bidt Bavarian Research Institute for Digital Transformation
AN INSTITUTE OF THE BAVARIAN ACADEMY OF SCIENCES AND HUMANITIES

ANALYSES AND STUDIES

The Role of AI in Counterspeech: Assessing Risks and Potential

Authors:
Magdalena Obermaier, Cathy Buerger, Lena Frischlich,
Lea Bund, Fay Carathanassis, Anne Clausen, Steliyana Doseva,
Jan Eissfeldt, Joshua Garland, Christian Grimme, Keyan Ghazi-Zahedi,
Mario Haim, Alina Herderich, Yuru Li, Hannah Oetting, Cornelius Puschmann,
Eugenia Rho, Anke Stoll, Andreas Wenninger, Marc Ziegele

Publisher: bidt – Bavarian Research Institute for Digital Transformation
en.bidt.digital

Imprint

bidt Analyses and Studies No. 21

The analyses and studies published by the bidt represent the views of the authors; they do not reflect the position of the institute as a whole.

bidt – Bavarian Research Institute for Digital Transformation
Gabelsbergerstrasse 4
80333 Munich
en.bidt.digital

Coordination
Margret Hornsteiner
Anna-Maria Esch
dialog@bidt.digital

Design
made in – Design und Strategieberatung

Publication
May 2026
ISSN: 2701-2379
DOI: 10.35067/xypq-kn79

Hate speech is a widespread phenomenon in digital environments, with severe consequences for individuals and society. Counterspeech by users witnessing the incidents, the so-called bystanders, can help mitigate these effects and foster a more benign digital public sphere without limiting the freedom of expression. However, most bystanders do not intervene due to various barriers. Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), offer new possibilities to support counterspeech efforts. At the same time, implementing AI for counterspeech poses immediate and long-term risks that require careful consideration.

This report provides a consolidated overview by authors from different academic disciplines as well as practitioners, discussing the risks and potential of AI for counterspeech. It was developed by this international and interdisciplinary group of experts at a workshop funded by the Bavarian Research Institute for Digital Transformation (bidt) and organised at LMU Munich in February 2025.

As an institute of the Bavarian Academy of Sciences and Humanities (BAdW), the Bavarian Research Institute for Digital Transformation (bidt) contributes to a better understanding of the developments and challenges of the digital transformation. It thus provides the basis for shaping the digital future of society in a responsible and public interest-oriented manner.

The authors

Magdalena Obermaier: Postdoctoral research associate at the Department of Media and Communication at LMU Munich, Munich, Germany. Email: magdalena.obermaier@ifkw.lmu.de

Cathy Buerger: Director of research at the Dangerous Speech Project, Washington DC, USA. Email: catherine.buerger@gmail.com

Lena Frischlich: Associate professor and vice-director of the Digital Democracy Centre, University of Southern Denmark, Odense, Denmark. Email: lefr@sam.sdu.dk

Lea Bund: Project manager at ichbinhier e.V., Berlin, Germany.

Fay Carathanassis: Research associate at the Technical University of Munich and associated researcher at the Bavarian Research Institute for Digital Transformation, Munich, Germany.

Anne Clausen: PhD candidate at the Digital Democracy Centre, University of Southern Denmark, Odense, Denmark.

Steliyana Doseva: Research associate at the Bavarian Research Institute for Digital Transformation, Munich, Germany.

Jan Eissfeldt: External applied complexity fellow at the Santa Fe Institute, New Mexico, USA.

Joshua Garland: Center director and associate research professor at Arizona State University, Arizona, USA.

Keyan Ghazi-Zahedi: Associated faculty member at Arizona State University, Arizona, USA.

Christian Grimme: Professor of computational social science and systems analysis at the University of Münster, Münster, Germany.

Mario Haim: Professor at the Department of Media and Communication at LMU Munich, Munich, Germany.

Alina Herderich: Postdoctoral research associate at the IDea Lab at the University of Graz, Graz, Austria.

Yuru Li: Postdoctoral research associate at the Institute of Communication Science at Friedrich Schiller University Jena, Jena, Germany.

Hannah Oetting: Research associate and PhD candidate at the Department of Communication at the University of Münster, Münster, Germany.

Cornelius Puschmann: Professor at ZeMKI at the University of Bremen, Bremen, Germany.

Eugenia Rho: Assistant professor at the Department of Computer Science at Virginia Tech, Blacksburg, Virginia, USA.

Anke Stoll: Postdoctoral research associate at the Faculty of Social and Behavioural Sciences at the University of Amsterdam, Amsterdam, the Netherlands.

Andreas Wenninger: Research coordinator and research project leader at the Bavarian Research Institute for Digital Transformation, Munich, Germany.

Marc Ziegele: Professor at the Department of Social Sciences at Heinrich Heine University, Düsseldorf, Germany.

Acknowledgements

The authors would like to thank Abhishek Roy for his valuable contributions to the workshop discussions.

Abstract

This report presents an interdisciplinarily consolidated overview of the risks and potential of artificial intelligence (AI) for counterspeech. Hate speech remains a widespread phenomenon, with severe consequences for individuals and society. Counterspeech by bystanders can mitigate these effects while preserving freedom of expression, but most witnesses do not intervene due to various barriers. Recent AI advancements, particularly large language models (LLMs), offer new possibilities to support counterspeech efforts and help overcome these barriers. However, implementing AI for counterspeech poses immediate and long-term risks that require careful consideration. This report was developed through an iterative process, provides an analysis of key opportunities and challenges, and offers recommendations for researchers, practitioners, and policymakers seeking to leverage AI responsibly for counterspeech.

Contents

1 Key Conclusions

This section synthesises the key conclusions of this interdisciplinary report. Stakeholder-specific implications are addressed comprehensively in the concluding implications section.

Key Conclusion 1:
AI could significantly support counterspeech and counterspeakers

AI tools have the potential to support all phases of the counterspeech process:

- Detection and Decision-Making: AI could help identify harmful content and raise awareness (Stoll et al. 2023), and may encourage bystanders to speak up and underscoring the relevance and effectiveness of counterspeech.
- Implementation: AI could help craft messages tailored to a speaker's goals and even act as a counterspeaker (Bilewicz et al. 2021; Costello et al. 2024).
- Handling the Aftermath: AI could support users after speaking up, for example by processing replies or managing emotional responses (Mun et al. 2024).

Key Conclusion 2:
To fulfil its potential, AI needs high-quality training data, valid and reliable models, access to real-time conversations, and models optimised for providing accurate responses

Realising the potential of AI requires efforts across the full AI pipeline: the data a model is trained on (Jarrahi et al. 2023), the model optimisation process, and how and where it is employed (Heuer et al. 2021; Zajko 2021):

- Clear definitions of hate speech and counterspeech need to be established, informed by systematic research on the harms of different types of hate speech, as well as the differential effectiveness of specific forms of counterspeech.

- Based on that, reliably annotated large-scale datasets for training that account for multiple platforms, languages, and modalities of modern digital communication should be created.

- Persistent weaknesses during data curation and AI training should be eliminated, such as discriminatory bias (Caliskan et al. 2017; Zajko 2021), LLM hallucinations, and the tendency to prioritise persuasiveness over accuracy to prevent the spread of misinformation.

- To ensure quality, corresponding models need to be made available for regular independent audits that account for multiple stakeholder perspectives.

- Individual counterspeakers need the ability to deploy valid, reliable, and transparent models to monitor digital environments. Such implementations, as well as the development and implementation of corresponding techniques, need to be carefully weighed against risks to users' privacy, freedom of speech, and to central values (e.g., human rights).

Key Conclusion 3:
AI bears significant immediate and long-term risks when implemented for counterspeech — and only a part of these risks could be mitigated by addressing technical challenges

The potential of AI must be weighed against a range of significant risks. These risks are not limited to immediate responses but have the potential to shape (digital) public spheres over time (Yakura et al. 2025).

Risks emerging from the technical setup and training of AI:

- Detection systems, for example, can carry built-in biases that reduce their reliability, risking both over- and underreporting. This risk is especially large for identifying subtle forms of hate speech (Schmid et al. 2024a), and in contexts that require cultural sensitivity (Davidson et al. 2019).

- Detection systems can also be abused for surveilling dissidents in authoritarian regimes or to mass-produce hate speech aimed at counterspeakers, making it easier to target those who speak up and discouraging future intervention.

Risks emerging from users' social interactions:

- Technical hurdles or a lack of motivation may prevent many users, particularly those not already inclined to intervene, from adopting AI tools for counterspeech. The added mental effort required to process AI-generated suggestions could overwhelm users, even those who are willing to engage.

- AI-generated counterspeech may lack perceived empathy and authenticity, appearing robotic or impersonal, that can fail to persuade or actively backfire (Mun et al. 2024; Rubin et al. 2025).

- Hateful or misleading counterspeech, whether by humans or AI, can actively diminish the quality of online discussions (Lasser et al. 2025).

- To mitigate potential risks, a comprehensive and ongoing socio-technical assessment of AI tools for counterspeech is needed.

2 Introduction

Many social media users have witnessed hate speech in their digital environments: posts, videos, or memes that spread contempt, hate, or insults targeted at people based on their (perceived) membership in social groups or categories (e.g., gender, race, or sexual orientation). For example, 45% of people in the USA reported witnessing severe online harassment in 2024 (Anti-Defamation League 2024). Hate speech often targets marginalised communities, but the attacks can also be directed against influential voices such as politicians (HateAid 2021).

The consequences can be devastating. Victims suffer from mental stress, online conversations become more aggressive (Cinelli et al. 2021), and bystanders who witness the attack refrain from digital debates out of fear of becoming victimised themselves (Keipi et al. 2017). Besides leading to an impoverishment of digital debates, hate speech can reduce empathy for victims and prosocial behaviour towards the attacked groups (Bilewicz/Soral 2020; Rösner et al. 2016; Ziegele et al. 2018). In extreme cases, it can even lead to offline violence (Leets 2002; Williams et al. 2019). Countering online hate speech is therefore of the utmost relevance.

Every countermeasure must carefully weigh the right to freedom of expression against the potential damaging effect of online hate speech, especially, but not limited to, the personal rights of the affected people. This is particularly true for restrictive interventions such as upload filters, platform bans or other direct interferences with public communication (Funk et al. 2024).

One measure that does not interfere with freedom of expression but builds on the idea of public exchange is counterspeech by bystanders. Counterspeech includes any direct response to hateful or harmful speech that seeks to undermine it (The Dangerous Speech Project n.d.). Counterspeech can improve online discussions (Garland et al. 2022) and reduce negative effects on the targets (Obermaier et al. 2023; Van Houtven et al. 2024).

Counterspeech also plays an increasingly central role in current company and legal frameworks that address hate speech. For example, in the USA, economic pressure and political shifts have led major social media platforms to reduce investments in professional content moderation (LTO 2025), increasingly favouring user-driven mechanisms like Community Notes (e.g., Meta 2025). In the European Union, where laws such as the Digital Services Act (DSA) and the Terrorist Content Online Regulation (TCO), often require platforms to remove illegal or extreme content after it has been reported (although the DSA does not explicitly require such removal in Art. 16(6) DSA, it is assumed in the case of (obviously) illegal content. In addition, according to Art 3(3) TCO, terrorist content must be removed as soon as possible and in any event within one hour of receipt of the removal order by a competent authority of an EU member state.

Counterspeech is also central to addressing the large amount of hostile online content that does not directly transgress legal boundaries (Schmid et al. 2024a) but can still have severe adverse effects. Such “awful but lawful content” is barely covered by legal regulations (an exception is the risk mitigation obligation for Very Large Online Platforms (VLOPs) according to Art 34 and 35 DSA) and remains mainly within the scope of voluntary commitment by platforms via their general terms and conditions (ToS).

By providing counterarguments or uplifting the targets, counterspeakers can mitigate these negative consequences (Obermaier et al. 2023), strengthen civil communication norms, and prevent the deterioration of digital discussions (Garland et al. 2022). However, most bystanders remain silent when witnessing online hate speech. For example, one study showed that only 25% of Germans who saw hate speech online also intervened (Schmid et al. 2024b).

Recent AI developments, specifically LLMs, offer new possibilities to empower bystanders and contribute to constructive digital debates. AI can help detect online hate speech (Stoll 2023; Stoll et al. 2023), produce counterspeech (Bilewicz et al. 2021), and even engage in conversations aimed at reducing potentially harmful beliefs (Costello et al. 2024). However, AI also poses severe risks, including technical risks (e.g., algorithmic bias, hallucination), and social risks such as individuals' overreliance on AI-generated advice (You et al. 2022). There are also potential consequences for how human-made counterspeech is perceived (Mun et al. 2024), and produced (Kosmyna et al. 2025). Finally, several legal questions remain open.

This report is based on an interdisciplinary workshop by the bidt and LMU Munich in March 2025. The workshop brought together 18 experts from Central Europe and the USA, representing the social sciences (anthropology, communication, media psychology), legal studies, and computer science, as well as four practitioners from counterspeech organisations and the tech industry.

The workshop sought to ascertain experts' views on four central topics:

(1) What are the potentials of AI in supporting counterspeech to diminish online hate speech?

(2) What are the challenges for implementing AI in the context of counterspeech?

(3) What are the immediate and long-term risks of AI when implemented to support counterspeech to diminish online hate speech?

(4) What are the implications of these potentials, risks, and open questions for policymakers, regulators, civil society, and internet users interested in AI and counterspeech?

The insights gained from the workshop and a subsequent iterative consensus-building process have been integrated in this report. To reflect the co-creative nature of the report, all workshop participants were added as co-authors. No one decided against this option and all co-authors read and agreed with the report.

The remainder of this report is organized as follows: We first introduce the central concepts relevant to this report (section 2): Online hate speech, counterspeech and AI. Then we summarise what we know (and do not) about the potential promises and pitfalls of AI in counterspeech applications targeted at countering online hate speech (section 3). Finally, we discuss the implications of our comprehensive summary of the experts' knowledge organised by central stakeholder groups (section 4).

3 Central Concepts

3.1 Hate Speech

At its core, hate speech describes communicative attacks on individuals because of their (perceived) membership in one or several social groups or categories, that is their *social identity,* respectively identities (Frischlich 2023). Typical examples include attacks based on people's race, religion, gender, or sexual identity (e.g., United Nations n.d.), or mixtures thereof *(intersectionality).* Hate speech is norm-violating, uncivil (Bormann et al. 2022) and intolerant (Rossini 2020) communication, although not necessarily illegal (there is also not yet a binding legal definition of hate speech). The term is rooted in Critical Race Theory (Matsuda 1989) but has been rapidly extended to include other social groups (Leets/Giles 1997). Societal minorities and disempowered groups are particularly targeted (Silva et al. 2021); however, influential individuals such as politicians or journalists can also become targets (HateAid 2021).

Hate speech has many forms. It can happen both online and offline and can include images, texts, videos, and nonverbal communication. Hate speech can be explicit, including linguistic utterances of verbal violence such as dehumanising metaphors (Frischlich 2023), but can also appear in more subtle and implicit forms, for example, through the promotion of negative stereotypes (Rieger et al. 2021), or sexist memes (Schmid 2025).

Given the diverse forms of hate speech, from implicit to explicit and across varied content, and the interdisciplinary perspectives examining it, prevalence estimates vary significantly across studies. This variation stems from differing definitions of incidents, different times of study but can also be traced back to perceptual differences. Perpetrators might attack someone intentionally or not, those attacked might perceive the attack or not, and observers might detect the attack or not (O'Sullivan/Flanagin 2003).

Cultural and contextual differences can further shape perceptions. Survey results may also reflect respondent bias, as individuals may over- or underestimate exposure depending on their familiarity with different forms of hate speech or desensitisation to recognising it (Schmid et al. 2024a). This underscores the need for precision when discussing or studying the phenomenon but also when applying AI to detect or counter hate speech. Explicit and extreme forms of hate speech are thereby easier to detect and evaluated similarly by different observers (Kümpel/Unkel 2023).

Despite uncertainties around concrete exposure rates, representative surveys consistently indicate that exposure to online hate is a common experience, particularly among young media users. A representative survey in Germany, which has strong regulations against online hatred that constitutes illegal content (the country's previous Network Enforcement Act (NetzDG) from 2017 can be seen as a role model for the DSA), showed that 45% of people aged 16 and older have experienced "hatred on the net" (Bernhard/Ickstadt 2024, 30). Among people aged 16 to 24, 69% reported exposure to online hate. Similar percentages were reported for the USA (Anti-Defamation League 2024).

There is typically less hate speech than non-hateful content in most online discussions. However, prevalence can rise after significant events such as terror attacks (Kaakinen et al. 2021) and varies across contexts. Content analyses of news comments in the USA and Germany showed that 8.6% to 10% of all comments were uncivil as measured by dictionary approaches, with an even smaller proportion qualifying as hate speech (Boberg et al. 2018; Muddiman/Stroud 2017). On social media, that share can be up to 25% (Haim/Hoven 2022), and especially relatively unmoderated spaces such as alternative social media platforms like Telegram or 4chan/8kun (vs. more moderated mainstream social media) allow for more hateful comments (Rieger et al. 2021). Across spaces, implicit hate speech is more prevalent than explicit hate speech (Paasch-Colberg et al. 2021; Rieger et al. 2021).

Detecting implicit hate speech is challenging for humans (Schmid et al. 2024a) as well as artificial systems (Benikova et al. 2018). Especially in ambiguous cases, people might also evaluate the same content differently. The sender of the hateful comments might or might not want to harm the receiver, the receiver might or might not feel harmed, and a bystander might judge the situation differently from either of the two (O'Sullivan et al., 2023).

Hate speech detection is often context-dependent. For instance, the same expression of disgust could be acceptable when discussing food, but problematic when directed at a group of people. The communicative norms are clearer for explicit hate speech, for example most Germans perceive it as substantially more problematic than simple impolite utterances (Kümpel/Unkel 2023).

Hate speech can have severe adverse effects. Potential effects include reduced wellbeing (Keipi et al. 2017; Leets, 2002), reduced social trust (Näsi et al. 2015; Schmid et al. 2024a), silencing (Weeks et al. 2024), reduced empathy (Bilewicz/Soral 2020; Soral et al. 2018) and prosocial behavior (Ziegele et al. 2018), as well as verbal (Álvarez-Benjumea/Winter 2018) and physical violence (Williams et al. 2019). Counterspeech is one measure that can help to reduce these adverse effects on both individuals and societies.

3.2 Counterspeech

Counterspeech includes any direct response to hate speech that seeks to undermine it (The Dangerous Speech Project n.d.). It can be understood as an intentional communicative response, triggered by the hateful attack.

Counterspeech is a process that involves four key phases (Mun et al. 2024). In the *detection phase,* the potential counterspeaker must notice an online incident and recognise it as hate speech that requires intervention. In the *decision-making phase,* the potential counterspeaker must be motivated and willing to act. Both personal and situational factors have to be considered (Latané/Darley 1970; Obermaier et al. 2016). In the *implementation phase,* counterspeakers need to select and formulate their response, considering what they want to achieve (e.g., support the victim, confront the perpetrator). This phase ends with the public or private utterance of the counterspeech. Finally, counterspeakers also have to *handle the aftermath,* including negative backlash, positive feedback, and further discussion.

Like hate speech, counterspeech is not restricted to speech but can have several modalities (e.g., images, text, emojis, memes), occur both online and offline, and take several forms. For example, Buerger (2021) interviewed members of the Swedish #Jagärhär Facebook group. She found that these regular counterspeakers frequently use three specific responses: providing *factual information* to counter attacks or correct misinformation, *criticizing the hateful tone,* and *offering support* to targets. Other forms of counterspeech include voicing *empathy, humour, pointing out contradictions, raising simple opinions,* but also aggressive, uncivil forms of counterspeech such as *insults* (Chung et al. 2019; Hangartner et al. 2021; Lasser et al. 2025).

Overall, counterspeech can be grouped into five broader strategies:

(1) *Empathy-based forms* include demonstrations of and appeals to the perpetrator's cognitive empathy, that is their ability to understand (not necessarily feel) the same as another person (Cuff et al. 2016). Examples include expressing understanding of a perpetrator's emotional state while simultaneously rejecting hate speech, so-called "high person-centred messages" (Masullo et al. 2022), as well as reminders of the targets' feelings (Bilewicz et al. 2021; Hangartner et al. 2021).

(2) *Reason-based forms* provide arguments or factual knowledge (Friess et al. 2021).

(3) *Norm-based strategies* remind the perpetrator of communication norms, for example through appeals to abstract moral virtues such as politeness, or more direct calls to use more adequate language (Bilewicz et al. 2021).

(4) *Sanction-based forms* directly target the perpetrator and use humor or insults to punish the norm transgression (Friess et al. 2021; Jia/Schumann 2025; Ziegele/Jost 2020).

(5) *Support-based forms* encourage victims or express solidarity with them. Table 1 summarises the different strategies and approaches.

Different forms of counterspeech vary in their effectiveness at improving online conversations. Counterspeech that appeals to people's empathy and reason, or that reminds them of civic communication norms, is effective in reducing hateful content (Bilewicz et al. 2021; Friess et al. 2021; Hangartner et al. 2021). Further, messages of support for targets can help mitigate the adverse effects on the victims and reduce the likelihood of escalating hostilities (Obermaier et al. 2023). In contrast, sanctions are less effective (Lasser et al. 2025).

Despite the effectiveness of counterspeech, only a small share of people who witness online hate decides to intervene (Schmid et al. 2024b). Barriers to interventions can emerge across all stages of the process (Buerger 2021; Mun et al. 2024). They can be personal, for example lack of skills or motivation (Obermaier et al. 2025; Schmid et al. 2024b) or situational, for example failing to detect implicit hate speech or hostile and discouraging platform environments (Leonhard et al. 2018; Ping et al., 2024; Sportelli et al., 2025).

Table 1: Exemplary Forms and Approaches of Counterspeech

Counterspeech Forms	Approaches	Examples	References
Empathy-based	High person-centred message	"I recognise that you are angry, but let's try to keep an open mind."	Masullo et al. 2022
	Empathy reminder	"Let's keep in mind that there are people of flesh and blood on the other side of the screen"; "This is just unnecessary hurtful"	Bilewicz et al. 2021; Hangartner et al. 2021
Reason-based	Argumentation	"Please also try to give reasons for your claims. We welcome critical contributions but also want discussions to be respectful and constructive."	Friess et al. 2021, Ziegele/Jost 2020
	Factual knowledge	"Sorry, but you are wrong. Discrimination again. [Gay people] will be perfectly capable of raising a normal child."	Naab et al. 2018
	Calling out disinformation	"These are just insinuations against Muslims that are totally inappropriate and wrong."	Obermaier et al. 2023
Norm-based	Virtue reminder	"Have you ever thought about how this discussion could be more enjoyable for all if we would treat each other with respect"	Bilewicz et al. 2021
	Reminder of relevant others' virtues	"Remember that those you care about can see this post too"	Hangartner et al. 2021
	Call for normative behaviour	"Using kinder words might be the way to go"	Bilewicz et al. 2021
Sanction-based	Implicit sanctions (e.g., humour, sarcasm)	"Please, Sir, stop tweeting" with two funny birds as image; "Of course you don't need to read the article! The countless arguments in your comment already prove that you have studied many reliable sources."	Hangartner et al. 2021; Ziegele/Jost 2020
	Explicit sanctions (e.g., insults)	"You are f**king crazy. Someone like you who just hides behind a keyboard looks like a total jackass"	Jia/Schumann 2025
Support-based	Encouragement	"Refugees are welcome here"	Leonhard et al. 2018

3.3 Artificial Intelligence

AI is "concerned with building smart machines capable of performing tasks that typically require human intelligence" (Shanthi et al. 2023, i). Such tasks can relate to learning, comprehension, problem-solving, decision-making, creativity, and autonomy (Stryker/Kavlakoglu 2024). One of the first legal definitions of AI is provided in Art 3 No 1 of the EU's Artificial Intelligence Act of 2024 (AI Act). It defines an AI system as a machine-based system designed to operate with varying levels of autonomy, that may adapt after deployment, and infers from input how to generate outputs such as predictions, content, recommendations, or decisions that influence both physical and virtual environments.

For counterspeech we can distinguish between two functional AI categories: **discriminative** AI systems are optimised to categorise data (e.g., analyse text, images or other online data to detect hate speech) and **generative** AI systems are typically trained on a large amount of internet texts or images and are optimised to generate text (or other content), such as a counterspeech response. Both have promises and risks.

Despite its name, AI (be it discriminative or generative) is, however, not comparable to human intelligence (Van Rooij et al. 2024). A widely cited paper in psychology (Gottfredson 1997, 13) describes human intelligence as a "general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience". Simply put, people with higher intelligence (as measured through the IQ score) tend to be better at all these different cognitive tasks (they might still be less able than others to do other things, of course).

This conceptualisation is not uncontroversial and other authors have argued for understanding human intelligence (also) in terms of adaptation to different environments (Sternberg 2021), as a multifaceted trait (Guilford 1968) and account for the distinction between processing speed (so-called "fluid intelligence") and world knowledge ("crystalline intelligence") as well as for the substantial influence of the environment on how human intelligence can unfold (Nisbett et al. 2012). Discussing these different conceptualisations in detail is beyond the scope of this report.

In contrast to intelligent humans, AI is typically very good at some tasks, but a model that is brilliant at computing the most complex mathematical models can still fail at detecting hate speech correctly. At its core, AI is a mathematical model that is optimised to reduce prediction errors. These models do not possess what might be conceived of as general intelligence (g) or cognitive intent. Even with the same task, small changes can lead to significant differences. For example, the same AI tool might detect 90% of hate speech in one language (such as English) correctly, but find only 60% in a smaller language, and can be rapidly overwhelmed when languages are mixed, such as in India, where people often mix Hindi with English in the same sentence (Mundra et al. 2021).

At the core of automating tasks that typically require human intelligence lies an algorithm, with varying degrees of complexity — a computation sequence that links one or more inputs (e.g., a racist slur) to an output (e.g., the classification of a comment as hate speech) — through statistical models. In its simplest form, *rule-based programming,* such an algorithm can take the general form of an "if a then b" statement. For example, a statement could be classified as hateful when it includes words that have been defined as racist, sexist or otherwise (Hayaty et al. 2020).

The same logic can be used to formulate more complex rules, for example by specifying conditions under which a statement is considered explicit hate speech versus implicit hate speech (for a manual codebook measuring this distinction, see Rieger et al. 2021). The benefit of such rule-based algorithms is that they are transparent and often easier to explain than other forms of AI (e.g., Cruz-Filipe et al. 2024). However, they require covering all potential cases, making the process resource-intensive and inflexible. Such simple analyses often perform poorly when it comes to more implicit forms of hate speech (Assenmacher et al. 2025), and they risk missing crucial context information (e.g., negations such as "no one should call you a slut").

3.3.1 Machine Learning

Machine learning promises to overcome these limitations (Chollet/Allaire 2018). In machine learning, the algorithm "learns" from a given dataset how to link the input (e.g., the words in a social media post) to the output (e.g., classification as hate speech). This learning process can be further distinguished into *unsupervised, supervised,* and *reinforcement learning* (Russell & Norvig 2010). In practice, these distinctions can be less clear-cut (e.g., in semi-supervised learning), but for the sake of this report, we focus on introducing the three main types.

Unsupervised machine learning works with data that has not been classified by a human before, and the algorithm learns patterns in the data without explicit input during this learning process (Russell & Norvig, 2010). Typical examples include cluster analysis, which aims to identify homogenous yet distinct groups of, for example, online posts; or topic modelling, which aims to identify groups of words (the topics) that often occur together in a dataset (Haim, 2023). Crucially, determining the "correct" number of topics, preparing the data, and validating the topics and clusters always require human oversight, as corresponding models are often very good at sorting large datasets but not always at finding meaningful categories (Hennig, 2015). Consider, for example, a dataset that includes different symbols that can be either black or white (Figure 1). An unsupervised machine learning process could sort them either by symbol or by colour. However, both solutions might be differentially meaningful in a specific context.

Supervised machine learning starts with a dataset that already includes the outcome of interest, for example a set of posts that human coders have labelled as hateful or not (Chung et al., 2019), or going back to Figure 1, as being black or white symbols. Supervised machine learning has two central phases: An exploratory training phase, during which the objective is to find the best model to predict the outcome of interest (e.g., whether a post is hateful or a symbol's colour), and a confirmatory evaluation phase during which this model is tested on another dataset that has not been used for training the model (Yarkoni & Westfall, 2017). This enables the researcher to predict the so-called "out-of-sample error" — a measure of how well the model works in data it has not been trained on. That is important, as a model that is "tailored" perfectly to the training data can perform poorly on other samples, just as a dress tailored to fit a specific person often does not fit other people.

During the training phase, the dataset is split into a training and a test set. The latter is put aside. The researcher then selects the range of characteristics, or features of the data that the algorithm can consider (so-called feature engineering, Figure 1). For example, one could decide whether words, images, or emojis should be considered when predicting if a post is hateful.

Afterwards, researchers often compare the performance of different algorithms (i.e., mathematical models) and settings to find the best model. A good model classifies a large number of posts correctly and only a few or ideally no posts incorrectly (according to the initial label). For example, posts that do not contain hate speech are classified as no-hate, whereas posts that do contain it are judged as hate speech (Figure 1). Afterwards, the model is applied to the test data to calculate the out-of-sample error.

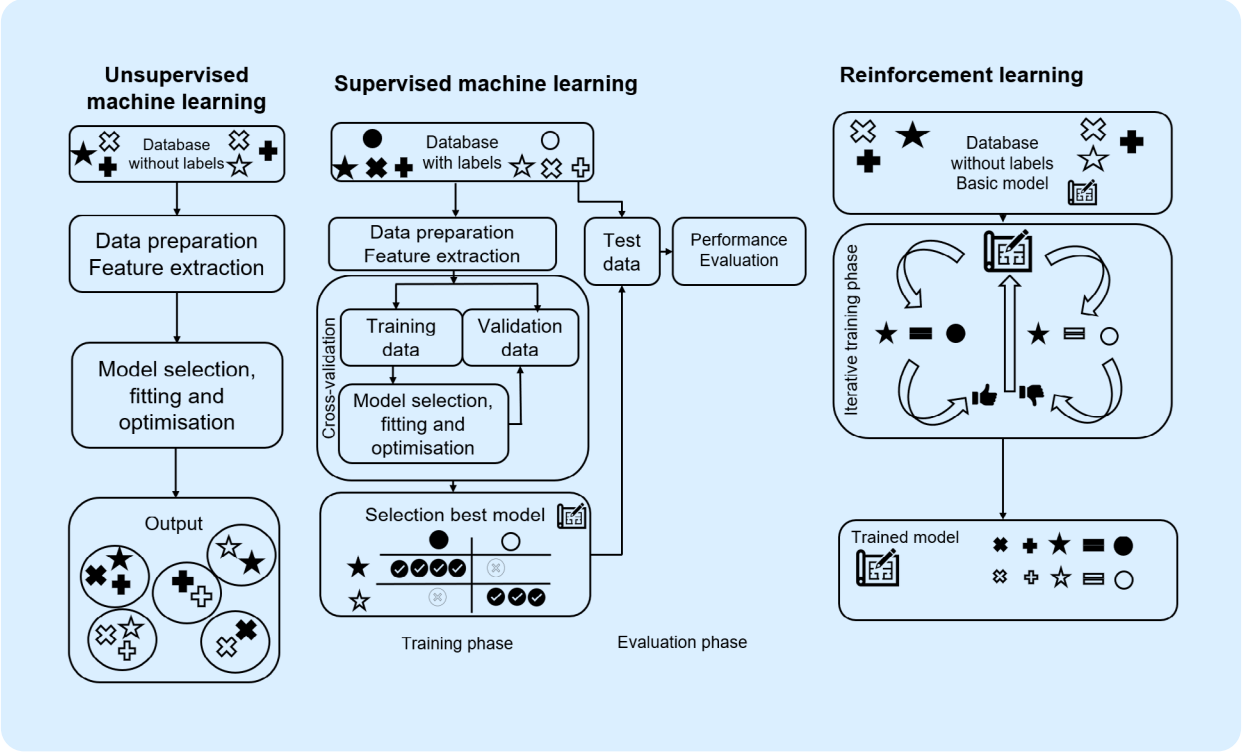
To ensure that the best model is not accidentally only the best because of how the training and test datasets were split, this process (training/test split, model selection) can be run several times using subsets of the data — a process called cross-validation. The training phase ends with selecting the best model to predict the outcome of interest, which is the one that classifies the content of interest as well as possible, has low out-of-sample error, and performs well across different data subsets.

In the final evaluation phase, the best model is applied to the test data. This allows the final out-of-sample error to be calculated. If the model also performs well on this sample (i.e., correctly classifies posts as hateful or symbols as black or white), it can be used to classify data that has not been manually labelled. Due to this training process, the quality of supervised learning outcomes crucially depends on the quality of their "gold-standard" training data. Such datasets are often resource-intensive to create (Kang 2023).

In **reinforcement learning,** the algorithm learns from a series of rewards or punishments for "correct" or "false" decisions without the need to label the data in advance. Russell and Norvig (2010) use the example of a player who is put in a situation in which he or she must play a new game with unknown rules. The player tries out all kinds of actions and gets feedback on either having "won" or "lost" the game. Based on this feedback (the reinforcement), the player tries to figure out the game's rules. Applied to the detection of hate speech, we can conceptualise the player as a virtual agent who tries out various ways to classify content as hateful, for example relying on the length of a text, or the presence of swear words. Based on another algorithm, the "reinforcer," the agent receives feedback on whether the classification action was successful. Through this trial-and-error approach, the algorithm can learn complex relations between input and output without requiring hand-labelled data (Russell/Norvig 2010). Figure 1 visualises the different learning approaches.

Figure 1: Simplified Depiction of Different Machine Learning Pipelines.

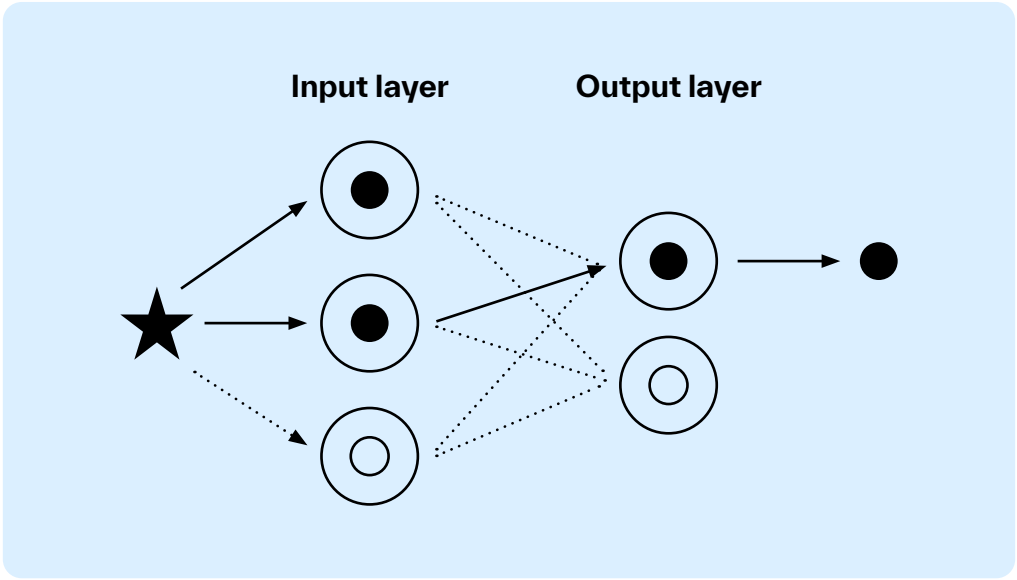
The symbols indicate the data classification that should be learned — in this case the colour of the symbols.



3.3.2 Deep Learning and Generative AI

Many recent AI models rely on **deep learning.** Deep learning is based on so-called *deep neural networks,* a technology inspired by the human brain. The human brain processes information via specific cells, the so-called neurons, that are connected and communicate via electrical impulses (Pinel/Barnes 2021). Deep neural networks are built on the idea of an artificial neuron that receives input from other artificial neurons in the form of numbers and signals this input to other neurons (McCulloch/Pitts 1943). This idea was first translated into a simple neural network, the *perceptron* (Rosenblatt 1958), which was designed with one input layer containing several such artificial neurons and one output layer. Analogous to the processes in the human brain, neurons in the input layer activate the layers in the output layer when the "correct" signal is detected, such as when a symbol is black instead of white (Figure 2).

Figure 2: Abstract Depiction of a Simple Neural Network with one Input and one Output Layer.



Nowadays, deep networks do not contain only an input and an output layer but several interconnected hidden layers (sometimes millions) between the input and the output in which more detailed information can be processed. Like reinforcement learning, deep learning is based on trial-and-error. After each trial, the output layer gives feedback to the model about its performance. The feedback is used to adapt the model for the next round of training, so-called "backpropagation" (Rumelhart et al. 1986). Over multiple rounds, deep learning models can thereby learn to detect complex patterns in data such as text, images, or other material.

Often, deep neural networks are combined with reinforcement learning (Matsuo et al. 2022), particularly when the goal is to make incremental improvements (i.e., finetuning the model) rather than training an entire model from scratch. For example, responding to ChatGPT messages with a thumbs-up or a thumbs-down provides human feedback to finetune the model further.

Deep learning also underlies another central process that is crucial for generative AI — **transformer models** (Vaswani et al. 2017). These models combine advanced methods for text processing via deep learning with an attention mechanism that provides finegrained context sensitivity, allowing them not only to predict the next word in a sentence but also to take the word's context into account. To inform generative AI applications such as ChatGPT, these transformer models are trained on very large amounts of data. Thus, they are often also referred to as *large language models* or *LLMs.*

Importantly, recently popular LLMs such as OpenAI's GPT family, Google's Gemini models, Meta's Llama, or open alternatives such as Mistral are developed and deployed as foundation models that can "be reused or repurposed across a broad range of tasks with minimal adaptation [by learning] to perform new tasks from input data" (Scott/Zuccon 2024, 705). Even though these models do not require labelled data, their quality nonetheless depends crucially on the data they have been trained on. Even the most advanced deep learning network can only work with the data it has received and been trained on. And while LLMs can mimic the syntax of complex argumentation, they lack the semantic understanding of social harm.

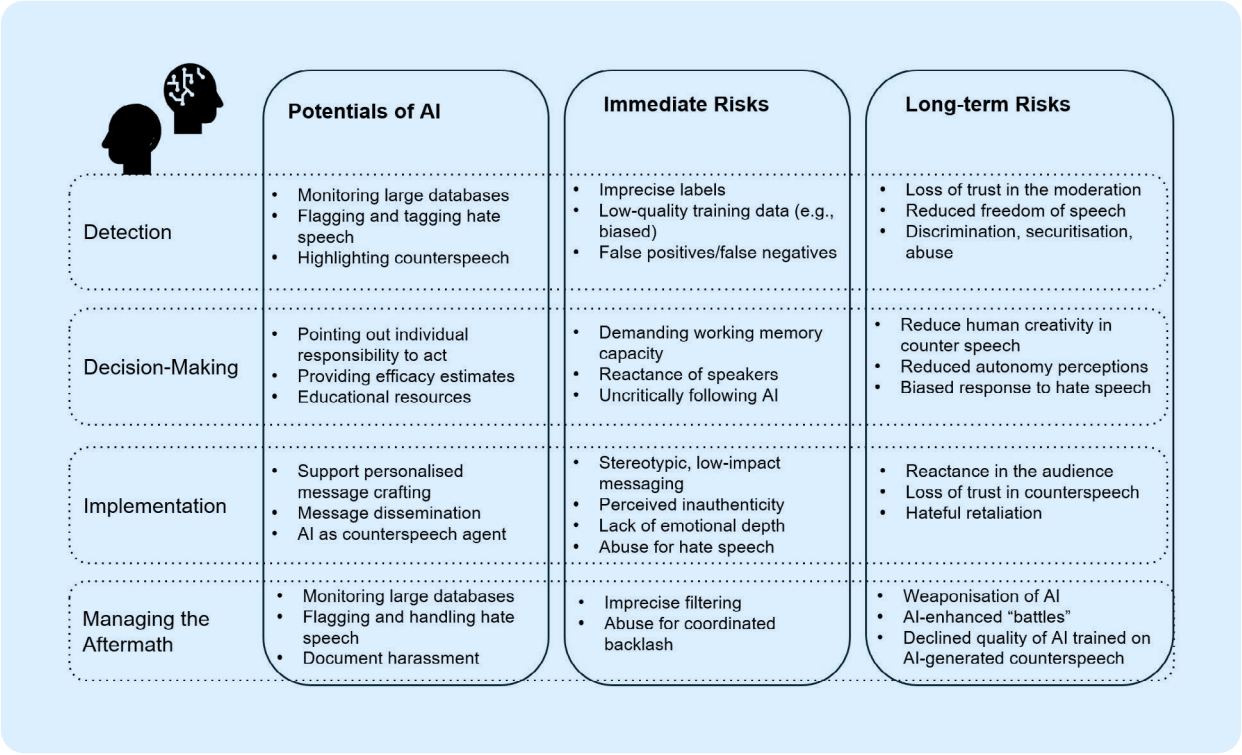
4 What We (Don't) Know About the Promises and Pitfalls of AI for Counterspeech

AI tools can potentially support users throughout each stage of the counterspeech process. During detection, AI can help monitor large amounts of data and identify hate speech that requires countering. In the decision-making phase, AI can offer encouragement to motivate action or support decisions about which action to take. In the implementation phase, AI could assist in writing effective, evidence-based counterspeech tailored to user goals, whether to support a target or foster constructive dialogue. To manage the aftermath, AI could support monitoring and processing reactions, or alert other counterspeakers to mobilise further support.

At all these stages, AI also carries specific risks that require careful assessment of its promises and pitfalls to ensure ethically responsible deployment. These risks can be grounded in the technical setup (i.e., the training data, the implemented model), but also in the social response to AI (e.g., blind trust in AI recommendations vs. loss of trust in the authenticity of digital communication). Furthermore, several questions of scientific rigour, such as whether AI systems are replicable, reliable, transparent and objective, remain open. Figure 3 summarises the promises and both immediate and long-term risks of AI in the counterspeech process. We will discuss these in detail in the following section.

It is important to recognise that many of the risks reflect broader concerns about AI and LLMs and that not all risks in the context of counterspeech are unique to AI. Ineffective counterspeech, communication missteps, or conflicting goals among counterspeakers can occur regardless of whether the message is generated by a human or a machine.

Figure 3: The Promises and Pitfalls of AI in the Counterspeech Process.



4.1 Detection of Hate Speech and Harm Using AI Support

To decide whether to step in against online hate speech, bystanders must first recognise a specific communication as a hateful norm transgression and understand the harm it can cause, for instance to targets or to the respectful tone of online conversations (Latané/Darley 1970; Obermaier et al. 2016; Ziegele et al. 2020).

However, implicit hate speech, such as humorous memes (Schmid et al. 2024a), is hard to detect, even for humans. Experienced counterspeakers often say that a major challenge is identifying the most harmful posts to respond to, since it is nearly impossible to sift through the huge amount of content manually (Mun et al. 2024). Adding to the difficulty, what people find offensive or harmful can vary depending on context. A comment that is viewed as hate speech on one platform or in one country might be seen as normal or harmless in another (Coe et al. 2014; Gibson 2019).

Potentials

Discriminative AI can potentially support hate speech identification and can be trained to provide estimates of its harmfulness. For example, AI could serve as an early warning system — a smart filter that classifies and flags potentially harmful content and highlights patterns of hateful behaviour online (Stoll et al. 2023). Major online platforms already use AI tools to filter and delete unwanted and hateful content, especially content prohibited by their ToS (Kühling 2023).

LLMs, in particular, are good at breaking down complex or hostile messages, helping to simplify them and point out the harmful parts (Van Royen et al. 2022). This makes it easier for bystanders to recognise when online content includes hate speech and deserves attention. With this kind of help, everyday users and regular counterspeakers could respond more effectively to hate speech without feeling overwhelmed by the flood of online content.

By clearly flagging hate speech, AI could also help shift how people perceive it. When bystanders realise that hateful comments are not the norm in communication but the acts of a loud minority, they may feel more empowered to speak out and support a more respectful, benign online dialogue. Clearly highlighting hateful content could also help make platform or context policies more explicit to all users, serve as a norm reminder, and be part of the counterspeech process. On the contrary, publicly flagging hate speech could result in a backfire effect of drawing more attention to it, in effect promoting hateful messages.

AI tools could also shine a light on thoughtful counterspeech that is already happening, making it easier for others to support those messages, even if they are not confident enough to speak up themselves. To this end, the algorithm would need to combine the detection of hate speech with the detection of counterspeech, potentially using metadata about the speaker (e.g., membership of a counterspeech group such as #IamHere) as additional features.

Challenges

Training AI to detect online hate speech is a challenging task, largely because of how detection systems are trained. A key ingredient in supervised machine learning (and most deep learning) is a high-quality "gold standard" dataset for training. Creating such a dataset requires a clear, universally accepted definition and a valid, reliable human agreement on what constitutes hate speech.

However, the field of hate speech research is plagued by a variety of partially overlapping terms and concepts (Paasch-Colberg et al. 2023). Some definitions focus on tone or word choice, such as the use of capital letters, emotional language, or exaggeration, while others look at different forms of harm or offense (e.g., Bormann/Ziegele 2023; Masullo Chen et al. 2019; Muddiman 2017; Papacharissi 2004; Rossini 2021; Sobieraj/Berry 2011; Stroud et al. 2015). These challenges are even bigger when counterspeech is to be detected. Due to the inherent relative nature of counterspeech and its partially dialogic form, detecting counterspeech solely based on message content, without accounting for the broader context (including hate speech) and different types of countermessages, risks undermining the validity of the resulting classification.

The labelling process has other inherent challenges. The formulation and perception of hateful comments is culture- and context-dependent, and high-quality datasets and models developed in one language or based on data from one platform do not automatically work in other language spaces or cultural contexts. In addition, machine learning models require large to massive datasets for effective training. They are often best trained on larger languages (particularly English), leading to lower performance and the need for finetuning or even retraining for smaller languages (e.g., Hashmi et al. 2024).

Human coders' characteristics also influence detection. For example, women and men were found to sometimes disagree on the hatefulness of implicit (but not explicit) sexist hate speech (Hettiachchi et al. 2023; Wojatzki et al. 2018). Creating the training datasets is also ethically complicated. Human annotators have to review large amounts of offensive and upsetting content, which can have severe psychological impacts on their well-being (AlEmadi/Zaghouani 2024). Finally, concept drift might result in lower model performance when trained on older data. To illustrate, new hateful terms such as *Lügenpresse* (German for "lying press") emerge over time with the progression of cultural and political events.

Risks

Employing AI to detect hate speech entails several short- and long-term risks arising from the socio-technological interplay between humans and machines. Because of the wide range of definitions, any dataset used to train AI is inevitably based on a small subset of these ideas. The labelling process in machine learning often relies on majority voting, so that only posts where several coders agreed on their hateful nature end up being classified in the dataset. Resulting detection systems carry the same biases as the data they were trained on. This can cause AI to focus too much on particular styles or types of hate speech, while missing others, especially those that are more subtle or context-dependent (Zhang et al. 2024), or are understood best by the attacked minorities.

When AI systems misclassify content, it may not only affect moderation decisions in individual cases. Over time, it may damage internet users' perceptions of the overall fairness and credibility of counterspeech, potentially leading to lower levels of trust. Ill-implemented solutions also risk threatening civic rights such as freedom of speech and can contribute to the discrimination of marginalised groups. For example, Black-American speech is classified more often as offensive than White-American speech (Davidson et al. 2019; Sap et al. 2022). This can lead to the securitisation of cultural practices, as common ways of communicating are increasingly viewed through a lens of danger and security (Eroukhmanoff, 2017).

4.2 Decision-Making: Fostering Responsibility and Efficacy Using AI Support

Even when bystanders or habitual counterspeakers recognise online hate speech and perceive it as potentially harmful, their decision to intervene depends on further factors. Most importantly, they need to feel personally responsible to act, believe that counterspeech can make a difference and feel confident in their ability to respond effectively (Dovidio et al. 2006; Latané/Darley 1970; Obermaier et al. 2016; Ziegele et al. 2020).

Certain situations can weaken that sense of responsibility. When many people witness an incident online, they often assume someone else will speak up, and so nobody does. This is known as the *online bystander effect* (Obermaier et al. 2016). Likewise, if counterspeech has already been posted, some may think the problem has been addressed, even though research shows that continued counterspeech can further encourage others to join in and help create a more respectful conversation (Bilewicz et al. 2021; Garland et al. 2020; Leonhard et al. 2018; Naab et al. 2018).

People may also hold back from intervening if they believe their actions will not help. They may feel the conversation will not change, the target will not feel supported, or they might fear negative consequences, like personal attacks. On the other hand, feeling supported by others — through encouragement, likes or positive comments — can make people more likely to speak out (Obermaier 2023; Ziegele et al. 2020). Just as importantly, knowing what to say and how to say it boosts both confidence and the willingness to speak out (Obermaier et al. 2025).

Potentials

Whether built into platforms, offered as browser extensions, or available as standalone apps, AI tools can be designed to encourage bystanders to take action. For example, algorithms might display encouraging prompts like "every voice matters" or suggest in the middle of a comment thread that a new perspective could be helpful. AI could also help users judge when and how to intervene by offering insights into the tone and reach of ongoing discussions, for instance highlighting how widespread a hateful comment is or how aggressive the language has become.

Another effective strategy is showing users real examples of successful counterspeech: stories where a supportive comment made a difference by comforting a target, swaying bystanders, and motivating them to speak up as well, or contributing to a more respectful dialogue. Platforms themselves could play a role by using features that signal which counterspeech efforts the community appreciates, such as upvotes that boost visibility, likes that show appreciation, or comments that stay at the top of the discussion. This kind of social validation can make others more likely to get involved (e.g., Walther 2024). Educational tools powered by AI could also inform about different forms of online hate and offer practical strategies for responding, while being clear about potential risks and rewards (Ding et al. 2024).

Challenges

Despite their potential, such AI support would often depend on users taking active steps, such as installing an app, enabling a browser extension, or paying attention to onscreen prompts. This means they may mostly reach people who are already motivated to engage, rather than drawing in new voices. Additionally, these tools could unintentionally add to users' mental burden. The extra information they provide, such as analysis of tone or reach, might overwhelm users' already strained working memory capacity, resulting in lower acceptance rates (Frischlich et al. 2024). While counterspeech is generally desirable, such prompts shift the burden of responsibility onto users rather than the platforms on which the discourse takes place.

Risks

If not carefully designed, nudges could encourage surface-level engagement, known as cognitive loafing (Weldon/Gargano 1988). In these cases, users might quickly accept AI-generated suggestions without thinking critically (You et al. 2022) or feeling ownership over their words. This can reduce the depth, relevance, and impact of counterspeech (Kosmyna et al. 2025). There is also a risk that overly prescriptive prompts or advice could backfire. Instead of encouraging action, they might trigger resistance, known as reactance (Miron/Brehm 2006) or make users feel that their freedom to express themselves is being limited — ultimately stifling the creativity and autonomy that make human counterspeech effective.

4.3 AI Support in Counterspeech Implementation

Once someone has identified hate speech and decided to respond, there are still potential barriers that can inhibit potential counterspeakers. Producing thoughtful, context-specific counterspeech often requires time and emotional labour — resources that are not always available, especially when hate spreads rapidly across platforms. Generative AI has the potential to offer real value in addressing these barriers. For example, Mun and colleagues (2024) found that a common barrier to posting (more) counterspeech is the time required to craft a convincing message. AI-based tools could help overcome this barrier by providing tailored suggestions.

While using AI to generate counterspeech is still in its early stages, recent research offers reasons for both careful optimism and caution. Much of the work so far has focused on developing high-quality datasets, exploring the best methods for generating effective responses, and determining what "effective" even means in this context. Studies such as those by Ashida and Komachi (2022) and Bonaldi and colleagues (2022) have made important contributions by evaluating the quality of AI-generated counterspeech and by constructing datasets that more closely reflect the multiturn nature of online dialogue.

As another example, Chung and colleagues (2019) focused on representing different types of counterspeech, annotating a dataset representing, for example, the provision of facts, pointing out hypocrisy or contradiction, and humour in counterspeech. These efforts are crucial in helping machines learn to intervene in ways that are informative, specific, and inoffensive. However, they are often limited to a few languages and seldom reflect the multimodal, multi-channel structure of the digital environments users navigate today. Researchers are also still largely the ones evaluating whether automated counterspeech is "effective", rather than testing its actual impact on audiences.

Only a handful of studies have begun to test how actual users respond to AI-generated messages. Bilewicz and colleagues (2021) conducted an intervention study on Reddit using a bot that monitored r/MensRights and r/TooAfraidToAsk and responded to personal attacks with one of three types of messages. Responses included disapproval, appeals to social norms, or empathetic reminders to encourage respectful dialogue. The study provides early evidence that AI-generated interventions can reduce the share of verbal aggression in online environments, suggesting that automated counterspeech may help de-escalate hostile discourse and foster more respectful interactions.

Another relevant study is Costello and colleagues (2024), in which the researchers found that in-depth conversational dialogues with AI can durably reduce belief in conspiracy theories. Users described a conspiracy theory they believed in, whereafter a chatbot engaged them in respectful (and prolonged) conversations that challenged their beliefs. The AI helped decrease conspiratorial thinking both immediately and several months later. Engaging in such in-depth, sustained, personalised conversations about conspiracy beliefs at scale would be impossible for human counterspeakers due to time constraints alone, while an AI-driven chatbot was successful.

The implementation phase carries both the greatest promise and the biggest risk of using AI to write counterspeech. AI can dramatically scale the reach of counterspeech interventions, addressing one of the field's most persistent challenges, namely volume. Human counter-speakers simply cannot keep pace with the proliferation of online incivility. On top of that, conversations with AI agents are in principle non-judgemental, potentially increasing the likelihood of perpetrators engaging in the conversation and changing their beliefs.

Challenges

There are substantial challenges and risks to successfully implementing AI-generated counterspeech. Besides the already-mentioned challenges of correctly detecting hate speech and the lack of evidence-based datasets for training, AI-generated (vs. human) counterspeech might be less effective in general. For example, people rate AI-generated and AI-supported content as less empathic than content generated purely by humans (Rubin et al. 2025). This raises meaningful and urgent questions about the socio-psychological processes that, alongside the technical questions, shape the effectiveness of AI-generated and AI-enhanced counterspeech. Bär and colleagues (2024) also expressed caution that generative AI may backfire. In their large-scale experiment on X (formerly Twitter), counterspeech that was perceived as inauthentic or that led users to feel deceived provoked backlash and further entrenched hateful views.

Risks

There are risks inherent in LLMs that have severe abuse potential and, in the worst case, can contribute to the deterioration of digital discourse. In a pre-print covering three large-scale experiments with crowd workers, Hackenburg and colleagues (2025) exposed over 79,000 participants to conversations about politics with LLMs. The authors noted that these conversations were on average more persuasive than static messages (+2.94 percentage points on a 1-100 scale). Persuasive effects were larger for larger models or models that were finetuned, or post-trained to be persuasive. In line with the elaboration likelihood model of persuasion (Petty/Cacioppo 1986), these increased with information the model provided. Yet, while most of the content provided by the LLMs was accurate (81%), AI models were better at being persuasive than accurate overall. In the communication of the most persuasive model and state-of-the-art models (e.g., ChatGPT 5) one in three fact-checkable claims turned out to be misinformation. The quality also depends on the model's training. Elon Musk's Grok (in)famously spread antisemitic and racist hate speech itself (Snoswell 2025).

We have very little evidence of how the use of AI-generated counterspeech at scale could impact larger social media ecosystems. For example, there is a risk that flooding platforms with AI-generated messages could either dilute genuine engagement or trigger backlash, especially if the messages are seen as inauthentic or intrusive. There is a certain danger to contribute to "reverse censorship" or "censorship by noise" (Tufekci 2017), drowning out genuine conversations through the sheer amount of content.

AI-generated counterspeech could also trigger reactance when being perceived as patronising or reduce people's trust in digital communication more generally. Studies have shown that people who are more familiar with X (formerly Twitter) were more likely to perceive accounts issuing statements they disagree with as "social bots" (Wischnewski et al. 2021). Given the widespread use of LLMs (e.g., 67% of Germans; see Bitkom 2025), similar biases will likely shape perceptions of digital messages, particularly counter-attitudinal ones such as counterspeech. Another question concerns the labelling of AI-enhanced social media contributions. Given that AI models could support the creation of effective counterspeech messages, at which level of human-AI collaboration would content flags be required and where would they not be needed?

As we move forward, we must not only improve the technical quality of automated responses but also deepen our understanding of how those responses are received, and which ethical frameworks should guide their deployment. We urgently need a better understanding of which processes guide effective counterspeech, which can be augmented or scaled using AI, and how effective they are for different audiences. For example, AI-generated counterspeech might be less effective via the "empathy" route but, if factually accurate, could allow for scaling reason-based counterspeech targeted towards the perpetrators of the attack.

There could also be other effects on perceived communication norms or targets. As such, a division of labour approach that is informed by both risks and potential seems most appropriate. For example, AI could support formulating logical or de-escalating counterspeech or patiently point out contradictions, while human authenticity is needed for empathic counterspeech, solidarity with targets and the enactment and validation of civic human communication norms.

4.4 AI Support in Handling the Aftermath of Counterspeech

Online counterspeech can lead to a wide range of outcomes — some intended by those who speak up, and others completely unexpected. These effects can impact everyone involved in the situation: the person spreading hate, the targets, and the bystanders watching the exchange (Obermaier et al. 2023; Ziegele et al. 2020). Counterspeech may encourage the offender to rethink their views, or it could backfire, triggering an even more aggressive response towards the victim or the person who intervened. It may offer emotional support to the target, helping them feel seen and defended, or it might not change their experience at all or be perceived as a shallow expression of sympathy directed at the perpetrator.

Bystanders can also react in different ways. Some may feel empowered to step in themselves or adopt the viewpoint expressed in the counterspeech. Others may remain passive or scroll past without engaging. In short, counterspeech can make a meaningful impact, but it can also fall flat or have unintended consequences, depending on how it is delivered and received.

People who engage in counterspeech are left to deal with its aftermath, whether it is rewarding or draining. AI tools can potentially shape this post-counterspeech landscape in two very different directions: decreasing negative backlash and supporting counterspeakers.

AI tools are being developed and deployed that *promise* to help users manage and respond to such backlash. For instance, technical features using discriminative AI such as Instagram's Hidden Words feature allows users to automatically filter offensive or abusive comments, preventing counterspeakers from having to read them in the first place. Platforms like YouTube and TikTok similarly offer personal moderation tools that can detect and block harmful language. On the generative AI side, LLMs could help users craft calm, clear responses in the face of hostility, or even provide supportive or affirming messages for people who engage in counterspeech (Mun et al. 2024).

Looking ahead, there could be potential to build AI companions that help counterspeakers document harassment, report abuse more efficiently (for example to obtain legal assistance) or simply process the emotional toll of their advocacy. In fact, tools like Spot (↗ talktospot.com) already exist to help people document and report workplace harassment. Other examples include Block Party, which helps users establish baseline security, provides expert recommendations in the case of attacks, and informs close others about the attacks (↗ https://www.blockpartyapp.com).

Challenges

For these tools to be effective and ethical, they must be developed with the input of the communities most affected, and designed not just to defend against harm, but to affirm and sustain the courage it takes to speak up. Moreover, the tools have to be transparent. Similar tools that, for example, allow reporting hate speech can also become targets of coordinated attacks aimed at mass-reporting counterspeakers or other users who caught a perpetrator's attention. A specific challenge in the context of counterspeech in multimodal, high-choice platform environments is that negative responses might appear across various social media platforms and not immediately catch a user's attention. While AI's ability to monitor data at scale could be valuable, there are still the central challenges of data access, as well as the need to adapt detection tools to the specific context discussed before.

Risks

Besides the already discussed risks emerging from mislabelling speech as hateful or harmful when it is not, bad actors could also weaponise generative AI to escalate negative outcomes like online hate speech. Personalised memes, manipulated images, deepfakes and rapid-fire comment floods can now be generated and disseminated at scale with minimal effort, making it easier than ever to retaliate against people who speak up.

In this way, the use of AI could increase the risks of counterspeech, making some people more hesitant to intervene again. Another risk arises from AI's dependency on training data. Current large language models often learn on online data, and an increasing amount of this data is AI-generated. This amplifies the potential risks of using AI for counterspeech (such as bias and hallucinations) as newer models learn from the mistakes of their predecessors. Unlike humans, they do not necessarily become better at avoiding these mistakes as their decisions are ultimately based on frequencies and statistical decisions.

5 Implications

AI bears both great potential and severe risks in the context of counterspeech. In the following section, we discuss the core implications of these observations for policymakers, AI developers, civic society, and research.

5.1 Implications for Policymakers and Regulation

To effectively foster counterspeech by online users to address hate speech with the support of AI tools, policymakers, and law enforcement must take a comprehensive approach that prioritises transparency and user empowerment. A first step would be to implement a comprehensive, uniform legal definition of hate speech (cf., for example, Section 192a of the German Criminal Code, which punishes hate-mongering insults).

The legal framework for using AI in the context of counterspeech must be communicated accurately to the parties concerned — especially since the freedom of speech of the person uttering hate speech can be affected depending on how the AI tool interferes in the communication process. European legal frameworks such as the Digital Service Act (DSA), the Terrorist Content Online (TCO), and the AI Act already provide several crucial regulations, tackling, for example the need for platforms to address illegal hate speech and terrorist content, and researchers' rights to access data (Art 40 DSA).

For example, the voluntary EU Code of Conduct on countering illegal hate speech online from 2016 has been integrated into the DSA since 2025, making it a code of conduct according to Art 45 DSA. This provides it with a more binding legal nature than a voluntary framework of self-regulation (from now on Code of Conduct+).

VLOPs that signed the code explicitly prohibit *illegal* hate speech in their terms-of-service review, valid notifications (e.g., by users) about illegal hate speech in a "timely, diligent, non-arbitrary and objective manner" (Clause 2.2, Code of Conduct+), and remove content that violates their policies or applicable law. At least 50% of the content should be processed in the first 24 hours after reporting (Clause 2.3, Code of Conduct+). Terrorist content must be removed within one hour after a removal order (Art 3(3) TCO).

The AI Act prohibits the use, dissemination, and sale of AI services that exploit the vulnerabilities of specific people or social groups, and that motivate or encourage behaviour that causes harm to those people or others (Art 5, AI Act). Yet, to reach users and lead to meaningful change, these frameworks must not only be refined by court decisions; they also require an active public that is aware of its individual rights and user-friendly mechanisms to claim them. We thus make three suggestions.

a) Harmonise legal frameworks and establish clear standards for online hate speech

Hate speech, counterspeech, and LLMs transgress national boundaries and legal frameworks in the digital realm. This poses significant challenges for lawmakers and leads to a scattered landscape burdened by various definitions and partially conflicting regulations emerging from national laws and supranational frameworks such as the European Union. Big platforms and AI companies have their own (business) interests and, therefore, oftentimes varying terms of service that can contribute to confusion on the side of the users as well as AI developers. Hate speech also flourishes on smaller platforms that are not covered by the Code of Conduct+ or the risk mitigation provisions of the DSA. Ignoring such smaller platforms risks weakening perceived norms around online hate speech in users' eyes.

Attempts to harmonise scattered legislation can be a meaningful step in providing regulatory clarity when it comes to explicit and illegal forms of hate speech. Such attempts must consider cultural and legal differences across regions, recognising that definitions and perceptions of hate speech can vary significantly worldwide and may evolve over time. Clear definitions of which aspects of speech fall under the regulation and which do not are thus pivotal for enabling the detection and deletion of hate speech.

b) Ensure data access while protecting individual rights for privacy and ownership

Ideally, legal frameworks, court decisions, and real cases relevant to hate speech should be available to train AI models. Corresponding data should be made available under the FAIR (findable, accessible, interoperable and reusable) principles, while balancing the privacy rights of those who utter hateful online content, for example by national data stewards ensuring safe, privacy-protecting access.

Access to diverse and representative datasets is essential for training reliable tools and to enable researchers and regulators to properly evaluate how enforcement works in practice. We also encourage initiatives that help users better understand how and what their data is used for and its effects. This includes user-friendly access to the data stored about them but could extend to information on how the platforms' AIs evaluate their own content.

At the same time, AI development is currently driven by the exploitation of user data at scale and the active circumvention of both privacy and proprietary rights. Big tech companies such as Meta have already indicated that they will not sign the voluntary EU Code of Conduct (Weatherbed 2025), demanding respect for copyright. Legal frameworks intended to enable data access need to balance digital innovation with the protection of legitimate business interests, while safeguarding individual citizens and counterspeakers.

c) Aim for a whole-of-society approach to digital regulations

To support and strengthen the counterspeech process, integrating AI tools into social media platforms can play a key role, especially in helping civil society actors, counter-speakers, and everyday users detect online hate speech more effectively. Doing so responsibly requires more than just technical solutions. We recommend that policy-makers adopt a process-based approach to regulation — one that looks not only at what content is removed but also at how decisions are made, why, by whom, and under what oversight.

Civil society organisations must be formally recognised as co-regulatory partners, not treated as afterthoughts. This means giving them a meaningful role in shaping and overseeing digital regulation, ensuring that they can both comply with the law and hold other actors accountable. Their inclusion must be embedded in law, not left to voluntary consultation. All regulatory frameworks must be anchored in the Charter of Fundamental Rights, which should serve as the foundation for the design, use, and governance of digital tools. These structures must be resilient enough to withstand efforts to undermine democratic principles and free expression.

Ultimately, a shared responsibility model is needed — one in which governments, platforms, and users have a clearly defined role in promoting respectful online discourse. Regulation should not only support the development of ethical and effective AI tools, but also guard against their misuse. Counterspeech efforts must be grounded in transparency, user consent, and the fundamental right to speak and be heard.

To this end, lawmakers must communicate openly with civil society, and strict enforcement of existing provisions, such as Articles 37 and 40 of the DSA, which require independent audits and platform data access, is non-negotiable. New legislation, like the AI Act, must also ensure that AI systems are explainable (XAI), so that their decisions can be understood, traced, and challenged when necessary.

5.2 Implications for AI Developers and Providers

AI systems can meaningfully support counterspeech and have great potential to contribute to a benign digital public sphere that users want to participate in. However, to fulfil these potentials, AI tools must overcome several challenges, requiring developers and providers to step in.

Hate speech and counterspeech are complex social constructs that span both explicit and implicit expressions, and several modalities. We thus encourage developers to engage in interdisciplinary work to harness the strengths of disciplines such as the humanities or the social sciences and develop tools that are based on solid research and valid and reliable classifications.

These tools must be thoughtfully designed, implemented, and evaluated to ensure that they help users intervene constructively. Promising examples include interface nudges that foster a sense of personal responsibility, such as prompts that encourage users to contribute constructively or approval mechanisms that highlight and reward thoughtful engagement. Other helpful tools could include personal moderation dashboards, affirmation systems that offer encouragement after speaking up, or mechanisms that provide peer and institutional support.

Crucially, these AI tools should be seen as trusted complements to human responses, not zero-shot replacements. Early research suggests that human-generated counterspeech often has a stronger emotional impact than AI-generated messages (Rubin et al. 2025), especially when empathy, nuance, or lived experience matters. Messages that feel genuine are more likely to resonate and inspire action, while overly polished or formulaic AI responses can come across as inauthentic — raising concerns about the diminishing returns of mass-produced counterspeech. At scale, the effectiveness of AI-generated responses may plateau, or even backfire, if they are perceived as inauthentic or manipulative (Mun et al. 2024).

Additionally, AI tools designed to support counterspeech and reporting must be accessible, intuitive, and protected from being undermined by complex or misleading platform terms of service. To ensure that these tools work in practice and not just in theory, it is essential to involve the people who use them the most directly: community managers, regular counterspeakers, and everyday users. Their input and experience are invaluable in evaluating the practicality and effectiveness of AI tools in supporting counterspeech.

5.3 Implications for Civil Society

AI systems can meaningfully support counterspeech and counterspeakers, but there are also both immediate and long-term risks and several open questions that require users and counterspeakers to employ these tools with care and caution. To effectively foster counterspeech with the support of AI tools, it is essential for civil society actors, along with counterspeakers and everyday users, to take an active role. We have three central suggestions:

a) Use AI to support human engagement, not to replace it

Relying on fully automated or unsupervised AI systems to generate counterspeech on their own can lead to errors, unintended consequences, or a reduced feeling of personal responsibility for intervention among counterspeakers. Instead, civil society and internet users must advocate for policies that promote "machine-in-the-loop" approaches. In these systems, AI supports and enhances human judgement rather than replacing it. AI tools should be designed to empower users, helping them respond more effectively and safely. They should not dictate what people say, as the goal is to amplify human agency, not automate it. For an example of how such a system could look in the domain of emotion regulation, see Sharma and colleagues (2023), who designed and deployed an application to improve people's own reframed suggestions for negative situations.

b) Know your own biases

Research has shown that we tend to trust AI recommendations, even when we know AI makes mistakes (You et al. 2022). Popular LLMs still provide a lot of factually wrong claims (Hackenburg et al. 2025), but people tend to think that it is only others who fall for misinformation (the *third-person effect,* e.g., Jang/Kim 2018; French et al. 2023). It is realistic to assume that these biases affect us all, and we should take them into account when using AI to alert us to hate speech or to formulate counterspeech. Reflecting carefully about the messages we send (e.g., by reading them aloud to ourselves or imagining that we would send them to a beloved family member) could help us think more carefully about the content and thus can help us make decisions that do not only feel more authentic to us, but are also more closely aligned with our values and priorities (cf. Schmitt et al. 2018; Van Royen et al. 2022).

c) Hold platforms and law enforcement accountable

Civil society should hold both platforms and law enforcement accountable for how AI tools that monitor digital communication, classify hate speech and support counterspeech are developed and regulated. This means demanding transparency — not only about how these systems work, but also about who designs them, what data they are trained on, and how decisions are made around content detection. Civil society has a critical role to play in pushing for greater openness, especially from law enforcement. Users will then understand the systems that shape their online experiences better and will be able to push for change when those systems fall short.

5.4 Implications for Future Research

When it comes to AI and counterspeech, several questions remain open and urgently require intensified research efforts that bridge multiple disciplines. We want to highlight the need for evidence on the harm caused by different types of hate speech, informing definitional approaches and legal frameworks as well as evidence on the effectiveness of different counterspeech strategies.

We are only at the beginning of understanding the processes underlying effective counterspeech. AI-enhanced research methods can provide meaningful insights into these questions but do not reduce the importance of nuanced qualitative and rigorous quantitative approaches following best ethical practices for working with human data. We also need to understand better how digital environments can be designed to create constructive interaction spaces.

Finally, most of the research on AI and counterspeech is rooted in the Global North. Research on mis- and disinformation interventions has demonstrated the need to test the generalisability of interventions in traditionally underresearched communities (Blair et al. 2024). Including global perspectives is even more relevant for counterspeech and the broader development of AI.

6 References

Al Emadi, M. M./Zaghouani, W. (2024). Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data. In: Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024, 66–72, Torino, Italy. ELRA and ICCL. ↗ https://aclanthology.org/2024.legal-1.10.pdf [retrieved 08.04.2026].

Álvarez-Benjumea, A./Winter, F. (2018). Normative Change and Culture of Hate: An Experiment in Online Environments. In: European Sociological Review 34 (3), 223–237. ↗ https://doi.org/10.1093/esr/jcy005.

Anti-Defamation League (2024). Online Hate and Harassment: The American Experience 2024. ADL Center for Technology & Society. ↗ https://www.adl.org/sites/default/files/documents/2024-06/online-hate-and-harassment-the-american-experience-v2024.pdf [retrieved 08.04.2026].

Ashida, M./Komachi, M. (2022). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 11–23. Seattle, WA, USA. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2022.woah-1.2 [retrieved 08.04.2026].

Assenmacher, D. et al. (2025). Beyond the Explicit: A Bilingual Dataset for Dehumanization Detection in Social Media. In: arXiv. ↗ https://doi.org/10.48550/arXiv.2510.18582.

Bär, D./Maarouf, A./Feuerriegel, S. (2024). Generative AI May Backfire for Counterspeech. In: arXiv. ↗ https://doi.org/10.48550/arXiv.2411.14986.

Benikova, D./Wojatzki, M./Zesch, T. (2018). What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In: Rehm, G./Declerck, T. (eds.). Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science(), vol 10713. Cham, 171–179. ↗ https://doi.org/10.1007/978-3-319-73706-5_14.

Bernhard, L./Ickstadt, L. (2024). Lauter Hass – leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung. Das NETTZ/Gesellschaft für Medienpädagogik und Kommunikationskultur/HateAid/Neue deutsche Medienmacher*innen als Teil des Kompetenznetzwerks gegen Hass im Netz. ↗ https://kompetenznetzwerk-hass-im-netz.de/wp-content/uploads/2024/02/Studie_Lauter-Hass-leiser-Rueckzug.pdf [retrieved 08.04.2026].

Bilewicz, M./Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. In: Political Psychology 41 (S1), 3–33. ↗ https://doi.org/10.1111/pops.12670.

Bilewicz, M. et al. (2021). Artificial Intelligence Against Hate: Intervention Reducing Verbal Aggression in the Social Network Environment. In: Aggressive Behavior 47 (3), 260–266. ↗ https://doi.org/10.1002/ab.21948.

Bitkom e.V. (2025). KI-Nutzung boomt – aber die Angst vor Abhängigkeit vom Ausland ist groß. ↗ https://www.bitkom.org/Presse/Presseinformation/KI-Nutzung-boomt-Angst-vor-Abhaengigkeit-Ausland-gross [retrieved 08.04.2026].

Blair, R. A. et al. (2024). Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South. In: Current Opinion in Psychology 55, 101732. ↗ https://doi.org/10.1016/j.copsyc.2023.101732.

Boberg, S. et al. (2018). The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum. In: Media and Communication 6 (4), 58–69. ↗ https://doi.org/10.17645/mac.v6i4.1493.

Bonaldi, H. et al. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2022.emnlp-main.549.

Bormann, M. et al. (2022). Incivility as a Violation of Communication Norms. A Typology Based on Normative Expectations toward Political Communication. In: Communication Theory 32 (3), 332–362. ↗ https://doi.org/10.1093/ct/qtab018.

Bormann, M./Ziegele, M. (2023). Incivility. In: Strippel, C. et al. (eds.). Challenges and Perspectives of Hate Speech Research. Berlin, 199–217. ↗ https://doi.org/10.48541/dcr.v12.12.

Buerger, C. (2021). #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. In: Social Media + Society 7 (4). ↗ https://doi.org/10.1177/20563051211063843.

Caliskan, A./Bryson, J. J./Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. In: Science 356 (6334), 183–186. ↗ https://doi.org/10.1126/science.aal4230.

Chollet, F./Allaire, J. J. (2018). Deep Learning mit R und Keras: Das Praxis-Handbuch von den Entwicklern von Keras und RStudio. Frechen.

Chung, Y. L. et al. (2019). CONAN – COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech. In: arXiv. ↗ https://doi.org/10.48550/arXiv.1910.03270.

Cinelli, M. et al. (2021). Dynamics of Online Hate and Misinformation. In: Scientific Reports 11 (1), 22083. ↗ https://doi.org/10.1038/s41598-021-01487-w.

Coe, K./Kenski, K./Rains, S. A. (2014). Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. In: Journal of Communication 64 (4), 658–679. ↗ https://doi.org/10.1111/jcom.12104.

Costello, T. H./Pennycook, G./Rand, D. G. (2024). Durably Reducing Conspiracy Beliefs through Dialogues with AI. In: Science 385 (6714), eadq1814. ↗ https://doi.org/10.1126/science.adq1814.

Cruz-Filipe, L./Gaspar, G./Nunes, I. (2024). Hypothetical Answers to Continuous Queries over Data Streams. In: ACM Transactions on Computational Logic 25 (4), 1–40. ↗ https://doi.org/10.1145/3688845.

Cuff, B. M. P. et al. (2016). Empathy: A Review of the Concept. In: Emotion Review 8 (2), 144–153. ↗ https://doi.org/10.1177/1754073914558466.

Dangerous Speech Project (n.d.). What Is Counterspeech? ↗ https://www.dangerousspeech.org/counterspeech [retrieved 08.04.2026].

Davidson, T./Bhattacharya, D./Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In: Proceedings of the Third Workshop on Abusive Language Online, 25–35, Florence, Italy. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/W19-3504.

Ding, X. et al. (2024). CounterQuill: Investigating the Potential of Human-AI Collaboration in Online Counterspeech Writing. In: arXiv. ↗ https://doi.org/10.48550/arXiv.2410.03032.

Dovidio, J. F. et al. (2006). The Social Psychology of Prosocial Behavior. Hillsdale, New Jersey.

Eroukhmanoff, C. (2017). Securitisation Theory: An Introduction. In: MgGlinchey, S./Walters, R./Scheinpflug, C. (eds.). International Relations Theory. Great Britain, 104–109.

French, A. M./Storey, V. C./Wallace, L. (2023). The Impact of Cognitive Biases on the Believability of Fake News. In: European Journal of Information Systems 34 (1), 72–93. ↗ https://doi.org/10.1080/0960085X.2023.2272608.

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. In: Political Communication 38 (5), 624–646. ↗ https://doi.org/10.1080/10584609.2020.1830322.

Frischlich, L. (2023). Hate and Harm. In: Strippel, C. et al. (eds.). Challenges and Perspectives of Hate Speech Research. Berlin, 165–183. ↗ https://doi.org/10.48541/dcr.v12.10.

Frischlich, L. et al. (2024). Fighting Fakes on WhatsApp – Audience Perspectives on Fact Bots as Countermeasures. In: Digital Journalism 12 (5), 700–720. ↗ https://doi.org/10.1080/21670811.2024.2341299.

Funk, V. et al. (eds.) (2024). Freedom on the Net 2024. Freedom House. ↗ https://freedomhouse.org/sites/default/files/2024-10/FREEDOM-ON-THE-NET-2024-DIGITAL-BOOKLET.pdf [retrieved 08.04.2026].

Garland, J. et al. (2020). Countering Hate on Social Media: Large Scale Classification of Hate and Counter Speech. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, 102–112, Online. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2020.alw-1.123.

Garland, J. et al. (2022). Impact and Dynamics of Hate and Counter Speech Online. In: EPJ Data Science 11 (1), 3. ↗ https://doi.org/10.1140/epjds/s13688-021-00314-6.

Gibson, A. (2019). Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. In: Social Media + Society 5 (1). ↗ https://doi.org/10.1177/2056305119832588.

Gottfredson, L. S. (1997). Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. In: Intelligence 24 (1), 13–23. ↗ https://doi.org/10.1016/S0160-2896(97)90011-8.

Guilford, J. P. (1968). Intelligence Has Three Facets. In: Science 160 (3828), 615–620. ↗ https://doi.org/10.1126/science.160.3828.615.

Hackenburg, K. et al. (2025). Comparing the Persuasiveness of Role-Playing Large Language Models and Human Experts on Polarized US Political Issues. In: AI & Society 41 (1), 351–361. ↗ https://doi.org/10.1007/s00146-025-02464-x.

Haim, M. (2023). Computational Communication Science: Eine Einführung. Wiesbaden.

Haim, M./Hoven, E. (2022). Hate Speech's Double Damage: A Semi-Automated Approach toward Direct and Indirect Targets. In: Journal of Quantitative Description: Digital Media 2, 1–37. ↗ https://doi.org/10.51685/jqd.2022.009.

Hangartner, D. et al. (2021). Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment. In: Proceedings of the National Academy of Sciences of the United States of America 118 (50), e2116310118. ↗ https://doi.org/10.1073/pnas.2116310118.

Hashmi, E. et al. (2024). Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration. In: IEEE Access 12, 121507–121537. ↗ https://doi.org/10.1109/ACCESS.2024.3452987.

HateAid (2021). Boundless Hate on the Internet. Dramatic Situation across Europe. ↗ https://hateaid.org/wp-content/uploads/2022/04/HateAid-Report-2021_EN.pdf [retrieved 08.04.2026].

Hayaty, M./Adi, S./Hartanto, A. D. (2020). Lexicon-Based Indonesian Local Language Abusive Words Dictionary to Detect Hate Speech in Social Media. In: Journal of Information Systems Engineering and Business Intelligence 6 (1), 9–17. ↗ https://doi.org/10.20473/jisebi.6.1.9-17.

Hennig, C. (2015). What Are the True Clusters? In: Pattern Recognition Letters 64, 53–62. ↗ https://doi.org/10.1016/j.patrec.2015.04.009.

Hettiachichi, D. et al. (2023). How Crowd Worker Factors Influence Subjective Annotations: A Study Tagging Misogynistic Hate Speech in Tweets. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 11 (1), 38–50. ↗ https://doi.org/10.1609/hcomp.v11i1.27546.

Heuer, H./Jarke, J./Breiter, A. (2021). Machine Learning in Tutorials – Universal Applicability, Underinformed Application, and Other Misconceptions. In: Big Data & Society 8 (1). ↗ https://doi.org/10.1177/20539517211017593.

Jang, S. M./Kim, J. K. (2018). Third Person Effects of Fake News: Fake News Regulation and Media Literacy Interventions. In: Computers in Human Behavior 80, 295–302. ↗ https://doi.org/10.1016/j.chb.2017.11.034.

Jarrahi, M. H./Memariani, A./Guha, S. (2023). The Principles of Data-Centric AI. In: Communications of the ACM 66 (8), 84–92. ↗ https://doi.org/10.1145/3571724.

Jia, Y./Schumann, S. (2025). Tackling Hate Speech Online: The Effect of Counter-Speech on Subsequent Bystander Behavioral Intentions. In: Cyberpsychology: Journal of Psychosocial Research on Cyberspace 19 (1), Article 4. ↗ https://doi.org/10.5817/CP2025-1-4.

Kaakinen, M. et al. (2021). Online Hate and Zeitgeist of Fear: A Five-Country Longitudinal Analysis of Hate Exposure and Fear of Terrorism after the Paris Terrorist Attacks in 2015. In: Political Psychology 42 (6), 1019–1035. ↗ https://doi.org/10.1111/pops.12732.

Kang, E. B. (2023). Ground Truth Tracings (GTT): On the Epistemic Limits of Machine Learning. In: Big Data & Society 10 (1). ↗ https://doi.org/10.1177/20539517221146122.

Keipi, T. et al. (2017). Online Hate and Harmful Content: Cross-National Perspectives. London.

Kosmyna, N. et al. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task. In: arXiv. ↗ https://doi.org/10.48550/arXiv.2506.08872.

Kühling, J. (2023). Der Einsatz von Künstlicher Intelligenz durch Unternehmen und Aufsichtsbehörden bei der Bekämpfung von Hassrede. In: Zeitschrift für Urheber- und Medienrecht (ZUM) 2023, 566–573.

Kümpel, A. S./Unkel, J. (2023). Differential Perceptions of and Reactions to Incivil and Intolerant User Comments. In: Journal of Computer-Mediated Communication 28 (4), zmad018. ↗ https://doi.org/10.1093/jcmc/zmad018.

Lasser, J. et al. (2025). Collective Moderation of Hate, Toxicity, and Extremity in Online Discussions. In: PNAS Nexus 4 (11), pgaf369. ↗ https://doi.org/10.1093/pnasnexus/pgaf369.

Latané, B./Darley, J. M. (1970). The Unresponsive Bystander: Why Doesn't He Help? New York, NY.

Leets, L. (2002). Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Antigay Speech. In: Journal of Social Issues 58 (2), 341–361. ↗ https://doi.org/10.1111/1540-4560.00264.

Leets, L./Giles, H. (1997). Words as Weapons – When Do They Wound? Investigations of Harmful Speech. In: Human Communication Research 24 (2), 260–301. ↗ https://doi.org/10.1111/j.1468-2958.1997.tb00415.x.

Legal Tribune Online (2025). Meta schafft Faktenchecker und Moderation ab. ↗ https://www.lto.de/recht/nachrichten/n/meta-moderation-faktenchecker-digital-services-act-zuckerberg-kehrtwende [retrieved 08.04.2026].

Leonhard, L. et al. (2018). Perceiving Threat and Feeling Responsible: How Severity of Hate Speech, Number of Bystanders, and Prior Reactions of Others Affect Bystanders' Intention to Counterargue against Hate Speech on Facebook. In: Studies in Communication and Media 7 (4), 555–579. ↗ https://doi.org/10.5771/2192-4007-2018-4-555.

Masullo Chen, G. et al. (2019). We Should Not Get Rid of Incivility Online. In: Social Media + Society 5 (3). ↗ https://doi.org/10.1177/2056305119862641.

Masullo, G. M./Riedl, M. J./Huang, Q. E. (2022). Engagement Moderation: What Journalists Should Say to Improve Online Discussions. In: Journalism Practice 16 (4), 738–754. ↗ https://doi.org/10.1080/17512786.2020.1808858.

Matsuda, M. J. (1989). Public Response to Racist Speech: Considering the Victim's Story. In: Michigan Law Review 87 (8), 2320–2381. ↗ https://doi.org/10.2307/1289306.

Matsuo, Y. et al. (2022). Deep Learning, Reinforcement Learning, and World Models. In: Neural Networks 152, 267–275. ↗ https://doi.org/10.1016/j.neunet.2022.03.037.

McCulloch, W. S./Pitts, W. H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. In: Bulletin of Mathematical Biophysics 5 (4), 115–133. ↗ https://doi.org/10.1007/BF02478259.

Meta (2025). Testing Begins for Community Notes on Facebook, Instagram and Threads. Meta Newsroom, 13.03.2025. ↗ https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/ [retrieved 08.04.2026].

Miron, A. M./Brehm, J. W. (2006). Reactance Theory – 40 Years Later. In: Zeitschrift für Sozialpsychologie 37 (1), 9–18. ↗ https://doi.org/10.1024/0044-3514.37.1.9.

Muddiman, A. (2017). Personal and Public Levels of Political Incivility. In: International Journal of Communication 11, 3182–3202.

Muddiman, A./Stroud, N. J. (2017). News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. In: Journal of Communication 67 (4), 586–609. ↗ https://doi.org/10.1111/jcom.12312.

Mun, J. et al. (2024). Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–22, New York, NY, USA. Association for Computing Machinery. ↗ https://doi.org/10.1145/3613904.3642025.

Mundra, S./Singh, N./Mittal, N. (2021). Fine-Tune BERT to Classify Hate Speech in Hindi English Code-Mixed Text. In: FIRE, 330–337. ↗ https://ceur-ws.org/Vol-3159/T1-32.pdf [retrieved 08.04.2026].

Naab, T. K./Kalch, A./Meitz, T. G. (2018). Flagging Uncivil User Comments: Effects of Intervention Information, Type of Victim, and Response Comments on Bystander Behavior. In: New Media & Society 20 (2), 777–795. ↗ https://doi.org/10.1177/1461444816670923.

Näsi, M. et al. (2015). Exposure to Online Hate Material and Social Trust among Finnish Youth. In: Information Technology & People 28 (3), 607–622. ↗ https://doi.org/10.1108/ITP-09-2014-0198.

Nisbett, R. E. et al. (2012). Intelligence: New Findings and Theoretical Developments. In: American Psychologist 67 (2), 130–159. ↗ https://doi.org/10.1037/a0026699.

Obermaier, M. (2023). Once Bitten, Twice Shy? Costs and Benefits Experiences Explaining Bystander Intervention against Online Hate Speech. In: 73rd Annual Conference of the ICA, Toronto, Canada.

Obermaier, M./Fawzi, N./Koch, T. (2016). Bystanding or Standing By? How the Number of Bystanders Affects the Intention to Intervene in Cyberbullying. In: New Media & Society 18 (8), 1491–1507. ↗ https://doi.org/10.1177/1461444814563519.

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. In: 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. In: New Media & Society 25 (9), 2339–2358.

O'Sullivan, L. F. et al. (2023). Did You Help? Intervening during Incidents of Sexual Assault among College Student Bystanders. In: Journal of College Student Development 64 (2), 208–224. ↗ https://doi.org/10.1353/csd.2023.0018.

O'Sullivan, P. B./Flanagin, A. J. (2003). Reconceptualizing "Flaming" and Other Problematic Messages. In: New Media & Society 5 (1), 69–94. ↗ https://doi.org/10.1177/1461444803005001908.

Paasch-Colberg, S. et al. (2023). Sharing Is Caring: Addressing Shared Issues and Challenges in Hate Speech Research. In: Strippel, C. et al. (eds.). Challenges and Perspectives of Hate Speech Research. Berlin, 11–22. ↗ https://doi.org/10.48541/dcr.v12.1.

Paasch-Colberg, S. et al. (2021). From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration. In: Media and Communication 9 (1), 171–180. ↗ https://doi.org/10.17645/mac.v9i1.3399.

Papacharissi, Z. (2004). Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups. In: New Media & Society 6 (2), 259–283. ↗ https://doi.org/10.1177/1461444804041444.

Petty, R. E./Cacioppo, J. T. (1986). Communication and Persuasion. Central and Peripheral Routes to Attitude Change. New York, NY.

Pinel, J. P./Barnes, S. J. (2021). Biopsychology Global Edition. London.

Ping, K. et al. (2025). Behind the Counter: Exploring the Motivations and Barriers of Online Counterspeech Writing. In: ACM Transactions on Computer-Human Interaction 32 (5), Article 55. ↗ https://doi.org/10.1145/3745769.

Rieger, D. et al. (2021). Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. In: Social Media + Society 7 (4). ↗ https://doi.org/10.1177/20563051211052906.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. In: Psychological Review 65 (6), 386–408. ↗ https://doi.org/10.1037/h0042519.

Rösner, L./Winter, S./Krämer, N. C. (2016). Dangerous Minds? Effects of Uncivil Online Comments on Aggressive Cognitions, Emotions, and Behavior. In: Computers in Human Behavior 58, 461–470. ↗ https://doi.org/10.1016/j.chb.2016.01.022.

Rossini, P. (2020). Beyond Toxicity in the Online Public Sphere: Understanding Incivility in Online Political Talk. In: Dutton, W. H. (ed.). A Research Agenda for Digital Politics. Cheltenham, 160–170. ↗ https://doi.org/10.4337/9781789903096.00026.

Rossini, P. (2021). More Than Just Shouting? Distinguishing Interpersonal-Directed and Elite-Directed Incivility in Online Political Talk. In: Social Media + Society 7 (2). ↗ https://doi.org/10.1177/20563051211008827.

Rubin, M. et al. (2025). Comparing the Value of Perceived Human versus AI-Generated Empathy. In: Nature Human Behaviour 9 (11), 2345–2359. ↗ https://doi.org/10.1038/s41562-025-02247-w.

Rumelhart, D. E./Hinton, G. E./Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. In: Nature 323 (6088), 533–536. ↗ https://doi.org/10.1038/323533a0.

Russell, S./Norvig, P. (2010). Artificial Intelligence – A Modern Approach. London.

Sap, M. et al. (2022). Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 5884–5906, Seattle, WA, USA. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2022.naacl-main.431.

Schmid, U. K. (2025). Humorous Hate Speech on Social Media: A Mixed-Methods Investigation of Users' Perceptions and Processing of Hateful Memes. In: New Media & Society 27 (3), 1588–1606. ↗ https://doi.org/10.1177/14614448231198169.

Schmid, U. K./Kümpel, A. S./Rieger, D. (2024a). How Social Media Users Perceive Different Forms of Online Hate Speech: A Qualitative Multi-Method Study. In: New Media & Society 26 (5), 2614–2632. ↗ https://doi.org/10.1177/14614448221091185.

Schmid, U. K./Obermaier, M./Rieger, D. (2024b). Who Cares? How Personal Political Characteristics Are Related to Online Counteractions against Hate Speech. In: Human Communication Research 50 (3), 393–403. ↗ https://doi.org/10.1093/hcr/hqae004.

Schmitt, J. B. et al. (2018). Critical Media Literacy and Islamist Online Propaganda: The Feasibility, Applicability and Impact of Three Learning Arrangements. In: International Journal of Conflict and Violence 12, 1–19. ↗ https://doi.org/10.4119/UNIBI/ijcv.642.

Scott, I. A./Zuccon, G. (2024). The New Paradigm in Machine Learning – Foundation Models, Large Language Models and Beyond: A Primer for Physicians. In: Internal Medicine Journal 54 (5), 705–715. ↗ https://doi.org/10.1111/imj.16393.

Shanthi, D./Madhuravani, B./Humar, A. (2023). Handbook of Artificial Intelligence. Sharjah.

Sharma, A. et al. (2023). Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 9977–10000, Toronto, Canada. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2023.acl-long.555.

Silva, L. et al. (2021). Analyzing the Targets of Hate in Online Social Media. In: Proceedings of the International AAAI Conference on Web and Social Media 10 (1), 687–690. ↗ https://doi.org/10.1609/icwsm.v10i1.14811.

Snoswell, A. J. (2025). How Do You Stop an AI Model Turning Nazi? What the Grok Drama Reveals about AI Training. In: The Conversation, 14.07.2025. ↗ http://theconversation.com/how-do-you-stop-an-ai-model-turning-nazi-what-the-grok-drama-reveals-about-ai-training-261001 [retrieved 08.04.2026].

Sobieraj, S./Berry, J. M. (2011). From Incivility to Outrage: Political Discourse in Blogs, Talk Radio, and Cable News. In: Political Communication 28 (1), 19–41. ↗ https://doi.org/10.1080/10584609.2010.542360.

Soral, W./Bilewicz, M./Winiewski, M. (2018). Exposure to Hate Speech Increases Prejudice through Desensitization. In: Aggressive Behavior 44 (2), 136–146. ↗ https://doi.org/10.1002/ab.21737.

Sportelli, C. et al. (2025). "Let's Make the Difference!" Promoting Hate Counter-Speech in Adolescence through Empathy and Digital Intergroup Contact. In: Journal of Community & Applied Social Psychology 35 (1), e70028. ↗ https://doi.org/10.1002/casp.70028.

Sternberg, R. J. (2021). Adaptive Intelligence: Intelligence Is Not a Personal Trait but Rather a Person × Task × Situation Interaction. In: Journal of Intelligence 9 (4), 58. ↗ https://doi.org/10.3390/jintelligence9040058.

Stoll, A. (2023). The Accuracy Trap: Or How to Build a Phony Classifier. In: Strippel, C. et al. (eds.). Challenges and Perspectives of Hate Speech Research. Berlin, 371–381. ↗ https://doi.org/10.48541/dcr.v12.22.

Stoll, A./Wilms, L./Ziegele, M. (2023). Developing an Incivility Dictionary for German Online Discussions – A Semi-Automated Approach Combining Human and Artificial Knowledge. In: Communication Methods and Measures 17 (2), 131–149. ↗ https://doi.org/10.1080/19312458.2023.2166028.

Stroud, N. J. et al. (2015). Changing Deliberative Norms on News Organizations' Facebook Sites. In: Journal of Computer-Mediated Communication 20 (2), 188–203. ↗ https://doi.org/10.1111/jcc4.12104.

Stryker, C./Kavlakoglu, E. (2024). What Is Artificial Intelligence (AI)? ↗ https://www.ibm.com/think/topics/artificial-intelligence [retrieved 08.04.2026].

Tufekci, Z. (2017). Twitter and Tear Gas: The Power and Fragility of Networked Protest. New Haven, CT.

United Nations (n.d.). Understanding Hate Speech. ↗ https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech [retrieved 08.04.2026].

Van Houtven, E. et al. (2024). "You Got My Back?" Severity and Counter-Speech in Online Hate Speech toward Minority Groups. In: Media Psychology 27 (6), 923–954. ↗ https://doi.org/10.1080/15213269.2023.2298684.

Van Rooij, I. et al. (2024). Reclaiming AI as a Theoretical Tool for Cognitive Science. In: Computational Brain & Behavior 7 (4), 616–636. ↗ https://doi.org/10.1007/s42113-024-00217-5.

Van Royen, K. et al. (2022). Think Twice to Be Nice? A User Experience Study on a Reflective Interface to Reduce Cyber Harassment on Social Networking Sites. In: International Journal of Bullying Prevention 4 (1), 23–34. ↗ https://doi.org/10.1007/s42380-021-00101-x.

Vaswani, S. et al. (2017). Model-Independent Online Learning for Influence Maximization. In: Proceedings of Machine Learning Research (PMLR), 3530–3539. ↗ https://proceedings.mlr.press/v70/vaswani17a.html [retrieved 08.04.2026].

Walther, J. B. (2024). The Effects of Social Approval Signals on the Production of Online Hate: A Theoretical Explication. In: Communication Research. ↗ https://doi.org/10.1177/00936502241278944.

Weatherbed, J. (2025). Meta Snubs the EU's Voluntary AI Guidelines. In: The Verge, 21.07.2025. ↗ https://www.theverge.com/news/710576/meta-eu-ai-act-code-of-practice-agreement [retrieved 08.04.2026].

Weeks, B. E./Halversen, A./Neubaum, G. (2024). Too Scared to Share? Fear of Social Sanctions for Political Expression on Social Media. In: Journal of Computer-Mediated Communication 29 (1), zmad041. ↗ https://doi.org/10.1093/jcmc/zmad041.

Weldon, E./Gargano, G. M. (1988). Cognitive Loafing: The Effects of Accountability and Shared Responsibility on Cognitive Effort. In: Personality and Social Psychology Bulletin 14 (1), 159–171. ↗ https://doi.org/10.1177/0146167288141016.

Williams, M. L. et al. (2019). Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. In: The British Journal of Criminology 60 (1), 93–117. ↗ https://doi.org/10.1093/bjc/azz049.

Wischnewski, M. et al. (2021). Disagree? You Must Be a Bot! How Beliefs Shape Twitter Profile Perceptions. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–11, New York, NY, USA. Association for Computing Machinery. ↗ https://doi.org/10.1145/3411764.3445109.

Wojatzki, M. et al. (2018). Do Women Perceive Hate Differently: Examining the Relationship between Hate Speech, Gender, and Agreement. In: 14th Conference on Natural Language Processing, KONVENS'18.

Yakura, H. et al. (2025). Empirical Evidence of Large Language Model's Influence on Human Spoken Communication. In: arXiv. ↗ https://doi.org/10.48550/arXiv.2409.01754.

Yarkoni, T./Westfall, J. (2017). Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning. In: Perspectives on Psychological Science 12 (6), 1100–1122. ↗ https://doi.org/10.1177/1745691617693393.

You, S./Yang, C. L./Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation?. In: Journal of Management Information Systems 39 (2), 336–365. ↗ https://doi.org/10.1080/07421222.2022.2063553.

Zajko, M. (2021). Conservative AI and Social Inequality: Conceptualizing Alternatives to Bias through Social Theory. In: AI & Society 36 (3), 1047–1056. ↗ https://doi.org/10.1007/s00146-021-01153-9.

Zhang, M. et al. (2024). Don't Go to Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 12073–12086, Bangkok, Thailand. Association for Computational Linguistics. ↗ https://doi.org/10.18653/v1/2024.acl-long.652.

Ziegele, M./Jost, P. B. (2020). Not Funny? The Effects of Factual versus Sarcastic Journalistic Responses to Uncivil User Comments. In: Communication Research 47 (6), 891–920. ↗ https://doi.org/10.1177/0093650216671854.

Ziegele, M./Koehler, C./Weber, M. (2018). Socially Destructive? Effects of Negative and Hateful User Comments on Readers' Donation Behavior toward Refugees and Homeless Persons. In: Journal of Broadcasting & Electronic Media 62 (4), 636–653. ↗ https://doi.org/10.1080/08838151.2018.1532430.

Ziegele, M./Naab, T. K./Jost, P. (2020). Lonely Together? Identifying the Determinants of Collective Corrective Action against Uncivil Comments. In: New Media & Society 22 (5), 731–751. ↗ https://doi.org/10.1177/1461444819870130.

List of Figures and Tables

en.bidt.digital

bidt – Bavarian Research Institute for Digital Transformation (bidt)
An institute of the Bavarian Academy of Sciences and Humanities (BAdW)
Gabelsbergerstrasse 4 | 80333 Munich