

The Medium Is the Menace: AI and the Platforming of Hate Speech

Houman Mehrabian

Blog post published April 27, 2026 in *Sounding Out!*

<https://doi.org/10.59350/9yh3z-qr683>

Abstract

The essays collected in this series (link to the Introduction) trace how nonhuman listening operates through sound, speech, and platformed media across distinct but interconnected domains. Across these accounts, listening no longer secures meaning or relation;

Keywords

Music (Subfield), Hate Speech and Cyberbullying Detection (Topic), Artificial Intelligence (Subfield), Article, Artificial Intelligence, Capitalism, Digital Humanities, Digital Media



Recommended citation

Mehrabian, H. (2026, April 27). The Medium Is the Menace: AI and the Platforming of Hate Speech. *Sounding Out!*. <https://doi.org/10.59350/9yh3z-qr683>

Copyright

Copyright © Houman Mehrabian 2026. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



The essays collected in this series (link to [the Introduction](#)) trace how nonhuman listening operates through sound, speech, and platformed media across distinct but interconnected domains. Across these accounts, listening no longer secures meaning or relation; it becomes a site of contestation, where sound is mobilized, processed, and weaponized within systems that privilege circulation, recognition, and response over truth. Last week, **Olga Zaitseva-Herz** examines how nonhuman listening operates under conditions of war, where AI-generated voices and deepfakes destabilize the very grounds of auditory trust. Through the case of Ukraine, she shows how platforms and political actors alike exploit algorithmic listening systems to amplify affect, circulate disinformation, and transform voice into a tool of psychological warfare. Listening, in this context, becomes not a means of understanding but a terrain of uncertainty. Today, **Houman Mehrbadian** turns to the dynamics of speech on social media, arguing that platforms do not simply fail to regulate hate but structurally amplify it through forms of proximity that render identity itself as a site of perceived threat. **—Guest Editor Kathryn Huether**

The Medium Is the Menace: AI and the Platforming of Hate Speech

DOI: <https://doi.org/10.59350/9yh3z-qr683>

During the Cold War, when the world was divided into two geopolitical poles, the *International Covenant on Civil and Political Rights* was drafted to enshrine individual rights. [Article 19](#) guarantees freedom of expression through any medium, "regardless of frontiers." The media landscape has significantly changed since then—we have transitioned from an information system where newspapers, radio, and television were dominant communication technologies to one where digital and online media play a central role, especially in transcending national frontiers. The Internet has amplified the right to free speech dramatically. Take, for instance, the decentralized hacktivist collective [Anonymous](#), gliding undetected past gatekeeping mechanisms to route confidential information to the public gaze or to rally digital protests, only to disappear again into the shadows of cyberspace. Yet, Article 19 also insists that this freedom carries "special duties and responsibilities": expression may be restricted when necessary to protect or prevent harm. The Internet, however, has challenged the enforcement of such laws in unprecedented ways.

What has changed is not only how speech circulates, but how it is heard—now increasingly by automated systems that register patterns rather than consider context. As Kathryn Huether explains in the [introduction](#) of this series, this shift marks the emergence of a new form of "nonhuman listening": a mode of perception in which speech is registered as data, classified and acted upon without ever being encountered as expression. Take, for instance, practices such as trolling, doxing, and flaming. Cyberbullies discover ever-new ways to propagate harmful content without raising the alarm bells of automated systems. [Tamar Mitts](#) explains how the digital ecosystem creates "safe havens" for online extremism: extremist groups persist by migrating to more permissive platforms, mobilizing aggrieved users to strengthen their group identity, or reformulating their messaging to slip past automated detection. As [major platforms](#) dial back their governance measures, those who disseminate toxic content grow ever more "[resilient](#)."

Digital technologies have opened new pathways for bad actors to take advantage of the protections of free speech. While this helps explain the growing volume of hate speech online, it addresses only the surface, the content itself. Even with robust content moderation tools in place, the deeper problem lies in the design of these platforms. Their very structure enables polarized expression and, its most pernicious manifestation, hate speech—precisely what Article 19 and related human rights frameworks seek to prevent. This severely hinders meaningful dialogue in the increasingly interconnected and interdependent world that these technologies have created.

In the [global social media economy](#), the United States sets the dominant tone. Its major platforms—Meta's Facebook, Instagram, and WhatsApp, alongside Google's YouTube, and others such as X, Pinterest, and Snapchat—shape what is circulated and amplified across the world. These companies operate under the protections of the [First Amendment](#) of the U.S. Constitution, which places relatively few restrictions on hate speech. By contrast, many other nations impose far stricter limits on online expression—from laws regulating hate speech in countries such as [Germany](#) and [France](#) to the extensive censorship regimes of states like [China](#) and Russia. Yet the question of free expression in the digital age exceeds the legal framework of any single nation, no

matter how powerful the prohibition or permission. More fundamentally, the medium of digital communication itself has transformed what speech is, and how it functions.



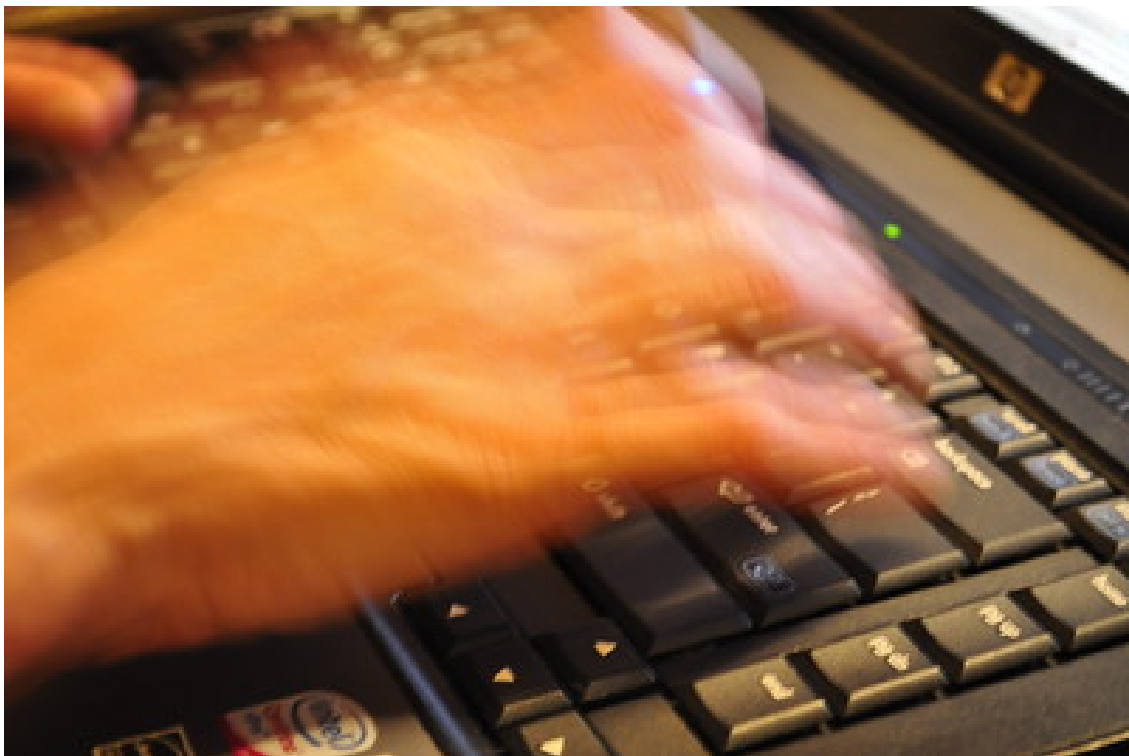
AVAAZ Protest in London, April 26, 2018, Photographer Rob Pinney, [PDM 1.0](#)

The era of private thinking and personal reasoning has given way to that of instant, public sharing, and everything shared on networking platforms is processed through algorithms running on binary coding. Algorithms should not be regarded simply as nonhuman—as alien intrusions into daily life—but rather as a sophisticated extension of a mode of human thinking that reduces complexity and nuance to mutually exclusive opposites. At the most basic level, these systems translate speech, images, and interaction into discrete units of data—encoded as sequences of zeroes and ones—and sort them through processes to classify. In doing so, they do not "listen" and interpret meaning in a human sense; they detect patterns and correlations across vast datasets. In this sense, nonhuman algorithms proliferate the dichotomy of "us" versus "them." They entrench what [Keith Stanovich](#) calls "myside thinking"—a widespread inclination to interpret the world through the lens of prior beliefs and loyalties. Appearing in every stage of thinking, across disciplines, and in all demographic groups, "myside bias," Stanovich argues, is more powerful than other types of bias because it involves "emotional commitment and ego preoccupation." Its greatest danger is that it prevents communities from converging on shared facts, even when evidence is available.

Algorithmic systems amplify thought attuned to binaries and, in turn, cultivate speech that gravitates toward extremes. Opposition intensifies into antagonism, nuance dissolves into simplicity, and complexity flattens into stark contrasts. To grasp this dynamic, it is essential to examine the underlying mechanisms of nonhuman listening that nudge speech in this direction. An illuminating lens is offered by [Judith Butler](#), whose account of "implicit" or "unspoken" modes of speech regulation reveal how discourse is shaped even before explicit prohibitions limit it.

These are conditions of intelligibility, frameworks that determine—in advance—what registers as meaningful speech, what recedes as noise, and what is never heard at all.

Online, discussion is always up and running. Breaking the silence or ending the conversation is almost unheard of in the digital realm. Ironically, this feature of Internet-mediated communication can itself function as speech control. It recalls what [Michel Foucault](#) describes as endless "commentary," in which discourse continually folds back on itself, repeating and reworking what has already been said. Silence becomes nearly impossible. In the words of [Gilles Deleuze](#), the user becomes "undulating," continuously "surfing" across interconnected spaces, each interaction rippling outward across the network. Repetition is vital for platforms that reward virality. Content creators, for instance, are encouraged to "[repurpose](#)" their old ideas and, in turn, encourage their audience to "co-create" their already-recycled ideas. In this sense, nothing truly begins or ends.



"Fast Typer" Image by Flickr user [jamiev_03](#) ([CC BY-NC-SA 2.0](#)).

Algorithms are not designed to propel free movement; [recommendation systems](#) learn from simple behavioral cues—a like or a skip, a pause or a quick swipe away—to incentivize us to go with the flow of [hyper-personalized data](#) and to affiliate with [echo chambers](#) of like-minded users. Even generative AI replies to each prompt in light of the ones that came before, becoming increasingly "[sycophantic](#)." This explains why a growing number of people, especially younger individuals, are [turning to artificial intelligence for friendship](#). These technologies offer something that mimics attentive listening, a feeling that the user's words do not go unheard. Artificial intelligence devices such as the [Friend](#) necklace are designed to make this type of connection effortless and always within reach.

Under these conditions, free speech comes to mean access to flows of information—the ability to move with them, rather than to analyze, interrupt, or challenge them. Listening becomes adaptive and reactive, attuned less to sound argumentation than to speedy circulation. Within insulated

echo chambers, expressions are encountered not as opinions to be evaluated but as signals to be affirmed or rejected. Memes, emojis, and abbreviated forms of expression condense complicated positions into immediate affective cues, eliciting responses of pride, indignation, gloating, mockery, delight, disappointment, disdain. The list goes on. What circulates most readily is not sustained reasoning but intensified feeling, shared across networks of both human and nonhuman participants.

This is not to say that debate has no place in the digital world. On the contrary, platform environments are configured to reduce nearly every issue (controversial or not) into a [rigidly polarized dispute](#). Algorithmic systems, optimized for engagement, sort content into recognizable positions, amplifying contrast and conflict. Issues are framed less as open questions than as preconfigured disputes, with sides already drawn and reiterated across countless iterations. One is either "woke" or dead set against it. Greta Thunberg's activism is either [inspiring](#) or [self-promoting](#). Online, users need only choose a side and signal agreement through simple actions—a like, a repost, a heart, an angry face. Digital debate becomes echoed: each side recycles familiar arguments that reinforce group identity rather than persuade others. This resembles the house war of the Montagues and the Capulets, with no hope of reconciliation.



"On the Tube" by Flickr User marneejill, London 2024 [CC BY-SA 2.0](#)

Even truth is drawn into this binary logic, as its validity now lies in how closely it aligns with one's viewpoint. Platforms like Truth Social—the social media site launched by Trump in 2022 and described as "[free from political discrimination](#)"—reinforce this dynamic by presenting "truth" as something to be claimed by one side, with opposing views dismissed as fake news.

The same pattern appears in responses to deepfakes. Also in 2022, a manipulated video of President Volodymyr Zelensky [falsely urging Ukrainian troops to surrender](#) circulated online. While widely debunked, its reception still followed partisan lines: dismissed as propaganda for some, and treated by others as plausible or strategically meaningful within existing narratives. [Olga Zaitseva-Herz](#) discusses other examples of AI-generated voices and videos used as psychological weapon in warfare. More broadly, deepfakes are often framed as satire or humor when they support one's perspective, and condemned as disinformation when they do not. Despite the apparent complexity of digital media, this dynamic reduces debate to a series of rigid oppositions. Under these conditions, dialogue becomes difficult to sustain—or even non-existent—as positions are evaluated less through exchange than through alignment.



"Fake_News-vor-Hacker" Image by Flickr User Christoph Scholz, (CC BY-SA 2.0)

Dialogue is often proposed as the answer to hate speech. The [United Nations Strategy and Plan of Action on Hate Speech](#) describes it as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are." This definition does not fully capture the underlying dynamic. While hate speech targets the identity of others, it is often driven by the perception that those identities pose a threat to one's own. [Research on hate suggests](#) that this perception is not tied to a single action but to a broader attribution of the other as inherently dangerous or malicious. Hate, in this sense, does not respond to behavior; it calcifies identity itself as a source of threat. Hate thus becomes, as [Daniele Battista](#) puts it, an ideal "communicative asset" for driving the digital economy.

Hate speech is the violent defense of an insecure self; it is Iago, ever sensitive to the closeness of the dissimilar other; it is yet another extreme manifestation of the us-versus-them mindset. But we should not confine our understanding of online hate speech to the level of content. The [amplification of harmful communication](#) is not merely due to mobilization of verbal violence by political figures or technical failures of content moderation systems—such as hate speech slipping through as free speech—but is more fundamentally a formal effect of the platforms themselves. Collapsing geographic and cultural distances, the Internet brings diverse users into unprecedented forms of closeness. This structure reflects what [Marshall McLuhan](#) diagnoses as the "implosive" character of modern media, in which boundaries contract and differences are forced into constant contact. Under these conditions, both users and automated systems are overwhelmed with volume, and listening—human and nonhuman alike—becomes reactive rather than responsive. The patient work of contextual understanding disappears beneath the flood of signals.

By their very design, these shared virtual spaces place the user's sense of self under continuous pressure. In response, users align with particular influencers and subscribe to particular channels to "[strengthen feelings of belonging and opposition](#)." Speech, then, tends to take on a defensive quality, reinforcing identity against perceived threat. Digital platforms do not simply host hate speech; they develop the very conditions in which it emerges. Prolonged interaction and sustained proximity in polarized environments make communication more likely to be shaped by anxiety than by dialogue. What follows is not a failure of communication, but a transformation of it: speech no longer seeks to understand the other, but to secure the self.



"Binary" / 12" x 12" / acrylic, pastel on fibreboard ©2011 [Braydon Fuller](#) (CC BY-SA 2.0)

Returning to the framework of this series, we can understand the shift to digital mediation as one in which listening collapses into the reiterative reception of preconstituted positions and oppositions, precipitating immediate, affectively saturated reactions that merely reproduce them. Increasingly "[detached from sensation, exposure, and accountability](#)," listening operates less as an encounter with speech than as a mechanism of bias confirmation by selectively sorting information.

Amid digital closeness in environments marked by binary thinking, the more users are "[silenced by speech](#)," the more listening becomes passive. We need to distance ourselves from communication as an instrument of pacification, or worse, suppression. Dialogue begins with attentive listening—not only to the speech of others, but also for polarizing mechanisms that surround us both online and offline. More importantly, we need to appreciate the formal effects of a listening that is not reduced to a rehearsal for rebuttal, a listening that is an antidote to the restless compulsion to react to speech that our digital devices incessantly fuel. In suspending the immediacy of response, listening evolves into a delaying tactic, a deliberate deferral that carves out an interval within which patient reflection may find form.

Featured Image by Flickr User Jeff Gates, [CC BY-NC-ND 4.0](#)

Houman Mehrabian earned his doctorate in English from the University of Waterloo (2020), where he focused on the history and theory of rhetoric. Currently, he is an Assistant Professor in the Arts, Communications & Social Sciences department at University Canada West. His research interests include exploring the rhetorical and technological mechanisms that regulate speech. Bringing together perspectives from critical media studies, philosophy, and rhetorical theory, his work investigates how the structural design of digital platforms and their economic logics can amplify harmful discourse, and how appeals to more free speech in online environments can operate as rhetorical cover for the proliferation and normalization of hate speech. Through this lens, his research aims to better understand the interplay between technology, power, and communication in the digital age.



REWIND! . . . If you liked this post, you may also dig:

Hate & Non-Human Listening, an Introduction—Kathryn Huether

Mimicked Voices and Nonhuman Listening: AI Deepfakes, Speech, and Sonic Manipulation in the Digital War on Ukraine—Olga Zaitseva-Herz

Impulsive: Bro-casting Trump, Part I—Andrew J. Salvati