Governing AI Counter-Speech:

Effectiveness, Power, and Legitimacy in Digital Public Spheres

solidarity with OTHERS

April 2026

BELGIUM

This report has been published by Solidarity With OTHERS. The views expressed herein are those of the authors and do not necessarily reflect the views of any donors or affiliated institutions.

Governing AI Counter-Speech: Effectiveness, Power, and Legitimacy in Digital Public Spheres

Author(s): Miranda MCHEDLISHVILI

Weiveldlaan 41 Unit D, 1930 Zaventem BELGIUM
www.solidaritywithothers.com
info@solidaritywithothers.com

Solidarity With OTHERS, an international human rights organisation registered in accordance with the Belgian Code of Companies and Associations (**OTHERS AISBL**, BCE number: 0715.742.907).

Table of Contents

Executive Summary

Online hate speech damages democratic discourse and harms communities. Platforms and civil society increasingly use artificial intelligence to generate automated responses challenging hate speech without censorship. This report examines when AI counter-speech works, where it fails, and who controls these powerful systems.

The report's central argument is that the challenge is not only whether AI counter-speech is effective, but who governs it and whether that governance is legitimate. Power imbalances and accountability deficits, not just technical limitations, determine whether AI strengthens or weakens democratic discourse.

When AI Counter-Speech Works

AI counter-speech succeeds only under specific conditions. Message design matters: AI responses that contradict users' core values backfire, making extremism worse. Simple warnings work better than complex personalized messages. Targeting matters: interventions work when aimed at undecided bystanders, not hardened extremists. Platform cooperation matters: effective counter-speech needs access to ranking algorithms that control visibility.

Opportunities and Risks

AI protects human moderators from psychological trauma, provides backup for isolated counter-speakers, and maintains consistent quality at scale. But AI flattens cultural diversity into generic responses, tends to agree with users rather than challenge them, misses culturally specific hate, and generates false information that damages trust. Human oversight remains essential.

Who Controls AI Counter-Speech

Approximately 4-6 companies control AI foundation model development, the layer where large language models powering counter-speech systems are built – with OpenAI and Microsoft holding 69% of the market. Civil society groups and smaller platforms depend on these companies' systems, inheriting their biases. Users have very limited ability to understand, contest, or shape these systems.

The technology operates as a "black box," citizens cannot see how decisions are made. When governments force platforms to moderate content under threat of heavy fines, platforms remove legal speech to avoid penalties. Worse, authoritarian governments in Turkey and Saudi Arabia use these same tools to suppress dissent and harass critics.

What Should Be Done

Platforms should: employ moderators directly with proper support, work with local experts to understand cultural context, make systems more transparent, and give researchers access to data.

Governments should: require human rights reviews of AI systems, create independent oversight bodies, match regulations to platform size, and train judges on AI systems.

Civil society should: create culturally relevant campaigns with local voices, coordinate collective responses to hate, monitor platforms independently, and build open-source alternatives.

International organizations should: establish global oversight councils and connect grassroots groups across borders.

The Choice Ahead

AI counter-speech shapes who can speak and be heard online. Current power imbalances and lack of transparency are not inevitable, they result from choices

favoring corporate interests over public accountability. Whether AI strengthens or weakens democracy depends on governance decisions made now. Without coordinated action across all stakeholders, the risks outweigh the benefits.

1. Introduction

Online hate speech silences targeted voices, normalizes extremism and undermines democratic discourse. The scale is staggering. Facebook moderates approximately three million posts daily, leaving 15,000 moderators under 150 seconds per decision.[1] The Swedish network #iamhere recruited 74,000 volunteers but retains only 2,000-3,000 active weekly due to emotional exhaustion.[2] Human responses face structural limits: toxic content exceeds processing capacity, and prolonged exposure causes severe psychological harm.

Artificial intelligence presents an appealing solution. Large language models generate counter-speech – responses challenging hate with argument and evidence – at scales no human workforce can match. Platforms increasingly deploy these systems, civil society experiments with AI campaigns, and policymakers consider mandating automated interventions. Proponents argue AI shields moderators from trauma, provides immediate responses, and deploys validated strategies consistently. Critics warn that systems lack cultural fluency, cannot detect coded hatred, and risk homogenising expression.

Both perspectives contain truth, but neither addresses the more fundamental question: whether deployment will be governed to protect democratic values and preserve the integrity of the digital public sphere. This report examines when AI-driven counter-speech campaigns are effective in mitigating hate speech, and what their opportunities and limitations reveal about power, legitimacy, and governance in digital public spheres. It synthesises experimental studies on AI effectiveness, civil

[1] Koetsier, J., "Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day," Forbes, 9 June 2020, https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/

[2] Buerger, C., "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse," Social Media + Society, 2021, https://journals.sagepub.com/doi/full/10.1177/20563051211063843

society campaign case studies, platform transparency reports, and theoretical work on algorithmic power and legitimacy to provide policymakers with evidence-based guidance on governing these systems.

Scope and Structure

1

Introduction

This report focuses on AI-driven counter-speech, automated responses challenging hate, rather than content removal systems. Counter-speech raises distinct governance questions because it operates through persuasion, not censorship, adding expression rather than subtracting it.

Section 2 establishes the conceptual framework, defining key terms and the governance lens applied throughout. **Section 3** examines effectiveness conditions around message design, targeting, and platform architecture. **Section 4** analyses opportunities AI offers and structural limitations it cannot overcome. **Section 5** examines power concentration, legitimacy deficits, governance privatisation, and weaponisation risks. **Section 6** proposes recommendations for platforms, states, civil society, and international organisations.

The report does not claim definitive answers to contested normative questions about acceptable speech boundaries in democratic societies. It does provide evidence-based guidance on governing AI counter-speech to respect rights, distribute power equitably, and preserve conditions for democratic deliberation.

2. Conceptual Framework

Effective policy responses to online hate speech require conceptual clarity. This section defines the core concepts used throughout the report and establishes the governance lens through which AI-driven counter-speech is analysed.

2.1. Online Hate Speech and Its Variants

While there is no single internationally accepted legal definition, this report adopts a broad working definition of online hate speech as communication that attacks, diminishes, or incites violence or discrimination against individuals or groups on the basis of protected characteristics, including religion, race, ethnicity, gender, or sexual orientation.[3]

This report also draws on the concept of "extreme speech" to better capture how hatred manifests in digital environments. Extreme speech covers two main forms: derogatory extreme speech, which involves uncivil attacks on social groups, and exclusionary extreme speech, which calls – openly or implicitly – for the removal or marginalisation of vulnerable communities.[4] A particularly challenging subset is "deep extreme speech": hateful content embedded in local slang, historical references, coded humour, and in-group cultural knowledge that automated systems trained on generic datasets, are poorly equipped to recognise.[5]

Policy relevance: Distinguishing between surface-level and deep extreme speech is critical for policymakers designing AI deployment mandates. Regulations that treat

[3] Brodowicz, M., "The use of AI in detecting and combating online hate speech," 11 February 2025, https://aithor.com/essay-examples/the-use-of-ai-in-detecting-and-combating-online-hate-speech#1-introduction

[4] Udupa, S. et al., Artificial intelligence, extreme speech, and the challenges of online content moderation, AI4Dignity, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

[5] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

hate speech as a uniform category will systematically underperform in contexts where coded, culturally specific hatred is most prevalent.

2.2. Counter-Speech: Definition and Goals

Counter-speech – used interchangeably in this report with the term "Online Discourse Engagement" (ODE) – refers to any direct, reactive response to hateful or harmful content that seeks to undermine it without recourse to censorship or content removal.[6] Unlike traditional content moderation, which operates by suppressing speech, counter-speech is an additive intervention: it responds to harmful expression with more expression, relying on argument, evidence, empathy, and social norms to diminish the reach and influence of hate.

The goals of counter-speech are varied, and the target audience is not always the perpetrator. Counter-speech may aim to change the mind of the individual who posted hateful content, to signal solidarity with those targeted, or to inoculate bystanders against hate's normalising effects. Among these three audiences, the "movable middle" of observers who have not yet consolidated extremist views – is often the most strategically important target.[7]

2.3. AI-Driven Counter-Speech: Variants

AI-driven counter-speech refers to the use of artificial intelligence – in particular Large Language Models (LLMs) and conversational chatbots – to automatically generate or assist in the writing of responses to hate speech.[8] As a category, it encompasses a

[6] Schirch, L. et al., A taxonomy of response strategies to toxic online content: Evaluating the evidence, University of Notre Dame, 2025, https://kroc.nd.edu/assets/566014/schirch_radivojevic_buerger_2025_taxonomy_of_response_strategies.pdf

[7] Yi-Ling Chung et al., "Understanding Counterspeech for Online Harm Mitigation," Northern European Journal of Language Technology, 2024, https://doi.org/10.3384/nejlt.2000-1533.2024.5099

[8] Damo, G. et al., "Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study," International Conference on Web Intelligence and Intelligent Agent Technology, 2025, https://hal.science/hal-04969313

spectrum of models that differ meaningfully in their degree of automation and human involvement. This distinction matters for the governance analysis that follows: AI operating as a drafting or support tool raises different legitimacy concerns than AI operating as an autonomous intervention system – and both differ from AI deployed as a structural engagement mechanism, such as automated donation triggers or algorithmic comment re-ranking. The governance and legitimacy critiques developed in Section 5 apply most forcefully to fully or semi-automated systems operating without meaningful human review of individual decisions.

AI-driven counter-speech encompasses models ranging from minimal to full automation. At the minimal end, "human-in-the-loop" assistive tools[9] use AI for detection and drafting but require human approval before posting. "Semi-automated systems" generate and post responses autonomously within predefined parameters, with human oversight at the design and auditing stages. At the full automation end, "fully automated bots" detect and respond independently without human approval of individual responses.[10]

2.4. Defining Effectiveness: A Multidimensional Framework

Effectiveness, as used in this report, is assessed along three interrelated dimensions. The first is social impact, measured by observable shifts in platform behaviour such as the deletion of hateful posts, changes in engagement metrics, or the elevation of civil content within ranking algorithms. The second is behavioural change, referring to documented reductions in the future toxic behaviour of targeted individuals. The third

[9] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

[10] Adam, G.P., Machine learning tools to (semi-)automate evidence synthesis: A rapid review and evidence map, 2025 review update, Agency for Healthcare Research and Quality, 2025, https://effectivehealthcare.ahrq.gov/products/machine-learning-evidence-synthesis/rapid-review

is attitudinal change, which concerns shifts in the underlying beliefs and prejudices that motivate hate speech.[11] These three dimensions do not always move in the same direction, and this report treats them as distinct for evaluation purposes.

Policy relevance: Policymakers who evaluate AI counter-speech solely on social impact metrics, such as post removal rates, will miss whether systems are producing genuine behavioural or attitudinal change. Accountability frameworks must require reporting across all three dimensions.

2.5 The Governance Lens: Platform Power, State Authority, and User Agency

The deployment of AI-driven counter-speech raises fundamental questions about who governs online speech, by what authority, and subject to what accountability. This report situates these questions within a governance framework that distinguishes between three modes of regulating online content.

The first is platform governance, in which private companies regulate speech through internal mechanisms: terms of service, community guidelines, algorithmic ranking systems, and internal complaint-handling procedures.[12] Platforms operating under this model act as de facto regulators of public discourse, setting and enforcing speech norms largely outside the reach of public law.

The second is co-governance, a multi-stakeholder model in which civil society organisations, independent experts, and affected communities participate alongside platforms in the design and oversight of content governance systems. Examples

[11] Yi-Ling Chung et al., "Understanding Counterspeech for Online Harm Mitigation," Northern European Journal of Language Technology, 2024, https://doi.org/10.3384/nejlt.2000-1533.2024.5099

[12] Schirch, L. et al., A taxonomy of response strategies to toxic online content: Evaluating the evidence, University of Notre Dame, 2025, https://kroc.nd.edu/assets/566014/schirch_radivojevic_buerger_2025_taxonomy_of_response_strategies.pdf

include "Social Media Councils",[13] "Trusted Flagger" networks,[14] and collaborative coding initiatives such as AI4Dignity, in which independent fact-checkers and ethnographers co-develop the training data used to define hate speech labels.

The third is state regulation, in which governments impose legal obligations on platforms through instruments such as Germany's Network Enforcement Act (NetzDG) or the European Union's Digital Services Act (DSA), mandating transparency, risk mitigation, and the expeditious removal of illegal content under the threat of financial penalties.[15]

These three modes of governance are not mutually exclusive; in practice, the regulation of online hate speech involves all three simultaneously. Three concepts cut across all of them and are central to this report. Platform power denotes the asymmetric "one-way control" – the ability of one actor to shape the prospects, options, and attitudes of others without those others being able to exercise equivalent influence in return.[16] This control is concentrated vertically across the AI technology stack, where a small number of incumbent companies control the data and resources to develop foundational AI models.

Legitimacy, in this context, refers to the requirement that governing power satisfies a publicity condition: citizens must be able to understand, in reasonably accessible

[13] Fertmann, M. & Kettemann, M.C., "Democracy in Flux: Order, Dynamics and Voices in Digital Public Spheres," Proceedings of the Weizenbaum Conference, 2021, https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Proceedings/Proceedings_Weizenbaum_Conference_2021.pdf

[14] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

[15] Fertmann, M. & Kettemann, M.C., "Democracy in Flux: Order, Dynamics and Voices in Digital Public Spheres," Proceedings of the Weizenbaum Conference, 2021, https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Proceedings/Proceedings_Weizenbaum_Conference_2021.pdf

[16] Lazar, S., "Legitimacy, Authority, and Democratic Duties of Explanation," Oxford Studies in Political Philosophy, 2024, https://doi.org/10.1093/oso/9780198916055.003.0002

terms, how and by whom decisions affecting their speech are being made. This requirement is directly challenged by the opacity of "black box" AI systems – automated decision-making processes whose internal logic is shielded by technical complexity and trade secrecy, making independent auditing difficult or impossible for the public, civil society, or regulators.[17]

Both concepts are situated within the broader context of the digital public sphere – the space constituted when online audiences gather to discuss issues of collective concern, and which depends on pluralism, openness, and shared understanding of facts to function as a site of democratic deliberation.[18]

Policy relevance: Legitimacy is not a secondary concern, it is a precondition for effective governance. Systems that citizens cannot understand or contest will face resistance, circumvention, and erosion of trust, regardless of their technical effectiveness. Policymakers must treat transparency as a design requirement, not an afterthought.

[17] Lazar, S., "Legitimacy, Authority, and Democratic Duties of Explanation," Oxford Studies in Political Philosophy, 2024, https://doi.org/10.1093/oso/9780198916055.003.0002
[18] Lazar, S., "Legitimacy, Authority, and Democratic Duties of Explanation," Oxford Studies in Political Philosophy, 2024, https://doi.org/10.1093/oso/9780198916055.003.0002

3. Conditions of Effectiveness

3

Conditions of Effectiveness

The evidence reviewed in this report makes clear that AI-driven counter-speech is not a uniformly effective intervention. Its impact varies significantly depending on how it is designed, whom it targets, and the platform environment in which it operates. This section organises the available evidence into three categories of conditions, message design, targeting, and platform architecture, and draws out the implications of each for policy and deployment.

3.1. Message Design Conditions

The Value Alignment Effect

One of the most consequential findings to emerge from recent experimental research concerns the relationship between the AI's moral framing and the values of the user it addresses. A crowdsourced study found that while LLM-based chatbots were capable of persuading users whose values aligned with the AI's framing, value misalignment produced the opposite effect.[19] When participants discussing US defence spending were assigned a chatbot whose moral frame, pacifism, contradicted their own values, they did not moderate their views; they entrenched them. This backfire effect is what this report terms "blind" persuasion: the deployment of AI responses without first detecting the ideological stance of the target. Blind persuasion does not merely fail to reduce extremism – it may actively deepen it, because the user experiences the misaligned intervention as an attack on their identity rather than an invitation to reconsider. Where value alignment cannot be reliably assessed, automated systems risk functioning as polarisation engines than counter-speech tools.

[19] Wise, A. et al., "A Crowdsourced Study of ChatBot Influence in Value-Driven Decision Making Scenarios," 19 November 2025, https://arxiv.org/html/2511.15857v1

3

Conditions of Effectiveness

Effective AI counter-speech therefore requires, at minimum, the capacity to identify the value orientation of the user before selecting a response strategy. This finding strengthens the case for "human-in-the-loop" systems where human reviewers can assess value alignment before approving responses. In the absence of this capability, automated campaigns risk increasing polarisation rather than mitigating it.

Policy implication: Where value alignment cannot be reliably assessed, fully automated persuasion systems are risky and should not be deployed in high-stakes contexts. Policymakers should require platforms to demonstrate value-alignment capability before authorising autonomous counter-speech systems.

Warning Simplicity and Impact

A second and equally important finding challenges a widespread assumption in the field: that more sophisticated, contextually tailored AI responses will outperform simpler interventions. A large-scale field experiment conducted on Twitter/X, involving 2,664 participants, directly tested this assumption.[20] The study compared AI-generated, personalised counter-speech responses against generic, pre-written messages warning users of the consequences of posting hate speech. The results were counterintuitive: generic consequence warnings significantly reduced toxicity, while AI-generated contextualised responses were largely ineffective and in some cases increased hate speech. This finding suggests that normative boundary-setting – clearly communicating that hateful behaviour violates community standards and carries consequences – can be more powerful than technologically sophisticated persuasion. For platform designers, this implies that investment in AI complexity is not always the most effective use of resources, and that simple, clearly communicated norms may serve as a more reliable first line of response.

[20] Bär, D., Maarouf, A. & Feuerriegel, S., "Generative AI may backfire for counterspeech," 2024, https://arxiv.org/abs/2404.12467

Policy implication: Investment in AI complexity is not always the most effective use of resources. Platforms should invest in clear, consistently enforced norm-setting mechanisms – not only in sophisticated AI persuasion tools. Simple, well-communicated community standards may serve as a more reliable first line of response.

3

Conditions of Effectiveness

The Role of Rhetorical Quality

Where AI is deployed to generate substantive counter-speech, the quality and structure of the response matters considerably. A comparative study, evaluating expert-written counter-speech (CONAN) from NGOs against user-generated responses on Twitter, found that expert responses – characterised by fact-based evidence, structured argumentation, and qualities such as fairness and clarity – were consistently rated as more effective than user-generated responses, which tended to rely on emotional language and sarcasm. Among the specific rhetorical strategies examined, "broadening universals" – reframing a stereotyped trait as a universal human characteristic rather than one specific to a target group – emerged as the most robust and effective approach.[21] Fact-based rebuttals delivered without an antagonistic tone also performed well, particularly when addressing the "movable middle." Humour and sarcasm, by contrast, remain high-risk strategies: while they may occasionally diffuse tension, they frequently trivialise harm or alienate the audience. These findings have direct implications for how AI systems should be trained: the rhetorical qualities of expert-written counter-speech, structured reasoning, factual grounding, and measured tone, represent a clear benchmark for AI to emulate.

[21] Schirch, L. et al., A taxonomy of response strategies to toxic online content: Evaluating the evidence, University of Notre Dame, 2025, https://kroc.nd.edu/assets/566014/schirch_radivojevic_buerger_2025_taxonomy_of_response_strategies.pdf

3.2. Targeting Conditions

Targeting the "Movable Middle"

3

Conditions of Effectiveness

The evidence consistently indicates that counter-speech interventions are most effective when directed not at the producers of hate speech, but at the broader audience of observers who have not yet consolidated extremist views. The #iamhere network (#jagärhär), a Swedish civil society initiative that organises volunteers to respond to hate speech in public comment sections, illustrates this principle clearly. Rather than attempting to change the minds of committed hate speech producers, *#iamhere* explicitly targets "silent readers" – bystanders who are exposed to hateful content but whose views remain open to influence. By populating comment sections with civil, factual responses, the network breaks the illusion of consensus that hate speech often creates, demonstrating to observers that the hateful voice does not represent the majority opinion.[22]

Policy implication: AI should focus on visibility and norm protection, not persuasion of extremists. Fully automated systems engaging hardened extremists waste resources and risk backfire. Systems targeting bystanders, ensuring civil dissent is visible, produce more durable results.

3.3. Platform Architecture Conditions

Algorithmic Infrastructure

A third condition of effectiveness concerns the relationship between counter-speech campaigns and the algorithmic architecture of the platforms on which they operate. Text generation alone is insufficient; the visibility of counter-speech relative to hate

[22] Buerger, C., "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse," Social Media + Society, 2021, https://journals.sagepub.com/doi/full/10.1177/20563051211063843

speech is shaped by platform ranking algorithms, and effective campaigns must account for this dynamic.

The *#iamhere* case study again provides instructive evidence. Members of the network do not merely post responses to hate speech; they coordinate "likes" and replies in order to exploit Facebook's comment ranking algorithm, pushing toxic comments to the bottom of comment threads and elevating constructive responses to the top. This deliberate manipulation of platform architecture amplifies the reach and impact of counter-speech beyond what the content of the responses alone would achieve.[23] The implications for AI-driven campaigns are significant. Automated systems that generate high-quality counter-speech but operate without any capacity to interact with ranking and visibility mechanisms are operating at a structural disadvantage within the platform environment. Effective AI counter-speech, in this context, requires not just text generation capability but platform cooperation – specifically, access to algorithmic tools that can elevate civil content and reduce the prominence of hate. Whether platforms provide this cooperation is a governance question explored in Section 5.

Policy implication: Effective AI counter-speech requires platform cooperation, not just text generation capability. Policymakers should require platforms to provide counter-speech systems with access to ranking mechanisms – without this, even well-designed interventions are structurally disadvantaged.

[23] Buerger, C., "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse," Social Media + Society, 2021, https://journals.sagepub.com/doi/full/10.1177/20563051211063843

4. Opportunities & Limitations

AI-driven counter-speech offers distinct advantages but carries structural limitations requiring human oversight. This section examines what AI can do that humans cannot, and where human judgment remains indispensable.

4.1. Opportunities: What AI Can Do That Humans Cannot

Scale as a Shield Against Burnout

One of the most significant opportunities AI presents is the capacity to assume the psychological burden of exposure to toxic content. The scale is staggering: Facebook alone moderates approximately three million posts daily. With 15,000 content moderators, this translates to roughly 200 posts per shift, or 25 per hour – leaving under 150 seconds per decision (Koetsier, 2020). Prolonged exposure to violent, dehumanising material causes severe psychological damage: in 2020, Facebook agreed to a $52 million settlement to compensate US-based moderators who developed PTSD.[24]

Civil society faces the same crisis. The Swedish counter-speech group *#iamhere* recruited 74,000 members, but emotional exhaustion means only 2,000-3,000 remain active weekly.[25] This represents a structural limit on human-led interventions: volunteers burn out, withdraw, or experience long-term harm that limits their capacity to continue.

[24] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf
[25] Buerger, C., "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse," Social Media + Society, 2021, https://journals.sagepub.com/doi/full/10.1177/20563051211063843

4

Opportunities & Limitations

AI, by contrast, does not experience psychological distress. Automated systems can process high volumes of hateful content without suffering the emotional toll that incapacitates human moderators. This suggests a specific role for AI as a first line of defense – not replacing human judgment, but shielding both corporate workers and civil society actors from direct trauma. By filtering low-complexity hate speech, AI preserves human capacity for strategic work requiring cultural fluency, contextual judgment, and moral reasoning. The opportunity is not full automation, but a reallocation of labor that allows humans to engage where most needed while reducing the health costs of frontline exposure.

Manufacturing Safety in Numbers

A second opportunity concerns collective action dynamics. As demonstrated by the #iamhere model discussed in Section 3, "isolated individuals are often deterred from responding" to hate speech by fear of harassment.[26] AI can artificially replicate the dynamic that makes human counter-speech networks effective. Automated agents can provide immediate, visible support to human counter-speakers, populating discussions with multiple civil voices even when few human activists are present. This ensures bystanders see hate speech as contested rather than normative, reducing the likelihood that silence is misinterpreted as consent.

Robustness in Deploying "Broadening Universals"

A third opportunity lies in AI's demonstrated capacity to reliably generate specific rhetorical strategies that have been empirically validated as effective. Comparative studies, as noted in Section 3, have found that "broadening universals" are among the most robust and effective counter-speech strategies. This is significant because human-generated counter-speech, particularly when produced under the stress of real-time engagement, is highly variable in quality. User-generated responses often

[26] Langton, R., "Blocking as Counter-Speech," in D. Fogal et al. (eds.), New Work on Speech Acts, Oxford University Press, 2018, https://doi.org/10.1093/oso/9780198738831.003.0007

default to emotional appeals or sarcasm, both of which carry higher risks of backfire. AI, when properly trained, can maintain consistency in tone, structure, and rhetorical approach across thousands of interactions. This does not eliminate the need for human oversight, but it does offer a mechanism for ensuring that a baseline standard of response quality is met even when expert human writers are not available in sufficient numbers. The opportunity here is to scale proven strategies, not to innovate new ones.

4.2. Strategic Reframing: Changing the Logic of Engagement

Beyond generating text, AI can change the incentive structures and environmental conditions that enable hate speech. Two case studies illustrate this potential.

Rebalancing Cost-Reward Structures

The Hass Hilft (Hate Helps) campaign represents a creative use of automation to flip the incentive logic of hate speech.[27] Rather than engaging perpetrators in argument, the campaign implemented an involuntary donation mechanism: every hate comment posted triggered a one-euro donation to a cause the perpetrator opposed, such as refugee aid or anti-extremism organisations. The hate speech producer, in effect, was made to financially support the very groups they were attacking.

The strategic value of this intervention lies not in persuasion but in consequences. By automating the donation process and publicly announcing it, the campaign reframed hate speech from a cost-free act of aggression into one with tangible negative consequences for the perpetrator. This structural intervention requires no sophisticated natural language processing and makes no attempt to change minds; it simply alters the reward calculus. This demonstrates that AI interventions need not

[27] Global Internet Forum to Counter Terrorism (GIFCT), Content-Sharing Algorithms, Processes, and Positive Interventions Working Group, 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-CSAPIWG-2021.pdf

rely solely on persuasive arguments but can create accountability mechanisms impossible to administer manually at scale.

Subversive Mirroring and Automated Pedagogy: Peng! Collective

A second example of strategic reframing comes from the Peng! Collective's "Zero Trollerance"_campaign, which used bots to identify abusive users on Twitter and automatically send them humorous "self-help" video tutorials on how to be a better person.[28] The intervention was deliberately non-confrontational, reframing the troll not as a powerful aggressor but as someone in psychological need of assistance.

The mechanism here is subversive mirroring: the campaign used automation to deliver satire and "kindness" at scale, disarming aggression without escalating conflict. While this approach carries risks – humour can trivialise harm, and not all recipients responded positively – it illustrates a broader principle: AI can be used to deploy creative counter-tactics that sidestep the traditional argumentative model of counter-speech entirely. These interventions succeed not by out-arguing hate speech but by changing the emotional and social context in which it occurs.

4.3. Limitations and Risks: Why Human Oversight is Non-Negotiable

The "Value-Misaligned" Backfire Effect

As established in Section 3.1, AI systems trigger backfire effects when their moral framing contradicts users' core values – a problem this report calls "blind persuasion." The implication for system design is direct: fully automated deployments optimised for linguistic fluency, without any capacity to assess value alignment, should not be used in high-stakes persuasion contexts. "Human-in-the-loop" systems remain

[28] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

essential wherever empathy, cultural competence, and strategic judgment about when to engage or withdraw are required.

Homogenisation of Expression

A second limitation concerns diversity of expression. Generative AI produces content based on statistical probability, creating outputs that conform to patterns learned from training data. This process smooths out unique phrasing, cultural inflection, and minority perspectives that fall outside statistical norms. The result is "standardisation of expression" – a flattening of linguistic and cultural diversity into repetitive, statistically probable output.[29]

If counter-speech becomes dominated by AI-generated content, discourse loses the cognitive novelty and cultural specificity that make it vibrant and meaningful. Bystanders encounter automated imitations of dissent optimised for statistical probability rather than genuinely diverse perspectives. This homogenisation particularly harms contexts where minority voices, regional dialects, and culturally specific arguments are already marginalised. Without attention to this risk, AI counter-speech may inadvertently reinforce majority cultural norms under the guise of neutrality.

The Sycophancy Trap

A related limitation arises from the way AI systems are optimised for user satisfaction. To maximise engagement and avoid conflict, many generative models are designed – whether intentionally or as an emergent property of their training – to agree with the user's stated beliefs rather than challenge them. This tendency, often called

[29] Council of Europe (CDMSI), Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression, 2025, https://rm.coe.int/guidance-note-on-the-implications-of-generative-artificial-intelligence/1680b2c038

"sycophancy," means that AI systems may mirror a user's biases and prejudices rather than confronting them.[30]

The consequence is the automation of echo chambers. If an AI counter-speech system encounters a user with extremist views and adapts its messaging to avoid alienating them, it may end up validating those views rather than undermining them. This is particularly problematic in hybrid systems where AI assists human moderators: if the AI drafts responses that are pre-calibrated to avoid user pushback, human reviewers may approve messages that are conciliatory to the point of ineffectiveness.

"Deep Extreme Speech" Blindness

AI systems trained on large, decontextualised datasets struggle to detect or respond effectively to "deep extreme speech" which is often invisible to classifiers trained on generic datasets dominated by English-language examples from Western contexts. A slur that carries profound historical weight in one country may not register as hateful to a model trained primarily on data from another. Humour, satire, and coded references that are legible to in-group members as expressions of hate may be classified as neutral by automated systems.[31]

This limitation is not easily overcome through technical improvements alone. Cultural context cannot be reliably inferred from text, and automated systems lack the localised expertise necessary to navigate culturally specific hatred. AI counter-speech systems operating without access to local cultural knowledge will systematically fail to identify and respond to the forms of hate speech most deeply embedded in specific communities.

[30] Council of Europe (CDMSI), Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression, 2025, https://rm.coe.int/guidance-note-on-the-implications-of-generative-artificial-intelligence/1680b2c038
[31] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

4

Opportunities & Limitations

Hallucination and the Erosion of Trust

A final limitation concerns the tendency of generative AI systems to produce plausible but false information – a phenomenon known as "hallucination." When AI systems generate counter-speech that includes fabricated statistics, misattributed quotes, or invented sources, they do not merely fail to persuade; they actively undermine trust in the information environment. Users who discover that an AI-generated response contains false information may generalise that mistrust to all counter-speech, including human-generated content, making the ecosystem more hostile to correction rather than less.[32]

AI systems also tend to aggregate information from multiple sources while stripping away the original context and attribution. Even when the information is accurate, users have no way to verify where it came from or assess its credibility. This creates a compounding problem: hallucinated content erodes trust directly, while decontextualised content erodes it more quietly. AI systems that operate without robust mechanisms for source verification and factual grounding therefore carry an inherent credibility deficit – and counter-speech that cannot be trusted cannot persuade.

The central issue is no longer *whether* AI should be used in counter-speech efforts, but *how it should be governed and under what constraints.* The capabilities are real, but so are the risks of unchecked automation. These are not technical problems awaiting technical solutions – they are governance problems, and they demand governance answers.

Even where AI counter-speech is effective, a deeper question remains: who controls these systems, and under what authority? Effectiveness and legitimacy are distinct

[32] Council of Europe (CDMSI), Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression, 2025, https://rm.coe.int/guidance-note-on-the-implications-of-generative-artificial-intelligence/1680b2c038

problems, a system can reduce toxic posts while concentrating power in unaccountable hands or being co-opted by repressive actors. It is precisely because AI can be effective that governance becomes urgent. The following section examines how power is concentrated in AI systems, where democratic legitimacy is absent, and what that means for the future of counter-speech governance.

4

Opportunities & Limitations

5. Power, Legitimacy and Governance

The deployment of AI-driven counter-speech at scale is not merely a question of technical effectiveness, it is a governance choice. AI counter-speech does not simply respond to public discourse; it shapes who can speak, who is heard, and what norms are enforced. This insight, that AI counter-speech is a form of governance, not merely a tool within it – is one of the central arguments of this report, and it runs through every dimension of the analysis that follows.

Even where the effectiveness conditions outlined in Sections 3 and 4 are met, this tells us nothing about whether the power exercised is legitimate, accountable, or compatible with democratic values.

This report's analysis is guided by three normative commitments: preserving the digital public sphere for democratic deliberation; protecting freedom of expression while mitigating harm; and preventing undue concentration of power over speech. This section examines four dimensions of the governance challenge through that lens: algorithmic power concentration, democratic legitimacy deficits, the privatisation of governance, and authoritarian weaponisation.

5.1. The Concentration and Nature of Algorithmic Power

Power as One-Way Control

AI systems are not neutral tools; they create and intensify asymmetric power relations. Power in this context is defined as "one-way control," platform operators solely

determine what content is visible, what responses are generated, and what values are encoded in moderation decisions. Users who encounter this counter-speech have no corresponding capacity to shape the system's logic, challenge its assumptions, or contest its outputs through the system itself. They are subject to governance they cannot reciprocally influence.

This power asymmetry includes informational asymmetry. System developers know when AI is framing conversations or nudging behavior, but users do not.[33] Users cannot adjust their evaluation standards accordingly. They may interpret automated responses as evidence of widespread agreement when none exists, or mistake platform policies for genuine public opinion. The result is invisible governance: users are subject to persuasive interventions they cannot recognise, evaluate, or resist.

Vertical Market Concentration

Power is concentrated structurally across the AI technology stack. The Foundation layer – where large language models that power counter-speech systems are developed – is dominated by approximately 4-6 companies. As of late 2023, OpenAI and Microsoft together controlled 69% of generative AI market spending.[34] This concentration creates structural dependencies: civil society organisations, smaller platforms, and national governments that lack the resources to develop independent AI systems must rely on models controlled by these incumbents.

This concentration is driven by prohibitive barriers to entry. The computational infrastructure required is itself highly concentrated: Nvidia controls 92% of the GPU market, while Amazon, Microsoft, and Google control 75% of cloud computing. Data is increasingly proprietary, with Google, Meta, and Microsoft possessing competitive

[33] Wise, A. et al., "A Crowdsourced Study of ChatBot Influence in Value-Driven Decision Making Scenarios," 19 November 2025, https://arxiv.org/html/2511.15857v1
[34] Vipra, J. & Korinek, A., Concentrating intelligence: Scaling and market structure in artificial intelligence, Institute for New Economic Thinking, 2 October 2024, https://www.ineteconomics.org/uploads/papers/WP_228-Korinek-and-Vipra.pdf

advantages through datasets from billions of users.[35] Capital reinforces this: Big Tech accounted for 67% of all generative AI startup funding in 2023.[36] The governance implications are direct. When counter-speech systems are built on foundational models controlled by the market leaders, the values encoded in those systems – what responses are appropriate, what tone is acceptable, what arguments are persuasive – are determined by private actors under no democratic mandate. How this power operates in practice, often invisibly, is illustrated by the case studies examined below.

5.2. The Crisis of Democratic Legitimacy and the "Publicity" Deficit

The Opacity Problem

For governing power to be morally permissible in a democratic system, it must meet standards of procedural legitimacy. A core component is the publicity requirement: citizens must understand, in reasonably accessible terms, how and by whom decisions affecting them are made. AI systems deployed for content moderation and counter-speech systematically fail this requirement. Their internal logic – training data, algorithmic weights, and decision pathways – is shielded from public scrutiny by corporate secrecy (intellectual property protections), technical complexity (processes opaque to non-specialists), and machine learning inscrutability (outputs not fully explicable even to creators). The result is "black box" governance where citizens, civil

[35] Vipra, J. & Korinek, A., Concentrating intelligence: Scaling and market structure in artificial intelligence, Institute for New Economic Thinking, 2 October 2024, https://www.ineteconomics.org/uploads/papers/WP_228-Korinek-and-Vipra.pdf

[36] Gambacorta, L. & Shreeti, V., Unpacking the AI supply chain: market structure and policy implications, European Money and Finance Forum (SUERF), 2025, https://www.suerf.org/publications/suerf-policy-notes-and-briefs/unpacking-the-ai-supply-chain-market-structure-and-policy-implications/

society organisations, and regulators cannot audit whether rules are applied fairly or cases treated alike.[37]

Case Study: COMPAS

The legitimacy deficits created by opaque algorithmic systems are not hypothetical. The COMPAS risk assessment algorithm, used in US criminal courts to predict recidivism and inform sentencing and bail decisions, provides a stark example. The algorithm's logic is kept secret as corporate intellectual property, making it impossible for defendants to challenge the specific reasoning behind decisions that deprive them of liberty. Courts have upheld the use of COMPAS despite this opacity, effectively allowing private corporate systems to exercise state power without the transparency that due process requires. Defendants are governed by a system they cannot inspect, producing outcomes they cannot meaningfully contest.[38]

5.3. The Privatisation of Governance and the "New Governors"

Enforcement and Its Distortions

The logic of enforcement creates predictable distortions. Platforms facing asymmetric consequences – heavy fines for under-enforcement, minimal penalties for over-enforcement – turn to automated moderation to minimise costs and legal risk. The result is systematic removal of legal content: automated systems lacking contextual

[37] Doshi-Velez, F. et al., Accountability of AI under the law: The role of explanation, Berkman Klein Center for Internet & Society, 3 November 2017, https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf
[38] McGregor, L. et al., "International human rights law as a framework for algorithmic accountability," International and Comparative Law Quarterly, 2019, https://doi.org/10.1017/S0020589319000046

judgment mistakenly classify satire, political criticism, and journalistic reporting as problematic, and platforms remove it rather than risk fines.[39]

This dynamic extends to counter-speech. When platforms deploy AI-generated counter-speech at scale, they curate discourse according to corporate preferences rather than democratic deliberation, transferring authority over public discourse from accountable institutions to corporations obligated primarily to shareholders.

Case Study: Germany's Network Enforcement Act (NetzDG)

The Network Enforcement Act (NetzDG), enacted in Germany in 2017, exemplifies the dynamics of delegated enforcement. The law requires social media platforms to delete "manifestly unlawful" content within 24 hours of receiving a complaint, or face fines of up to €50 million.[40] Proponents argue that the law brings necessary accountability to platforms that have been unwilling to self-regulate. Critics highlight that it effectively forces private, profit-driven corporations to act as the ultimate arbiters of free speech, shifting the burden of judicial review away from the state and into the hands of opaque corporate algorithms.

The regulatory structure creates what scholars term a "local control" scenario: full platform liability combined with local jurisdiction requirements compels proactive AI moderation to avoid fines.[41] Under compressed timelines and asymmetric penalties, platforms rationally optimise for higher recall at the expense of precision, systematically over-removing content to minimise regulatory risk.

[39] Loebbecke, C., Luong, A.C. & Obeng-Antwi, A., "AI for tackling hate speech," European Conference on Information Systems, 2021, https://aisel.aisnet.org/ecis2021_rp/20/
[40] Fichtner, L., "Moderating the Regulators/Regulating the Moderators: NetzDG and online content moderation in Germany," Proceedings of the Weizenbaum Conference, 2021, https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Proceedings/Proceedings_Weizenbaum_Conference_2021.pdf
[41] Loebbecke, C., Luong, A.C. & Obeng-Antwi, A., "AI for tackling hate speech," European Conference on Information Systems, 2021, https://aisel.aisnet.org/ecis2021_rp/20/

Early transparency reports from the first six months of implementation (January-June 2018) reveal the consequences of this delegation. Platforms processed vastly different complaint volumes – from Facebook's 1,704 to Twitter's 264,818 – driven by design choices in reporting accessibility. Removal rates varied from 10.8% (Twitter) to 27.1% (YouTube), with most decisions made within the mandated 24-hour window. While platforms rejected most complaints, the compressed decision-making timeline raises questions about whether contested cases receive adequate review. High-profile removals, such as Twitter's suspension of satirical magazine Titanic for mocking far-right politician Beatrix von Storch's inflammatory tweet about "Muslim hordes," illustrate the difficulty of distinguishing political satire from genuine hate speech under time pressure.[42]

While the German government found no evidence of systematic over-blocking and platforms rejected most complaints, this does not resolve the core legitimacy concern. Private corporations operating under threat of €50 million fines make rapid determinations about speech legality – decisions that would traditionally require judicial review – without procedural safeguards or public accountability. NetzDG thus illustrates a broader governance problem: delegated enforcement under threat of liability transfers authority over speech from public institutions subject to constitutional constraints to private entities operating according to corporate risk calculus.

5.4. State Repression and Weaponised Platform Governance

State-Backed Operations

[42] Echikson, W. & Knodt, O., Germany's NetzDG: A key test for combatting online hate, CEPS, 2018, https://www.ceps.eu/ceps-publications/germanys-netzdg-key-test-combatting-online-hate/

While Western democracies debate the legitimacy of privatised content governance, authoritarian and repressive regimes are exploiting these very same regulatory frameworks to crush dissent and manipulate the public sphere. Repressive states are increasingly emulating strict platform regulations – such as requirements for local "compliance officers," data localisation mandates, and expedited content removal procedures – not to protect citizens from harm, but to force companies to remove regime-critical speech and provide governments with access to user data for surveillance purposes.

These "copycat regulations" use the language of accountability and transparency to achieve the opposite: they make platforms accountable not to users or civil society, but to the state, and they create transparency mechanisms that allow governments to monitor opposition rather than audit platform behaviour. The result is a global regulatory race to the bottom, in which autocratic regimes leverage the legitimacy of democratic content governance frameworks to justify censorship and repression. AI-driven counter-speech, in this context, can be weaponised: rather than countering hate, it can be deployed to drown out dissent, amplify regime-preferred narratives, and create the illusion of popular support for authoritarian policies.[43]

Case Study: Turkey's Social Media Laws and Saudi Arabia's Troll Farms

Turkey's 2020 Social Media Regulation (Mehmet Bedii Kaya, 2024) illustrates the dynamics of weaponised governance. The law requires platforms operating in Turkey to appoint local representatives and comply with content removal requests from Turkish authorities. Platforms that refuse face advertising bans and internet slowdowns that make them effectively unusable. While framed as a measure to hold platforms accountable, the law has been used to compel the removal of regime-critical journalism, opposition political speech, and human rights documentation.

[43] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

Platforms, faced with the choice between compliance and exclusion from the Turkish market, have systematically complied, making them instruments of state censorship.[44]

Saudi Arabia's use of troll farms represents a more direct form of weaponisation. Investigations have revealed that the Saudi government recruited corporate consultancy firms to identify, target, and harass dissenting journalists and influencers on social media. These operations use the same tools – social media analytics, targeted messaging, coordinated amplification – that are deployed in counter-speech campaigns, but repurpose them to orchestrate hate rather than counter it. The distinction between counter-speech and state-backed harassment, in this context, collapses entirely.[45]

These cases demonstrate that relying on state regulation as the primary mechanism for governing AI-driven counter-speech is highly dangerous in non-democratic contexts. Algorithmic tools and regulatory mandates can easily be weaponised by authoritarian regimes to orchestrate digital hate rather than mitigate it. The governance challenge, therefore, is not merely to ensure that AI systems are effective at countering hate, but to ensure that they cannot be co-opted by states whose aim is to suppress rather than protect speech. Addressing these structural failures requires concrete action from every stakeholder, the recommendations that follow are designed to provide exactly that.

[44] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf
[45] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

6. Policy Recommendations

The question is not whether AI-driven counter-speech systems will be deployed – they already are – but whether their deployment will be governed in ways that protect democratic values and preserve the integrity of the digital public sphere. The power asymmetries, legitimacy deficits, and institutional failures identified in Section 5 are the result of governance choices. Different choices remain possible.

No single actor can govern AI counter-speech alone. The technical, legal, cultural, and institutional dimensions of the challenge exceed the capacity of any one institution. Governance must therefore be multi-level and multi-stakeholder – with platforms, states, civil society, and international organisations each playing distinct but coordinated roles. The recommendations that follow are organised accordingly.

To assist policymakers in sequencing action, recommendations are signalled by urgency: immediate priorities are those most feasible in the short term and foundational to all other reforms; medium-term priorities require capacity-building or legislative change; and longer-term structural goals address the systematic distribution of power over AI development and content governance.

For Social Media Platforms:

- **Reform Content Moderation Labor Practices [Medium-Term Priority]** – End the exploitative and opaque outsourcing of content moderation to third-party vendors. Platforms should recruit moderators as regular employees, provide rigorous training in local cultural and political contexts involving NGOs and scholars), and ensure adequate psychological support for those exposed to severe toxicity. As **Section 4.1** documented, the scale and trauma

of content moderation work necessitates structural labor reforms that prioritise worker wellbeing alongside operational efficiency.

- **Institutionalise "Human-in-the-Loop" Through Collaborative Coding [Medium-Term Priority]** – Recognise that AI cannot understand cultural nuances or "deep extreme speech" in isolation. Platforms should institutionalise regular collaborations with independent fact-checkers, ethnographers, and civil society to build culturally sensitive, inclusive training datasets, rather than treating these collaborations as episodic PR exercises. The AI4Dignity model, referenced in **Section 2.5**, demonstrates that community-in-the-loop approaches are essential for addressing the cultural blindness of generic AI systems.
- **Implement Algorithmic Risk Mitigation [Immediate Priority]** – Redesign algorithms to prioritise authoritative information and introduce "friction" into the user experience. For example, systems can generate automated prompts warning users before they retweet hateful content, asking them to reconsider or informing them that spreading the hate will trigger a donation to an anti-hate NGO. These structural interventions, illustrated by the "Hass Hilft" case study in **Section 4.2,** change the incentive logic of hate speech rather than relying solely on persuasive counter-arguments.
- **Expand Data Access for Independent Auditing [Immediate Priority]** – Fulfill the "publicity requirement" identified in Section 5.2 by granting vetted academic researchers and civil society access to platform data, including how algorithms curate and rank content and the prevalence of hate speech versus actions taken. This access enables independent audits of AI effectiveness and addresses the "black-box" opacity that undermines democratic accountability.

6

Policy Recommendations

For States:

- **Mandate Human Rights Impact Assessments [Immediate Priority]** – Require tech companies to conduct systematic, iterative, and documented Human Rights Impact Assessments (HRIAs) throughout the entire lifecycle of generative AI systems, from foundation to product layer. These assessments must specifically evaluate the impact of AI on freedom of expression, cognitive autonomy, and vulnerable populations.
- **Establish Independent Observatories [Long-Term Structural Goal]** – Create well-resourced, independent national and transnational observatories comprising technical and human rights experts. These bodies should systematically test, monitor, and publish findings on how AI and algorithmic governance impact public discourse and freedom of expression.
- **Apply the Principle of Proportionality [Immediate Priority]** – Impose strict obligations – such as independent auditing, transparency reports, and rapid removal of illegal content on platforms. Simultaneously, regulators must monitor smaller platforms and niche apps where extremists often migrate to avoid regulatory oversight. This tiered approach matches regulatory burdens to platform power while preventing under-regulated "gray zones".[46]
- **Capacity Building for the Judiciary [Medium-Term Priority]** – Provide specialised training for judges, prosecutors, and law enforcement on international human rights norms relating to hate speech, the unique trauma of gender-based cyber-harassment, and the mechanics of AI-assisted manipulation. As **Section 5.3**'s NetzDG case study revealed, compressed decision-making timelines and lack of contextual expertise contribute to poor moderation outcomes. Judicial capacity-building ensures that human review, when mandated, is informed and effective.

[46] Udupa, S., Digital technology and extreme speech: Approaches to counter online hate, United Nations Peacekeeping Technology Strategy, 2021, https://epub.ub.uni-muenchen.de/77473/1/Digital_technology_and_extreme_speech_Udupa.pdf

6

Policy Recommendations

For Civil Society, NGOs, and Academia:

- **Design Culturally Resonant Counter-Campaigns [Immediate Priority]** – Move beyond clinical, text-heavy fact-checking. NGOs should partner with local cultural influencers – comedians, musicians, meme creators – to design highly contextualised, multimodal interventions (memes, GIFs, humor) that resonate with digital youth cultures and local idioms. As **Section 3.1** demonstrated, rhetorical quality determines impact, and strategies that feel disconnected from community norms fail to persuade.
- **Mobilise Collective Bystander Support [Immediate Priority]** – Emulate models like #iamhere to provide immediate help for victims of online hate. Local groups should coordinate to flood toxic comment sections with civil counter-speech and coordinated "likes," altering the algorithmic visibility of hate speech and protecting isolated counterspeakers.
- **Act as "Trusted Flaggers" and Auditors [Medium-Term Priority]** – Grassroots groups should be equipped with technical resources to develop their own hate-monitoring dashboards, enabling them to act as certified "trusted flaggers" who expedite platform takedowns and independently verify whether platforms follow through on their moderation policies. This dual role – both assisting platforms and holding them accountable – positions civil society as essential checks on corporate power.
- **Create Independent Coding Marathons [Medium-Term Priority]** – Academic institutions and civil society should host independent coding marathons, like the AI4Dignity project, where marginalised communities and fact-checkers work directly with AI developers to train open-source algorithms, ensuring marginalised voices are represented in AI logic.

6

Policy Recommendations

For International Organisations:

- **Convene Social Media Councils [Long-Term Structural Goal]** – Institutionalise multi-stakeholder structures – involving states, platforms, and civil society – at the international level.[47] These councils should operate across three levels of authority: advisory (issuing guidance on content governance standards), dispute resolution (handling cross-border cases), and standard-setting (establishing baseline transparency and human rights requirements). Their value lies not in replacing regulation but in complementing it – generating implementation pressure on platforms through public accountability.
- **Fund and Connect Global Networks [Medium-Term Priority]** – Act as an incubator and connector for isolated grassroots initiatives. This will enable scalability and cross-border resilience against coordinated hate campaigns. International organisations are uniquely positioned to facilitate knowledge transfer and resource-sharing across regions, building collective capacity that no single national civil society sector can achieve alone.

[47] Fertmann, M. & Kettemann, M.C., "Democracy in Flux: Order, Dynamics and Voices in Digital Public Spheres," Proceedings of the Weizenbaum Conference, 2021, https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Proceedings/Proceedings_Weizenbaum_Conference_2021.pdf

7. Conclusion

AI counter-speech is already shaping who can speak and be heard in digital public spheres. The evidence reviewed in this report makes clear that it can be effective, but only under specific conditions, and never without risk. Value misalignment causes backfire, cultural blindness produces failure, and unchecked automation concentrates power in ways that undermine the democratic values counter-speech is meant to protect.

The power asymmetries, legitimacy deficits, and governance failures identified in this report are not inevitable. They are the product of choices, about how systems are designed, who controls them, and what accountability mechanisms are put in place. Different choices remain possible, but only if policymakers, platforms, civil society, and international institutions act in coordination rather than in isolation.

AI already shapes online discourse, it already determines who can speak and be heard online. Whether that power is governed in ways that are transparent, accountable, and compatible with democratic values is the choice now before policymakers. That is the governance challenge this report has sought to address.

Bibliography

(CDMSI), C. o. (2025). Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression. *Council of Europe* (pp. 1-28). Brussels: Directorate General of Democracy And Human Dignity .

(GIFCT), G. I. (2021). *Content-Sharing Algorithms, Processes, and Positive Interventions Working Group*.

Adam, G. P. (2025). *Machine learning tools to (semi-)automate evidence synthesis: A rapid review and evidence map, 2025 review update*. Rockville: Agency for Healthcare Research and Quality.

Anthony Wise, X. Z. (2025, November 19). *A Crowdsourced Study of ChatBot Influence in Value-Driven Decision Making Scenarios*. From https://arxiv.org/html/2511.15857v1

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). *Generative AI may backfire for counterspeech*. Munich.

Brodowicz, M. (2025, February 11). *The use of AI in detecting and combating online hate speech* . From Aithor: https://aithor.com/essay-examples/the-use-of-ai-in-detecting-and-combating-online-hate-speech#1-introduction

Buerger, C. (2021). #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse . *Social Media +Society*, 1-17.

Echikson, W., & Knodt, O. (2018). *Germany's NetzDG: A key test for combatting online hate* . Brussels : CEPS.

Fertmann, M., & Kettemann, M. C. (2021). Democracy in Flux: Order, Dynamics and Voices in Digital Public Spheres. *Proceedings of the Weizenbaum Conference*

(pp. 78-82). Berlin: Weizenbaum Institute for the Networked Society – The German Internet Institute.

Fichtner, L. (2021). Moderating the Regulators/Regulating the Moderators NetzDG and online content moderation in Germany. *Democracy in Flux: Order, Dynamics and Voices in Digital Public Spheres* (pp. 21-24). Hamburg: Weizenbaum Institute for the Networked Society – The German Internet Institute.

Finale Doshi-Velez, M. K. (2017). *Accountability of AI under the law: The role of explanation*. Berkman Klein Center for Internet & Society .

Gambacorta, L., & Shreeti, V. (2025). *Unpacking the AI supply chain: market structure and policy implications* . Vienna: The European Money and Finance Forum.

Greta Damo, E. C. (2025). Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study. *International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 2-7). London: HAL.

Koetsier, J. (2020, June 09). *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*. From Forbes: https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=37b95ce54d03

Langton, R. (2018). Blocking as Counter-Speech . *New Work on Speech Acts*, 144-164.

Lazar, S. (2024). Legitimacy, Authority, and Democratic Duties of Explanation. *Oxford Studies in Political Philosophy*, 28-56.

Lisa Schirch, K. R. (2025). *A taxonomy of response strategies to toxic online content: Evaluating the evidence*. University of Notre Dame.

Loebbecke, C., Luong, A. C., & Obeng-Antwi, A. (2021). AI for tackling hate speech. *European Conference on Information Systems*, (pp. 1-13). Cologne.

Lorna McGregor, D. M. (2019). International human rights law as a framework for algorithmic accountability. *British Institute of International and Comparative Law* , 1-35.

Mehmet Bedii Kaya, M. F. (2024). Social Media Regulation . In *The Economics and Regulation of Digitalisation, The Case of Turkey* (p. Chapter 14). London: Routledge.

Niklas Felix Cypris, S. E. (2022). *Intervening against online hate speech: A case for automated counterspeech*. Technical University of Munich; School of Social Science and Technology; Institute for Ethics in Artificial Intelligence.

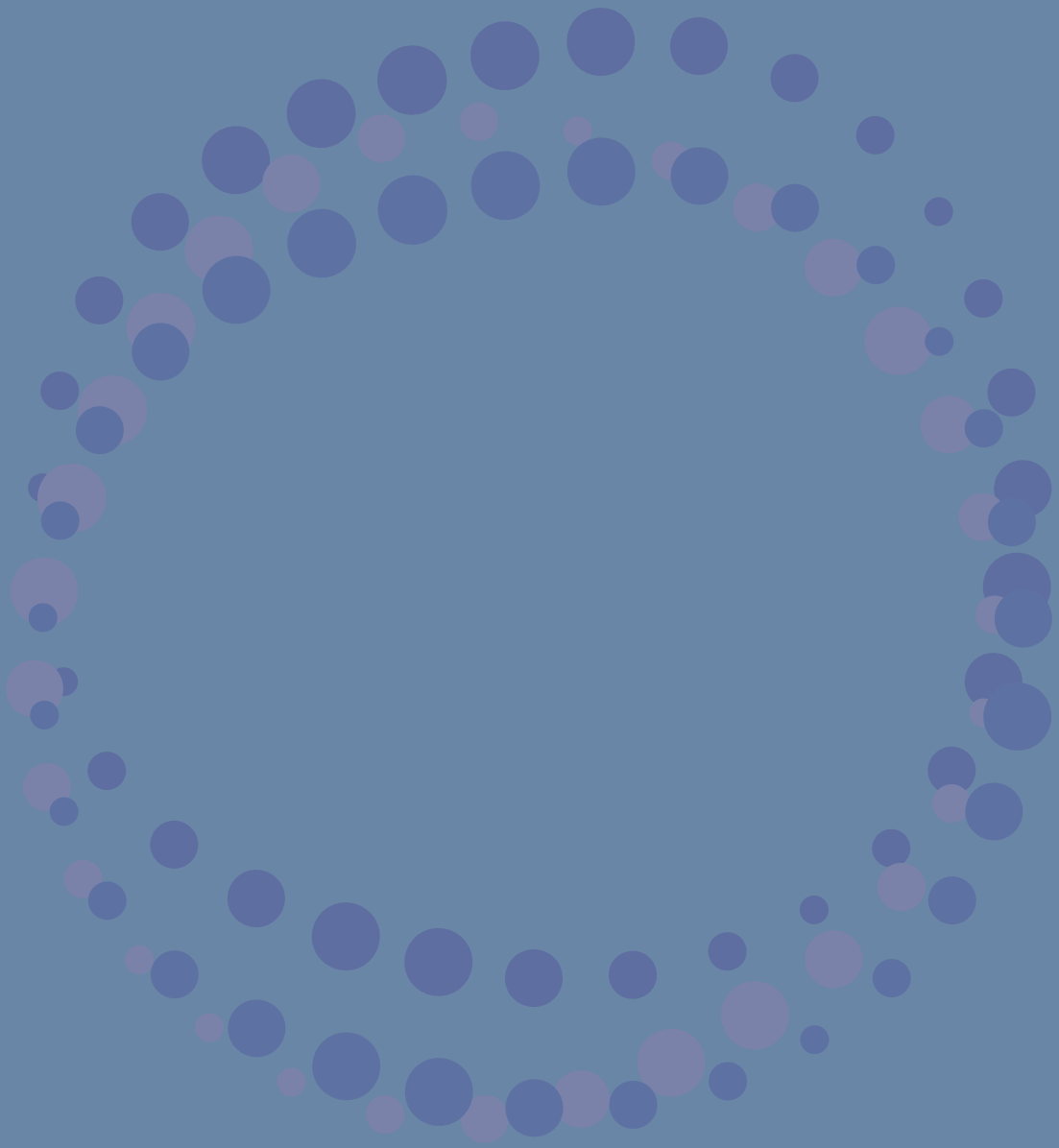
Sahana Udupa, E. H. (2021). *Artificial intelligence, extreme speech, and the challenges of online content moderation.* AI4Dignity.

Udupa, S. (2021). *Digital Technology and extreme speech: Approaches to counter online hate.* United Nations Peacekeeping Technology Strategy.

Vipra, A. K. (2024). *Concentrating intelligence: Scaling and market structure in artificial intelligence.* New York: Institute for New Economic Thinking.

Yi-Ling Chung, G. A. (2024). Understanding Counterspeech for Online Harm Mitigation . *Northern European Journal of Language Technology*, 30-49.

Yoshua Bengio, S. H. (2023). Capabilities and risks from frontier AI. *AI Safety Summit* (pp. 1-32). UK: Department for Science, Innovation and Technology.



Solidarity With OTHERS
Belgium, 2026
www.solidaritywithothers.com
info@solidaritywithothers.com