(4), 499–524. https://doi.org/10.5771/2192-4007-2018-4-499

Obermaier, M., Schmid, U. K., & Rieger, D. (2025). Empowerment is key? How perceived political and critical digital media literacy explain direct and indirect bystander intervention in online hate speech. *Social Media + Society*, *11*(1). https://doi.org/10.1177/20563051251325598

Obermaier, M., Schmuck, D., & Saleem, M. (2023). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, *25*(9), 2339–2358. https://doi.org/10.1177/14614448211017527

Paltieli, G. (2023). Re-imagining democracy: AI's challenge to political theory. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 333–342). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00036

Pamment, J., & Tsurtsumia, D. (2025). *Beyond Operation Doppelgänger: A capability assessment of the Social*

Design Agency (MPF Report Series No. 8/2025). Swedish Psychological Defense Agency. https://mpf.se/psychological-defence-agency/publications/archive/2025-05-15-beyond-operation-doppelganger-a-capability-assessment-of-the-social-design-agency

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Ping, K., Hawdon, J., & Rho, E. H. (2025). Perceiving and countering hate: The role of identity in online responses. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), Article CSCW147. https://dl.acm.org/doi/10.1145/3711045

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication*, *14*, 514–534.

Richards, R. D., & Calvert, C. (2000). Counterspeech 2000: A new look at the old remedy for "bad" speech. *BYU Law Review*, *2000*(2), 553–586.

Rieger, D., Greipl, S., Schmid, U. K., Hohner, J., & Schulze, H. (2024). Hassrede als Merkmal von (Online-)Radikalisierung. In T. Rothmund & E. Walther (Eds.), *Psychologie der Rechtsradikalisierung* (pp. 125–134). Kohlhammer.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media + Society*, *7*(4). https://doi.org/10.1177/20563051211052906

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisted. *Educational and Psychological Measurement*, *81*(2), 340–362. https://doi.org/10.1177/0013164420940378

Rothut, S., Sacher, A.-L., Strohmeier, R., & Reinemann, C. (2023). Meinungsfreiheit in Gefahr? Wie politische Einstellungen und individuelle Erfahrungen die Wahrnehmung der Meinungsfreiheit in Deutschland prägen. *Studies in Communication and Media*, *12*(1), 48–91. https://doi.org/10.5771/2192-4007-2023-1-48

Rothut, S., Schulze, H., Rieger, D., & Naderer, B. (2024). Mainstreaming as a meta-process: A systematic review and conceptual model of factors contributing to the mainstreaming of radical and extremist positions. *Communication Theory*, *34*(2), 49–59. https://doi.org/10.1093/ct/qtae001

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Saha, P., Garimella, K., Kalyan, N. K., Pandey, S. K., Meher, P. M., Mathew, B., & Mukherjee, A. (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, *120*(11), Article e2212270120. https://doi.org/10.1073/pnas.2212270120

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Schieb, C., & Preuss, M. (2018). Considering the elaboration likelihood model for simulating hate and counter speech on Facebook. *Studies in Communication and Media*, *7*(4), 580–606.

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube: Recommendation algorithms. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Schneiders, P., Stegmann, D., Stark, B., Zieringer, L., & Reinemann, C. (2024). Meinungsmacht unter der Lupe: Ein Ansatz für eine vielfaltssichernde, holistische Plattformregulierung. In M. Prinzing, J. Seethaler, M. Eisenegger, & P. Ettinger (Eds.), *Regulierung, Governance und Medienethik in der digitalen Gesellschaft* (pp. 97–120). Springer.

Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Ressa, M.,

Schweitzer, F., Song, D., Summerfield, C., Tang, A., . . . Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. *Science*, *391*(6783), 354–357. https://doi.org/10.1126/science.adz1697

Schulze, H., Greipl, S., Hohner, J., & Rieger, D. (2024). Social media and radicalization: An affordance approach for cross-platform comparison. *Medien & Kommunikationswissenschaft*, *72*(2), 187–212. https://doi.org/10.5771/1615-634X-2024-2-187

Shin, D. (2024). *Artificial misinformation – Exploring human-algorithm interaction online*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-52569-8_3

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), Article eadq2852. https://doi.org/10.1126/science.adq2852

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876. https://doi.org/10.1038/s41562-024-02001-8

Van Houtven, E., Acquah, S. B., Obermaier, M., Saleem, M., & Schmuck, D. (2024). 'You got my back?' Severity and counter-speech in online hate speech toward minority groups. *Media Psychology*, *27*(6), 923–954. https://doi.org/10.1080/15213269.2023.2298684

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, Article 105567. https://doi.org/10.1016/j.clsr.2021.105567

Wardle, C. (2018). The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, *6*(8), 951–963. https://doi.org/10.1080/21670811.2018.1502047

Wiesner, A. (2026). Tackling (misleading) incivility online: A user-centric evaluation of different comment moderation strategies. *Information, Communication & Society*, *29*(5), 1495–1513. https://doi.org/10.1080/1369118X.2025.2557548

Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social media? *Telematics and Informatics*, *101*, Article 102304. https://doi.org/10.1016/j.tele.2025.102304

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. https://doi.org/10.1177/1461444819870130

Zieringer, L., & Rieger, D. (2023). Algorithmic recommendations' role for the interrelatedness of counter-messages and polluted content on YouTube—A network analysis. *Computational Communication Research*, *5*(1), 109–140. https://doi.org/10.5117/CCR2023.1.005.ZIER

Zuckerman, E. (2020, January 17). The case for digital public infrastructure. *Knight First Amendment Institute at Columbia University*. https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure

About the Authors

Diana Rieger (PhD) is full professor of communication science with a focus on media psychology at the Department of Media and Communication at LMU Munich, Germany. Her research interests revolve around the characteristics and effects of online radicalization, hate speech, extremist communication, and conspiracy content, as well as potential countermeasures such as counter speech. More information: https://www.diana-rieger.net. Photo by Katharina Hajek.

Mario Haim (PhD) is full professor of communication science with a focus on computational communication research at the Department of Media and Communication at LMU Munich, Germany. His research interests revolve around algorithmic influences, such as in political communication, journalism, health communication, or media use, as well as on (computational) methods and meta-science. More information: https://haim.it. Photo by Lena Fleischer.