PDF Download
3757589.pdf
24 February 2026
Total Citations: 0
Total Downloads: 237

Published: 16 October 2025

Citation in BibTeX format

RESEARCH-ARTICLE

Empowering Creators in the Fight Against Online Hate: A Qualitative Exploration of AI-Mediated Counterspeech Tools

PHOEBE YIQING HUANG, University of Washington, Seattle, WA, United States

JIAMING DENG, University of Washington, Seattle, WA, United States

YINGCHEN YANG, University of Washington, Seattle, WA, United States

SPENCER WILLIAMS, University of Washington, Seattle, WA, United States

Open Access Support provided by:

University of Washington

Proceedings of the ACM on Human-Computer Interaction, Volume 9, Issue 7 (October 2025)
https://doi.org/10.1145/3757589
EISSN: 2573-0142

Empowering Creators in the Fight Against Online Hate: A Qualitative Exploration of AI-Mediated Counterspeech Tools

PHOEBE YIQING HUANG, University of Washington, USA
JIAMING DENG, University of Washington, USA
YINGCHEN YANG, University of Washington, USA
SPENCER WILLIAMS, University of Washington, USA

Content creators face heightened risks of online harm due to their public visibility and the increasing hostility in digital spaces. Traditional moderation tools, such as reporting and deletion, often fall short in addressing the emotional and reputational impact of online hate. In this study, we explore how community-driven approaches like counterspeech may support creators in navigating hateful comments, and how AI-powered tools could assist in this process. We conducted qualitative interviews and concept testing with 15 active content creators from Chinese social media platforms to understand their experiences and expectations. Our findings reveal that creators' approaches to counterspeech are shaped by their professional roles, the need to maintain authenticity, and the pressures of audience management. While creators saw value in AI-assisted counterspeech for improving efficiency and tone, they expressed concerns about misrepresentation, loss of control, and unintended harm. They preferred AI tools that augmented their voice rather than replaced it, and emphasized the importance of balancing emotional nuance with scalability. These insights inform the design of creator-centered AI moderation tools that integrate human agency and context sensitivity, contributing to healthier and more sustainable online environments.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Content Creators, Counterspeech, Hate Speech, Content Moderation, AI-Mediated Communication

ACM Reference Format:
Phoebe Yiqing Huang, Jiaming Deng, Yingchen Yang, and Spencer Williams. 2025. Empowering Creators in the Fight Against Online Hate: A Qualitative Exploration of AI-Mediated Counterspeech Tools. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW408 (November 2025), 27 pages. https://doi.org/10.1145/3757589

1 Introduction

On social platforms, creators - those who produce and share digital content to engage and grow audiences - play a major role in shaping online conversations. However, despite their creative freedom and influence, creators are increasingly vulnerable to online hate due to their amplified visibility and the growing hostility of online space [53, 81]. Research indicates that 95% creators have encountered some form of online hate or harassment [20, 56, 80, 83], considering it an inevitable aspect of their public presence.

To cope with this unavoidable challenge associated with their career, creators rely on a range of measures, both manual and platform-provided, including community guidelines [86, 93], algorithmic

Authors' Contact Information: Phoebe Yiqing Huang, phoebeh.pd@gmail.com, University of Washington, Seattle, Washington, USA; Jiaming Deng, miked232@uw.edu, University of Washington, Seattle, Washington, USA; Yingchen Yang, yingcy2@uw.edu, University of Washington, Seattle, Washington, USA; Spencer Williams, spencer1918@live.com, Information School, University of Washington, Seattle, Washington, USA.

ACM 2573-0142/2025/11-ARTCSCW408
https://doi.org/10.1145/3757589

filters [6, 58], moderation tools [78], and reporting mechanisms [83]. Yet, these mitigation strategies face notable shortcomings, such as inadequate follow-through in the reporting process [11], vague and incomplete guidelines [65], and emotional burden on moderators [83]. Moreover, algorithmic detection systems are plagued by biases and inaccuracies [5, 22]. In particular, deletion-based tools fail to address the creators' deeper needs, offering neither the emotional support they require [24, 92] nor a permanent solution to silence persistent haters [65]. As a result, many creators opt for self-censorship or even withdraw from social platforms entirely [83].

While efforts to improve censorship algorithms and moderation tools continue, researchers are also exploring new, sustainable community-led approaches to reducing online harm. Counterspeech, defined as directly responding to hate speech [8, 36], is widely recognized for fostering empathy and understanding in online communities [33, 34] by promoting positive comments over hate [90]. Prior research shows that, when used effectively, counterspeech can correct subtle forms of stereotypes [8, 27], prevent misinformation [14], and comfort victims [59] by changing the mindsets of haters or influencing bystanders. The rapid advancement of AI enables this moderation strategy to scale, addressing challenges encountered during implementation [34, 95]. We argue that counterspeech can serve as a powerful method to protect creators from online harm by countering hate speech, without significantly compromising their mental health or resorting to avoidance.

Despite extensive exploration on counterspeech—including responses [34], counterspeakers [88], haters [4], bystanders [62], and AI-powered tools [75]—little is known about creators' perspectives on this moderation strategy as a response to hateful comments, as well as their perception of AI involvement in this process. As highly visible figures who manage larger volumes of comments and navigate more complex interactions than typical users [18], creators face unique challenges when dealing with online hate. Their goals often extend beyond simple participation, as their livelihoods and reputations are closely tied to how they handle online discourse [41, 57, 64]. While AI-powered tools hold promise for scaling these efforts, their potential to specifically assist creators remains underexplored. Specifically, the degree of AI involvement can significantly influence users' acceptance of AI-powered tools [2, 42]. Therefore, understanding creators' experiences and views on these tools will address existing gaps and guide the development of more targeted, impactful solutions.

To add content creators' viewpoints to the existing body of research, we conducted qualitative interviews with 15 creators who are active on Chinese social media, as well as a concept testing featuring three AI-powered tools with varying AI involvement to address the following questions:

- **RQ1**: What are creators' perspectives on counterspeech, particularly regarding their experiences and perceptions of its effectiveness?
- **RQ2**: How do creators perceive AI-powered counterspeech tools with varying AI involvement?

Through iterative coding and thematic analysis of interview and concept testing data, this study provides new insights into how content creators navigate online hate using counterspeech and how they evaluate AI-mediated tools designed to support this practice. We find that creators' counterspeech strategies are shaped by their public visibility, the need to maintain authentic audience relationships, and the ongoing pressure to manage their online persona. Concept testing further reveals that creators assess AI-mediated counterspeech tools through a nuanced set of priorities, including agency, autonomy, emotional tone, moderation efficiency, and audience perception. Although many recognized the potential of AI assistance, they also expressed concerns about misrepresentation, overreach, and unintended consequences. These tensions point to a need for balanced moderation approaches that integrate both proactive and reactive strategies.

In summary, this paper presents findings on creators' perspectives, strategies, and needs regarding counterspeech and AI-mediated moderation tools. We contribute:

(i) A qualitative study that examines content creators' motivations and strategies for engaging in counterspeech, highlighting how their needs differ from those of typical users;
(ii) A conceptual framing of creators' perspectives on AI-driven counterspeech, situated at the intersection of AI-Mediated Communication and content moderation; and
(iii) Design implications for developing AI counterspeech tools that support creator agency, maintain community trust, and foster healthier online interactions.

2 Related Work

2.1 Creator's Challenges

Social media platforms have become central to the growth of the creator economy, an industry now valued at over $200 billion [74]. However, the structure and operation of these platforms vary significantly in geopolitical and cultural contexts. In this economy, individuals known as content creators, build dedicated communities and generate income from their content [48], which they share across various social platforms [9]. However, as their visibility increases, so does their exposure to hateful content [36], often targeted by online users. Research indicates that nearly all creators report facing some form of hate or harassment, with 70% enduring recurrent bullying, trolling, and identity-based attacks [56, 83]. Consequently, online harassment often leads content creators to mental health issues such as low self-esteem, sleep disorders, increased anxiety, and feelings of fear and insecurity [15, 17, 45].

The pervasive nature of online hate presents unique challenges for content creators, who face a dual burden [52]: they must simultaneously act as moderators [63], protecting their community [96], while coping with their own experiences as targets of harassment [47]. Unlike typical users who can simply block or avoid hostile interactions, creators operate within a complex ecosystem [96] where content moderation decisions interweave with their need to build a stable fanbase [67, 71], monetize their content [43, 69], and maintain authenticity amidst external pressures [19, 38]. These overlapped responsibilities create multiple layers of complexity in the way creators handle online harassment. They must simultaneously weigh the impact on their mental well-being [10, 68], community health [55], brand reputation [41, 64], and business sustainability [21] when addressing hate speech. Given the heightened complexity in creators' experiences with online hate, our research aims to provide an in-depth understanding of the unique challenges content creators face and explore more effective moderation strategies that acknowledge both personal and professional dimensions from their particular perspectives.

2.2 Online Moderation Tools

Traditional platform-provided moderation tools, such as content removal [7, 78], user banning [50], and automated detection filtering systems [1, 6], have been widely implemented across social media platforms to combat online hate speech. However, these deletion-based approaches are often criticized for their lack of transparency and effectiveness due to their platform-driven nature [89]. As a result, recent innovations have expanded beyond platform-driven tools, exploring solutions such as Squadbox for designated member review [54], customizable moderation interfaces [42], and curated keyword databases [35] to promote nuanced content management. Nevertheless, these approaches remain limited by their primary focus on content deletion, quarantine, and filtering, without addressing the deeper social dynamics within online communities. This underscores the need for more sustainable and community-centered approaches.

A shift toward community-led moderation offers promising alternatives and complements to traditional methods. Unlike platform-driven moderation, community-led moderation allows users and communities to actively participate in humane content moderation [78, 91], promoting healthier and more inclusive online spaces [82]. This approach emphasizes empathy and understanding, enabling users to collaborate, maintain respectful interactions, and create a more welcoming digital environment. Examples include bystander intervention programs that encourage self-directed counterspeech [72], peer-based moderation systems [28, 33], and community-generated blacklists like Block-together [12]. Although community-driven approaches show promise for fostering more sustainable online environments, their effectiveness in addressing the specific needs of content creators remains largely unexplored. It is essential to examine how these emerging community-led solutions, particularly counterspeech, can support creators in combating online hate while promoting values of transparency, empathy, and freedom of expression [14, 46]. This represents a critical gap in current research, and our study aims to explore counterspeech tools specifically designed to support content creators in these efforts.

2.3 AI-Mediated Communication

As artificial intelligence becomes increasingly embedded in online platforms, a growing body of research has examined how AI agents modify, augment, or generate messages on behalf of users, a process known as AI-Mediated Communication (AI-MC) [32, 73, 76]. Unlike traditional computer-mediated communication, where users retain full authorship over their messages, AI-MC introduces a third agent into human interaction. Hancock et al. propose that AI-MC systems operate in multiple dimensions, including the magnitude of AI intervention, the type of media (text, audio, video), optimization goals (e.g. trustworthiness or humor), and the level of autonomy granted to the AI agent [32]. These dimensions raise important questions about how communicative agency, authenticity, and intent are perceived by both senders and receivers.

In particular, prior work has identified tensions in AI-MC between the benefits of AI assistance and the risks of reduced trust and authenticity. For instance, the "Replicant Effect" suggests that audiences often view AI-generated messages—such as self-presentations or responses—as less trustworthy than those authored by humans, even when the content quality is comparable [39]. While such findings offer useful insights, it remains unclear how these perceptions manifest in the context of counterspeech, where tone, personal voice, and cultural alignment are especially critical. This is particularly relevant for content creators, who may rely on a consistent communicative style to maintain audience engagement and trust.

To address this gap, we explore how creators perceive and evaluate AI-generated counterspeech across varying levels of AI involvement. Specifically, we investigate whether known dynamics from AI-MC—such as concerns over authenticity, agency, and alignment—apply in the emotionally charged and socially sensitive context of online hate and its mitigation.

2.4 Counterspeech and AI Involvement

Building on the conceptual foundations of AI-mediated communication, one particularly relevant application is the use of AI to support counterspeech—an emerging strategy for addressing online hate in more dialogic and scalable ways. Counterspeech, where bystanders actively respond to hate speech through rational argumentation, fact check, or emotional appeals [16], has proven effective in countering online hate by directly challenging instigators [90]. Strategies of implementing counterspeech vary widely, from individual volunteer efforts to coordinated group actions [26, 87] such as the #iamhere movement, which amplifies the impact of counterspeech by presenting a unified front against hate. However, limitations exist within applying counterspeech in practice, including that it is time intensive (Mun, 2023), lacking immediate feedback or rewards [16], potential for backfires [30], and feeling burdensome to implement [30]. In addition to automation and AI-assisted tools, AI-driven counterspeech presents promising solutions to mitigate these challenges, enabling more efficient and scalable interventions [14]. Recent advances have allowed AI to detect harmful content and suggest relevant counterspeech strategies using natural language processing and machine learning [34, 95]. However, AI's limitations in understanding context, sarcasm, potential bias, and nuanced language can lead to inappropriate or ineffective counterspeech responses [3, 77, 84], which, for content creators, may feel inauthentic or misaligned with their style, potentially impacting audience trust and engagement (Mohamed, 2024). Furthermore, creators' perspectives on AI-driven counterspeech tools remain underexplored, as they must balance audience expectations, brand identity [94], and community management when utilizing such tools [25]. Given these complex challenges, gaining insight from content creators is invaluable, particularly by involving them in the design process of counter-speech tools. Our research investigates creators' viewpoints on AI-powered counterspeech tools to enhance counterspeech efficacy while supporting creators in preserving their reputation and fostering a positive community.

3 Method

In this research, we used a qualitative approach to explore content creators' perspectives on counterspeech and the potential role of AI-powered tools in online moderation. The study involved two parts: semi-structured interviews with 15 creators to understand their experiences navigating online hate and counterspeech, followed by a concept test in which participants evaluated three speculative AI-powered counterspeech tools. This two-part design allowed us to examine both current practices and future design opportunities.

3.1 Study Design

3.1.1 Interview design. We employed semi-structured qualitative interviews to understand content creators' experiences with counterspeech and its perceived effectiveness in managing online hate. This approach encouraged participants to freely share their thoughts and experiences while allowing for follow-up questions to explore emerging themes. Interview questions were primarily open-ended and explored creators' motivations, challenges in managing online hate, and their engagement with counterspeech. We also asked participants to reflect on counterspeech as a potential form of content moderation.

3.1.2 Concept test design. Originally, we planned participatory design sessions to co-develop counterspeech tools with content creators. However, pilot tests revealed that participants often relied on familiar tools (e.g., ChatGPT, Google Smart Reply) or needed extensive guidance, limiting open-ended exploration. We therefore shifted to a concept testing approach.

The three concept ideas were developed through a collaborative brainstorming process within our research lab, involving three researchers with a background in HCI and information science. During this process, we considered factors such as the identity of the counterspeaker (e.g., AI, creator, bystander), existing AI tools in AI-mediated communication, and different ways counterspeech strategies could be supported. Each concept represented a different level of AI involvement:

(C1) AI as an independent counterspeaker
(C2) AI assisting creators in crafting counterspeech
(C3) AI prompting bystanders to engage

During the session, each concept was described by the interviewer using a short scenario and script, prompting participants to imagine how they might use each tool within the context of their own communities and the platforms they regularly engage with. For example, the script for C2 began:*"Imagine you receive a harmful comment on one of your usual posts. This tool suggests* *replies or helps you revise your own, aiming to improve your response for counterspeech purposes. It's a co-writing assistant that supports you in crafting replies using different counterspeech strategies."* (see Appendix A for full scripts)

After each presentation, participants were asked open-ended questions such as: *"What do you like or dislike about this concept?"* Participants occasionally raised clarifying questions, such as whether the AI operates automatically or is user-controlled, which we welcomed as part of the open exploration. We invited them to share thoughts, concerns, and imagined use cases to understand how they envisioned using such tools in their own practice. After all three were presented, participants were asked additional reflective questions, such as: *"What are your thoughts on AI participating in comment sections to handle harmful content?"* and *"How might these tools affect your experience as a content creator?"*

3.2 Study Procedure

We conducted studies with 15 content creators recruited from various Chinese social media platforms. All sessions were conducted in Mandarin, approved by the IRB, and audio-recorded with participant consent. Three sessions took place via video call and twelve via audio call. Sessions lasted 30 to 60 minutes, typically split evenly between the interview and the concept test. One participant (P13) extended their session to 120 minutes due to strong interest, while another (P9) completed their session across two days due to an emergency. All sessions were conducted in May 2024.

Each session began with informed consent and a brief discussion of the creator's background, content type, and general moderation practices. The interview then focused on their experiences with hateful comments, prior use and perceptions of counterspeech, and comparisons with deletion-based approaches. To provide shared context, we introduced research on counterspeech strategies [8] and invited their thoughts before transitioning to the concept test.

In the concept test, the interviewer verbally described three speculative AI tools one by one. After each concept, participants were invited to reflect on its potential use, including perceived benefits and drawbacks. Participants sometimes raised clarifying questions, which we welcomed to encourage open-ended discussion. The interviewer also followed up on emerging ideas or concerns. After all three concepts were discussed, we asked broader questions about the role of AI in moderation and how such tools might affect their experience as content creators.

3.3 Recruitment

We distributed a screening survey through personal connections and a multi-channel network (MCN) agency—organizations that manage and support content creators across platforms—which shared the survey in creator group chats. The survey collected basic demographics and gauged interest in participation. Of the 45 valid responses that provided contact information, we selected 15 creators to ensure diversity in content type, follower count, gender, and platform use.

We gathered data on Platform mainly Post, Amount of Followers, Content Type, and Gender directly from participants during the interviews, and this information is summarized in the accompanying Table 1.

3.4 Chinese Social Media Platforms

We conducted this study in the context of Chinese social media platforms, including Xiaohongshu (RedNote), Douyin, Bilibili, WeChat, and Zhihu—all commonly used by our participants. Most participants in the study managed accounts across multiple platforms, reflecting the cross-platform nature of content creation in China. These platforms differ from Western counterparts in both technological affordances and content norms. Xiaohongshu (RedNote) is a lifestyle-focused platform popular for short posts and product recommendations, often blending personal narratives with influencer marketing. Douyin is the Chinese counterpart to TikTok, centered on short-form video and algorithm-driven content discovery. Bilibili is known for its youth-driven culture and video content, particularly in gaming, anime, and commentary. WeChat combines messaging with public accounts and social sharing, making it a hybrid of private and public interaction. Zhihu is a Q&A platform similar to Quora, where users engage in long-form discussions and knowledge sharing. These platforms are subject to a mix of platform moderation and state regulation, which shapes creators' experiences with content visibility, moderation, and community dynamics.

3.5 Qualitative Analysis

We analyzed the data using an iterative qualitative approach, drawing on principles from Iterative Categorization [60]. All interviews were transcribed verbatim in Mandarin. Prior to coding, the research team collaboratively developed an initial codebook based on our research questions and early familiarization with the data. This codebook provided a shared structure to ensure consistency during the initial round of open coding.

Three researchers independently coded a subset of transcripts using this preliminary codebook. After the first round, we met to compare coded excerpts, refine code definitions, and resolve discrepancies through discussion. This process allowed us to iteratively update the codebook to better capture emerging themes from the data. Coding was conducted in Mandarin to preserve linguistic and cultural nuance.

After finalizing the codes, we grouped related codes into broader thematic categories using affinity diagramming techniques. A final thematic framework was developed collaboratively to synthesize key insights across participants. The codebook and themes were then translated into English for reporting purposes, with all translations reviewed by bilingual researchers to ensure accuracy and consistency.

3.6 Translation

All interview quotes included in the paper were manually translated from Mandarin to English by two bilingual researchers. We focused on preserving participants' tone and intent while ensuring clarity for an academic audience. All translations were cross-checked to ensure accuracy and consistency.

3.7 Ethical Considerations

All participants provided their informed consent and were informed of their right to withdraw at any time. Data confidentiality and anonymity were ensured throughout. Audio recordings and transcripts were securely stored, accessible only to the research team.

4 Interview Findings

In this section, we present insights into creators' perspectives on counterspeech as a strategy for mitigating online harm. We examine their prior experiences with counterspeech, their evaluation of its effectiveness, and their views on its role within broader content moderation practices.

4.1 Counterspeech Experiences

We found that creators generally prefer to avoid engaging with hateful comments but may occasionally use positive, fact-based, or sarcastic responses depending on the situation and their concerns about public image. Beyond countering online hate independently, some participants adopted community-driven strategies, such as seeking help from friends or pinning comments to encourage broader input from bystanders, both of which were perceived as effective.

Table 1. Summary of participants interviewed

Participant ID	Major Platforms	Amount of Followers	Content Type	Gender
P1	RedNote (Xiaohongshu)	30,000+	Cover songs, Vlog	Female
P2	RedNote (Xiaohongshu)	13,000+	Work, Life	Female
P3	Tiktok (douyin), WeChat, RedNote (Xiaohongshu)	2,000,000+	Cultural comedy	Male
P4	RedNote (Xiaohongshu)	700,000+	Skincare	Male
P5	RedNote (Xiaohongshu)	38,000+	Study abroad, Feminism	Female
P6	Tiktok (douyin)	3,000+	Daily life, Dance	Female
P7	RedNote (Xiaohongshu), WeChat, Bilibili, Tiktok (douyin)	800,000+	Outdoors, Opinions	Female
P8	Tiktok (douyin)	1500+	Bartending, Singing	Male
P9	RedNote (Xiaohongshu)	3000+	Photography, Parenting	Male
P10	WeChat, RedNote (Xiaohongshu), Tiktok (douyin), Weibo	4000+	News, Education	Male
P11	Weibo, WeChat, RedNote (Xiaohongshu), Tiktok (douyin)	200,000+	Work, Life	Male
P12	RedNote (Xiaohongshu), Tiktok (douyin)	2000+	Daily life, Food	Female
P13	RedNote (Xiaohongshu)	10,000+	Home, Photography	Female
P14	RedNote (Xiaohongshu), Bilibili, Tiktok (douyin)	320,000+	Outdoors	Female
P15	RedNote (Xiaohongshu), Zhihu, WeChat	25,000+	Education, Study abroad	Female

4.1.1 Creator-led counterspeech.

Minimal engagement as a strategic approach. Most participants avoided responding to hate comments or engaged only minimally. They believed that real haters rarely sought meaningful dialogue because their goal is to attack. As P2 said, *"Well, it depends on what they're trying to do. If they just want to insult people,… "* One participant noted that engagement often felt futile, as even when commenters were persuaded, *"it doesn't mean they'll delete the comment. Like, maybe next time they see your video, they won't [comment]"* (P3). Several participants also viewed harsh comments as personal opinions rather than outright harm, and felt that, unless they involved personal attacks or discrimination, differing perspectives should be tolerated. As one explained, *"But sometimes I feel like... you're just the one expressing it, you know? Once you say it, all the explaining and interpreting, that's not yours anymore"* (P13). A few also framed silence as a resistance tactic that leaves room for reflection, as they described, *"This kind of space… invites reflection to dissolve tension—using silence to blur the lines between opposites"* (P13).

For those with large followings, public engagement carried reputational risk. They felt that any response could be seen as a personal stance, increasing the likelihood of backlash. One shared an example of a peer's comment from years ago still circulating as a meme: *"Once you respond, you've already kind of lost the situation. If I'm thinking ahead, I'd rather not give them anything they can screenshot or use against me"* (P11). There was also concern about how excessive counterspeech might be perceived. One participant cautioned that being overly responsive could be over-analyzed: *"It's like one of those 'the more you explain, the guiltier you look' situations. I probably wouldn't change their minds anyway"* (P8).

Positive and tactful responses. When participants did respond, they often relied on sincerity or humor. They might express understanding (P13), apologize (P3), agree with the commenter (P1), or use humor to defuse negativity (P9). For example, when wealth-shamed, one creator replied, *"Yeah, that's right. I'm really thankful to have parents who love me"* (P5). These responses were often used strategically—some believed that sincere acknowledgment could disarm the commenter. P14 noted that a few even deleted their comments afterward. More importantly, creators saw positive replies as a way to signal their stance to bystanders. As P10 put it, *"Just let the audience decide, whether they're on his side or mine."* Many also felt pressure to uphold a moral standard, particularly those with large followings: *"For most public figures, they don't really have that freedom. People expect you to be morally perfect. They want you to turn the other cheek, never fight back, always stay graceful and composed."* (P11).

Fact-based responses to misunderstandings. Most participants responded when hate came from misunderstanding or misinformation. They typically pointed out the error and provided evidence. For example, P7 responded to a makeup-related accusation by saying, *"Can you please watch it again? I'm literally not wearing any makeup at all. I'm completely barefaced in that video."* When misinformation spread widely, participants would pin a key comment with a rational reply to prevent confusion and increase visibility (P5). They observed that such replies often received more likes than the original hate comment (P2). If explanation failed, some felt justified in reporting (P3).

Rare countering cases. In rare cases, participants used sarcasm or biting humor, labeling the commenter a *"clown"* (P7), mimicking their tone (P6), or referencing personal details from their profile (P5). However, such responses were usually limited to smaller creators. P9 noted they felt freer to respond aggressively as a less prominent figure but would avoid this approach to protect their public image if they gained more visibility.

4.1.2 Community-led counterspeech.

Seeking external help. Some participants mentioned seeking support from trusted others. One who runs a consulting business on social media asked loyal clients to post honest reviews to balance out negative ones: *"I'll reach out to people I'm close with… ask them to share their real experience. It's kind of like balancing out the comments section with some positive feedback"* (P15). Others described receiving help from friends in group chats who would step in to counter unfair comments (P9). Some participants found it effective to pin hate comments to allow more people to see them and judge for themselves (P5). One participant explained that the pressure from others' reactions often leads the original commenter to delete their post: *Like, I'll pin their comment so others can go off on them. It's kind of like… letting them feel what it's like to be on the receiving end of hate. I mean, maybe then they'll realize—oh, so this is what it feels like to have someone throw that kind of stuff at you* (P7).

Bystander intervention. Most participants appreciated bystander support, describing it as *"like someone stood up for me"* (P9). However, they generally chose not to engage with these defenders, with some only liking their comments. As P1 explained, *"Most of the time I'd just think, 'Oh, someone's already speaking up for me—then I don't have to say anything myself.' It's kind of like that."* Others avoided any interaction to reduce visibility (P8) or to maintain neutrality. As P11 noted, *"When a public figure jumps into the comments and takes a side, it often just sparks another round of fighting"*, which might escalate the conflict. This is further illustrated by P14: *"If it's just people attacking each other, like, they're really no different from the person who started it, then I don't see the point in responding. Even if someone's defending me, they're still part of that same cycle…If I respond… it kind of drags me back into the conflict, and puts me in opposition to the original commenter again. It just keeps the fight going."*

Their decisions around whether to engage with bystanders were also shaped by platform norms and moderation rules. Some participants feared that aggressive comment sections could trigger penalties from the platform, harming their account's reach and visibility (P4). Conversely, a few participants were indifferent to the tone of the comment section and even embraced controversy to boost engagement. As P3 noted, *"I don't care. It's all just attention to me. As long as they're not coming at me directly, I'm fine with it."* Another participant shared that some platforms boost content with higher interaction and observed that some users even post extreme comments to attract attention, describing, *"some fans, sure, and also regular users, but a lot of the people posting are just trying to get attention off the buzz"* (P12).

4.2 Perceptions of Counterspeech

Participants generally viewed counterspeech as a more human-centered and constructive alternative to traditional platform moderation. They appreciated its potential to foster dialogue and reflection, especially when compared to punitive measures like content removal or user bans. However, they expressed mixed views on its overall effectiveness, as well as concerns about its emotional toll, practical feasibility, and cultural limitations, especially in algorithm-driven environments where hate and misunderstanding can be amplified.

4.2.1 Effectiveness depends on intent. Participants generally believed that the effectiveness of counterspeech depends heavily on the commenter's intent. Strategic responses were seen as helpful primarily when the commenter did not deliberately aim to cause harm. Some commenters were described as venting or joking without realizing the emotional impact of their words. As P6 noted, such individuals may just be *"bored and had nothing better to do,"* unaware of the consequences. P7 echoed this sentiment: *"There are people who… how do I put it… like you can kinda tell, they're just saying what they think, and they don't even realize they're hurting anyone. But then there are people who say stuff just to be mean. So yeah, it really depends on who you're dealing with."*.

In contrast, most participants believed that many haters act with clear intent to harm. Some were described as professional trolls who deliberately spread negativity to disrupt online spaces (P4). In these cases, counterspeech was seen as ineffective and even counterproductive. *"If you respond, you're basically giving them what they want… it's like feeding into their whole attention game,"* P14 warned. P7 mentioned that such commenters might only back down when faced with overwhelming public pushback, for example, by pinning their comment and letting bystanders respond. However, this tactic raised ethical concerns: *"That's basically like, hit me, I hit you back. Just a never-ending loop"* (P8).

Some creators also described facing coordinated attacks motivated by commercial competition or conflicting interests. P14 reported being targeted by paid trolls or bots leaving mass negative comments. P15 recalled an encounter where reasoning with facts had no effect—the attacker remained evasive and manipulative: *"I replied, but it didn't help at all… He doesn't even address the issue. I asked him to show evidence, and he just said, 'See? This teacher isn't professional at all.'… It's like, you say your part, he says his, and you're not even having the same conversation. That kind of comment is the hardest to deal with."* In these cases, where the intent was clearly malicious, participants felt that counterspeech not only failed but could entrap the creator in prolonged, emotionally draining interactions.

4.2.2 Counterspeech as a humane alternative. Despite these concerns, many participants still saw value in counterspeech, particularly when compared to harsher forms of moderation. P10, for instance, felt that automated platform bans—such as month-long bans—were often excessive. It was also seen as providing opportunities for reflection and dialogue. P14 appreciated that counterspeech allowed users to learn from differing views, while others felt it contributed to a more open environment by keeping critical comments visible rather than deleting them: *"I think it's great. Everyone gets to speak, and everyone has to think a little too"* (P2). One participant shared a notable example of a hater turning into a fan after receiving a thoughtful response (P3), suggesting that respectful communication can sometimes defuse hostility.

However, participants emphasized that counterspeech should complement, not replace, platform moderation. They felt that relying solely on creators to manage harmful content would be burdensome and emotionally taxing. As P4 noted, creators already face constant negativity and need system-level support to manage toxicity.

4.2.3 Time and emotional labor. While counterspeech was seen as potentially valuable, participants repeatedly stressed the practical and emotional burdens it imposes. Many said they lacked the time to respond to every negative comment, especially those with large audiences. Crafting thoughtful replies takes considerable energy (P3), and even then, the outcome is often uncertain. As P15 noted, *"Even if like, they think you make a good point, it doesn't mean they'll delete their comment,"* leading to lingering frustration. Moreover, engagement can invite further replies, which prolongs the conflict and drains emotional energy (P11). In some cases, creators experienced worsened hostility, eventually resorting to deleting, blocking, or reporting the commenter (P3).

Several creators preferred shifting their focus to positive interactions rather than *"holding back the toxic stuff"* (P11). Others echoed this sentiment, emphasizing emotional preservation over confrontation. *"I just want to keep the harm to a minimum, get it over with, and take it off my radar as fast as I can."* (P13).

4.2.4 Mixed-view on empathy-based approach. When introduced to research highlighting the effectiveness of various counterspeech strategies [8], participants expressed mixed views, particularly about the empathy-based approach. Some found it persuasive. P9 shared a successful experience using emotional appeal: *"Like, helping them see there's an actual person behind the screen. That really matters."*

However, others were skeptical. They felt that many hate commenters lack empathy and are unlikely to relate to creators. As P5 described, *"I mean, everyone's comment takes effort, right? So if someone actually takes the time to leave a hate comment, they really mean it. They want it to hurt. And honestly, if you tell them, 'That really hurt me,' and expect them to delete it… yeah, that just doesn't happen."* Some believed the root issue was a deeper social divide or miscommunication, and a few replies wouldn't resolve that, *"No one's mind ever gets changed in the comments"* (P14). Others felt the approach felt like self-defense than a solution and was outdated. P13 reflected that while such strategies may have worked in the past, today's echo chambers and rigid views can make even empathy-based counterspeech reinforce division: *"That's why you get hate comments. Later on, when you push back or say something, it's more like a way to protect yourself. I mean, it's probably not exactly a weapon, but maybe it works that way sometimes. And it's only when someone uses it on you that you realize, oh, I could do that too."*

Cultural context also shaped participants' skepticism. Several noted that empathy-driven counterspeech, while effective in Western contexts, might not work well in Eastern cultures. P15 observed that communication tends to be more indirect, explaining, *"Westerners are more direct, but Chinese people speak more subtly... When I heard that method, I got nervous. If I used it, it might come off too strong and trigger the commenter even more."* P8 added that showing vulnerability to invite empathy may not be well received on Chinese social media.

5 Concept Test Findings

This section presents insights from creators' evaluations of three AI-powered counterspeech tools, each involving different levels of AI intervention. Participants reflected on the perceived benefits and drawbacks of each concept from their unique position as content creators (see Table 2 for a summary).

Across participants, we observed a consistent preference for human-centered approaches that preserve creators' autonomy and foster authentic interpersonal connection, rather than fully automated, AI-driven interventions. Rather than viewing AI as a substitute for human voice, creators favored designs where AI plays a supportive role, assisting them with self-expression while keeping human agency at the forefront. Participants emphasized that effective counterspeech depends not just on message content but on who delivers it, highlighting the importance of empathy, tone, and relational nuance that AI alone cannot provide. This preference was also rooted in a broader concern over loss of control and a deep distrust of platform-driven automation, particularly when it risks misrepresenting their intent or undermining their relationships with audiences.

We structure our findings across three key dimensions: perceived usefulness in moderating online hate, alignment with creator goals, and impact on community dynamics.

Table 2. Summary of participant perceptions across AI-mediated counterspeech concepts *(Concept 1: AI as an independent counterspeaker; Concept 2: AI assisting creators in crafting counterspeech; Concept 3: AI prompting bystanders to engage)*

	Perceived Benefits	**Perceived Drawbacks**
Concept 1 (C1)	- Enhances efficiency in managing high comment volume (5.2.5) - Provides emotional support by preventing impulsive replies (5.1.2) - Serves as a neutral mediator in resolving factual or heated conflicts (5.1.2)	- Lacks emotional nuance and may sound insincere (5.2.5) - Ineffective in changing haters' behavior or building empathy (5.1.1) - Blurs accountability and may misrepresent the creator's voice (5.2.5) - Undermines authenticity and trust in creator-audience interaction (5.2.5) - Raises concerns about manipulation and platform control over discourse (5.2.1)
Concept 2 (C2)	- Supports creators in drafting thoughtful and effective responses (5.2.2) - Helps articulate thoughts when creators feel stuck or uninspired (5.2.2) - Aids emotional regulation and reduces impulsive reactions to hate (5.1.2) - Enhances tone and clarity, making counterspeech more reasoned (5.1.1) - Encourages more intentional and reflective engagement with commenters (5.3.1)	- May reduce efficiency due to added emotional and time investment (5.2.4) - Risk of generic or repetitive responses that lack personal touch (5.3.2) - Could lead to prolonged back-and-forth interactions with haters (5.2.4) - AI suggestions may feel intrusive or manipulative if not well-positioned (5.3.2) - Raises concerns about autonomy when AI influence feels too strong (5.2.2)
Concept 3 (C3)	- Encourages authentic, interpersonal counterspeech from real users (5.2.1) - Human responses are seen as more credible and impactful (5.1.1) - Supports independent thinking and situational judgment (5.2.2) - Empowers community members and fosters long-term cultural change (5.1.1)	- Difficulty in accurately identifying willing and capable counterspeakers (5.1.3) - Risk of bothering uninterested users, potentially reducing engagement (5.3.3) - Counterspeech from untrained users may lack effectiveness (5.1.3) - Potential to escalate conflict if AI invites biased or polarizing groups (5.1.3)

5.1 Perceived Usefulness in Moderating Online Hate

5.1.1 Human presence matters in countering hate. Participants viewed human involvement as essential to the usefulness of counterspeech in mitigating online hate. Analysis of their evaluations revealed a consistent belief that counterspeech is most effective when it fosters interpersonal connection, something they felt AI acting independently, could not offer. Designs that preserved human agency, such as AI assisting creators (C2) or enabling bystander participation (C3), were favored over those where AI responded autonomously (C1).

Many participants questioned whether independent AI agents could adequately address the emotional nuance and social complexity of hate-related interactions. Nearly all participants described the idea of an AI agent entering the comment section as *"strange,"* with one calling it *"unnatural"* (P3). Without the ability to express empathy or understand context, participants doubted that AI-generated responses could prompt meaningful reflection or defuse hostility. P7 noted, *"Some people might feel like the AI doesn't even get what they're saying."* Others viewed AI corrections as dismissive or even offensive: *"Who wants to be 'educated'? Especially by a machine. That's not what I asked for"* (P11). P4 noted this concern as well, speculating that *"If it's a hate comment and you let AI reply… it might make things worse. They might think, 'You won't even respond yourself? Hiding behind AI? Come say it yourself.'"* As P1 remarked, *"At some point people are like, wait, why am I arguing with a bot?"*

In contrast, human-in-the-loop approaches like C2 and C3 were viewed as more effective in moderating hate because they preserved authentic human involvement. Participants appreciated that C2 supported the process of crafting responses while allowing them to maintain their personal voice and control. This assistance was particularly valued in moments when participants felt misunderstood but lacked the right words. As P5 noted, *"…you feel the urge to reply, but sometimes you really don't know what to say. That's when having someone, or even an AI, to help would be nice…"* C3 was similarly appreciated for amplifying the voices of real users. As P6 explained, *"AI just gives you whatever you ask for, but people, you know, they've got their own minds. They can tell if it's something they should respond to or if it's even helpful."* Others highlighted the collective power of community-driven counterspeech: *"When more people jump in to push back, I just feel like it works better"* (P3).

Overall, participants believed that reducing online harm requires not just well-crafted responses, but the visible presence of real human voices behind them. They stressed that empathy, authenticity, and relational awareness are critical to de-escalating hate and promoting constructive dialogue. As P11 summarized: *"People like you because of that little bit of human touch… If everything you post is all polished and perfect, then what's the point?… In chasing efficiency and perfection, you lose that human side."*

5.1.2 AI's role in navigating conflict. Participants identified both opportunities and challenges in using the three AI concepts (C1–C3) to moderate online hate, emphasizing that effectiveness depended heavily on how these tools were implemented within specific conflict contexts.

C1 was seen as potentially useful in diffusing heated exchanges by serving as a neutral third party. Several creators likened it to a digital mediator. As P5 described, *"It's kind of like the AI plays the role of a mediator, trying to calm things down… online arguments aren't like two sides just yelling until one gets KO'd. The AI just steps in to mediate, that's it."* Participants suggested that, in cases of fact-based disagreements or misunderstandings, calm and reasonable AI responses might prompt users to reflect or de-escalate. For example, P9 noted, *"I think AI can help in some cases… Like when hate comments come from misinformation or echo chambers, it could step in to fact-check. That might actually reduce some of the arguing,"* suggesting that clarifying facts could reduce tension. However, this approach was seen as fragile. Participants stressed that this only worked when users perceived the AI as fair and trustworthy. Once that trust was compromised, its moderating power diminished.

C2 received the most positive feedback for managing emotionally charged situations. Participants appreciated that it supported users in transforming impulsive reactions into thoughtful replies, especially when hateful comments provoked strong emotions. By preserving user agency while offering AI-generated suggestions, C2 emerged as the most adaptable and dependable option for managing conflict.

On the other hand, C3 raised significant concerns. While the idea of mobilizing bystanders seemed viable in theory, many worried it would escalate rather than ease conflict. Participants questioned the unpredictability of crowd behavior: *"Like, who knows if the people getting the notification will really back you up… Sometimes they just start yelling at me too. Honestly, that probably makes the creator feel even worse"* (P5). Others warned of unintended pile-ons, where *"a mob comes after the person"* (P3, P9), or heightened polarization: *"Everyone thinks they're righteous… A lot of people who* *leave hate comments don't even see them as hate"* (P13). P14 described it as potentially triggering a *"call to arms"* mentality that made resolution more difficult.

Ultimately, these reflections highlight that the perceived effectiveness of AI-driven moderation rests not just on the presence of a response, but on its ability to ease tension and maintain a sense of stability during conflict.

5.1.3 Limitations of AI in contextual hate detection. Participants expressed concerns about AI's ability to accurately identify hate speech, given its inherently subjective nature. Many worried about two key failure modes: the system could be overly sensitive, flagging harmless comments and causing unnecessary conflicts (P11), or too lenient, missing subtle insults. As one participant noted, *"Sometimes they're clearly insulting you, but the AI doesn't catch it"* (P12). These challenges were further complicated by cultural nuances. Participants explained how certain slurs or insults rely on subtle contextual cues that AI systems often miss (P3). Even for clear-cut cases, they questioned whether truly neutral standards were possible, since interpretations vary across different communities (P8).

These concerns became particularly salient during evaluation of C3. Participants doubted the AI could reliably classify the comments and determine appropriate responders to controversial content, as P5 noted that content acceptable to one group might offend another. Some mentioned that imbalanced participation could amplify polarization rather than facilitate constructive dialogue.

5.2 Alignment with Creator Goals

5.2.1 Maintaining autonomy in decision-making. Participants consistently emphasized that effective counterspeech tools must align with their broader goals as content creators, particularly maintaining control over their public presence, protecting their reputations, and communicating with intentionality. While they appreciated that bystander-based approaches like C3 preserved a human presence, many expressed discomfort with AI taking initiative on their behalf, as imagined in C1.

A core concern was the potential loss of control. As P15 described, *"It's like you're handing your fate over to someone else… You have no idea what they're gonna say,"* while P8 noted, *"I don't think I can trust AI 100%… I'm not even sure if it's more objective than people."* Participants feared that automated responses, especially those generated without their involvement, could escalate conflict, misjudge tone, or misrepresent their intent. These concerns also touched on issues of accountability. When AI acts autonomously, participants questioned whether the message would be interpreted as coming from the creator or the platform, raising the risk of reputational harm. *"If it replies the wrong way, it could mess up your account… People should know whether it's AI replying or a real person."* (P4). Even with customization options, some remained skeptical. As P7 explained, *"What AI says might not fully match what I actually want to say."*

In contrast, C2 was appreciated for supporting creators without undermining their autonomy. Participants valued the ability to draft responses with AI assistance while retaining full control over whether and how to engage. As P15 noted, *"I still want to have control, not leave it in the hands of a machine… Even if I use it, maybe I just don't feel like replying, and that should be my choice."* This control was seen as essential to achieving their goals, whether protecting their brand, sustaining their relationships with audiences, or managing emotional energy.

Concerns deepened when automation was perceived as platform-controlled. Some participants expressed skepticism about platforms' motivations, suggesting that automated moderation tools might serve commercial interests rather than creators' needs. As P5 described, *"From the platform's point of view, they have no real reason to go against cyberbullying, because that would hurt engagement. The only time they step in is when something blows up so badly they can't control it anymore."* Others echoed this distrust, questioning whether social platforms would genuinely invest in tools for counterspeech as a public good, given that their business models rely on monetizing engagement and selling exposure for profit. *"I pay so more people see my videos… Now AI is part of that, buying and selling traffic. It's not about helping creators… it's just business. You pay, they deliver, whether the attention is good or bad."* (P10). As P2 concluded, *"Nobody asked you [platform] to offer this service… and you did it anyway. In the end, it just makes me not want to use the platform anymore."*

Taken together, these perspectives underscore that creators want tools that align with their goals, not ones that override their agency or serve opaque platform agendas.

5.2.2 Preserving content creator's voice. As content creators, participants viewed self-expression as central to their role, not just in producing content, but also in how they interact with their audiences. Many expressed concern that AI tools, if not carefully designed, could interfere with this core aspect of their identity. Rather than replacing their voice, they wanted AI to support their expression in ways that remained subtle and flexible.

In discussing C2, participants appreciated AI support for drafting responses but emphasized that it must preserve the creator's authorship and tone. They valued the ability to tailor messages and maintain a personal voice, which they believed was essential for genuine engagement. As P5 related, *"If I have any concern, it's whether I can choose the tone… like if I want to praise someone or clap back, can I decide that? I still want to be the one directing the AI."* P3 warned against overly generic replies: *"Eventually, it might all start sounding the same… That's the problem. It might not feel personal enough. Like, if every comment is just 'good job, good job,' then yeah, something feels off."* Others stressed that suggestions should remain unobtrusive and avoid feeling like instructions or judgments about what they should say. *"If it's just polishing, that's fine… but if AI steps in before I even reply, like flagging comments or warning me, that actually makes me feel worse… It feels like the platform is judging me before I've even said anything."* (P4). This preference extended to interface design: P14 proposed that AI suggestions should not appear inside the input box to avoid overriding the creator's own wording, explaining, *"I don't want other people's words, or what the system wants me to say, showing up in my own comment box."*

More broadly, participants saw replying to comments, especially technical or nuanced ones, as part of their creative expression and community-building. Fully automated tools like C1 were seen as potentially diminishing both their motivation and sense of authorship:*"Here's the thing: creators post online because they have something they want to say… You can't let AI speak for them. That completely goes against the point. The drive to create comes from people wanting AI to shut up, not speak for them… It's not about freedom of speech, it's about respecting the value of human expression."* (P14) One participant also worried that relying too much on AI could lead to dependency and weaken their ability to express themselves—something they were unwilling to trade as creators (P15).

Overall, participants made it clear that AI should empower—not replace—their voice. Preserving authentic authorship, emotional nuance, and intentionality in communication was essential not only for maintaining credibility, but also for sustaining their creative purpose and sense of agency.

5.2.3 AI as a buffer against harm. Participants frequently described these AI-driven counterspeech tools as a form of protection that could shield creators from direct emotional harm and reduce the risk of reactive conflict. Creators, they noted, are uniquely exposed on social platforms and face heightened scrutiny compared to typical users: *"The creator is out in the open, but commenters are hidden… You're the one in the spotlight, the one being consumed"* (P14). Within this dynamic, AI tools were seen as capable of intercepting aggressive comments before they reach the creator directly, either by responding on the creator's behalf or by shifting and reframing the tone of public conversation. By posting empathic or reasonable replies in response to harmful comments, AI could model healthier dialogue among the audience. As P6 suggested, *"I think AI can draw on algorithms and suggest healthier ways to talk… maybe even help people say things others will agree with. It might be better at helping resolve conflicts between people."*

Some participants also viewed this protective function as a way to strategically shift public attention from the individual to the topic, helping reduce personal attacks and ease the emotional burden on creators (P1). For many, this buffering role was not only about immediate relief but also long-term sustainability—helping creators conserve emotional energy and stay focused on content creation (P3).

5.2.4 The cognitive load of counterspeech. While participants acknowledged the value of AI-supported counterspeech, many questioned whether it aligned with their need to manage interactions efficiently and preserve energy for content creation—their primary focus. Some participants found C2 potentially burdensome rather than helpful. As P7 noted, *"Basically, you still have to deal with each comment manually… So in terms of time, it might not actually make our work more efficient."* Compared to simpler moderation actions like reporting or deleting, crafting thoughtful counterspeech—even with AI assistance—demanded greater effort. P3 described the cognitive load involved: *"You have to read what they [audience] give you, take it in, understand it, and figure out the right way to reply."* Others worried that AI-generated replies could initiate ongoing exchanges that creators would feel obligated to continue. *"If AI helps me craft replies, of course there's some expectation… You want to see if it works, if it has an effect… And once you care about the outcome, you're stuck in it. If the other person keeps going, you have to keep creating, right?"* (P13). This potential for prolonged back-and-forth made some participants hesitant, given the emotional and time-related costs.

More broadly, participants worried that overanalyzing hate comments—especially with constant AI monitoring—could become a psychological burden. *"Sometimes AI becomes a burden… Like before, I might just glance at comments and move on. But with the tool, I have to check what the AI filtered or flagged every day… "* (P9). For the sake of mental well-being, some suggested that creators should focus more on genuine, positive interactions rather than harmful ones (P11, P14).

5.2.5 Balancing authenticity and efficiency. Beyond handling online hate, participants emphasized that these AI tools could also shape their relationship with followers, something they deeply valued as creators whose careers rely on audience trust and engagement. Participants saw potential for AI to strengthen these relationships by helping them respond more frequently, especially during periods of high comment volume. A few noted that timely, appropriate replies—even to negative comments—could boost engagement and foster stronger community ties (P4). One participant shared, *"If I receive a hate comment and reply like, 'Yeah, sorry… ' that actually feels powerful… Like, 'wow, I'm not even a fan and they still took the time to respond.' Even if it's AI-assisted and just a quick reply, for the user, it still feels like real attention"* (P3). C1, in particular, appealed to participants managing large followings. Many appreciated its potential to improve communication efficiency by automating replies while staying engaged with followers.

However, deeper discussion revealed a tension between efficiency and authenticity. Participants worried that responses written by AI might feel insincere or overly formal, especially in emotionally sensitive situations. One creator explained, *"I don't know if the AI can match my personal tone or not… It might come off a bit stiff"* (P1). This concern was especially prominent when evaluating C1; several compared it to frustrating experiences with AI-driven customer service (P3, P4). At the heart of these concerns was authenticity. Many stressed that audiences want a real connection. *"When people comment, they're talking to you. They want to hear from you, not something else,"* said P14. P15 described this as a lack of emotional connection: *"The comments are supposed to be people talking to people… so why are we talking to a bot? It's like getting some auto-reply… there's no feeling in it."*

No matter how logical AI responses might be, they lacked emotional presence. Over-relying on AI risked damaging the trust creators worked hard to build and reducing the audience's willingness to engage. As P9 warned, *"If people notice, they might think, 'Wow, even your replies are from AI? What's the point?'... If this becomes common, it could even attract sarcasm, like 'Did AI write this?'"*

In short, participants saw AI as a double-edged sword: helpful for scaling engagement, but only if used with care and transparency. Preserving authenticity and a human touch was essential to maintaining audience trust.

5.3 Impact on Community Dynamics

5.3.1 Counterspeech as an additional layer of moderation. Participants commonly believed that AI-driven counterspeech tools could complement existing platform moderation by addressing moral issues beyond legal violations. As P10 observed, *"Platforms only care about what's legally banned... They use the law as the line, because of free speech. So anything that's morally wrong, they usually ignore... and there's still a lot of that kind of harmful content out there."*

In this context, C1 was seen as a promising addition—offering neutral, timely interventions that could defuse tensions without the severity of punitive measures like bans or content removal. In parallel, participants viewed C2 as a more humanized approach to moderation. By supporting creators in crafting empathetic responses, it enabled gentle interventions that could encourage self-reflection among offenders. As P14 explained, *"gives them space to reflect,"* fostering opportunities for behavioral change and, over time, contributing to a less toxic community dynamic.

5.3.2 Concerns about escalation and manipulation. While participants acknowledged the potential of AI-driven counterspeech tools to support healthier online discourse, they also raised significant concerns regarding the risks of abuse. Without thoughtful design and oversight, these tools could inadvertently exacerbate online harm and disrupt platform dynamics.

A key concern centered on how tools like C2, despite being designed to support empathetic responses, could be exploited to generate harmful or sarcastic counterspeech with minimal cognitive effort. Some participants worried that the ease of response generation might encourage users to escalate minor provocations (P6), explaining, *"You don't even have to think, just copy and paste replies... That could make things spiral even more."* They expressed parallel concerns about C1, particularly the risk that automated counterspeech agents could be repurposed to efficiently carry out large-scale, coordinated antagonistic attacks.

Others emphasized the possibility of users manipulating AI to produce inappropriate tones. As they warned, *"Think about it—are people smarter, or algorithms? They are written by people... If AI replies are public, they'll be tested by thousands of users who might outsmart the system... There's a risk it could get twisted... like, trained by bad actors to sound sarcastic or even toxic"* (P13). Similarly, some participants questioned whether AI could maintain consistent judgment in the face of strategic manipulation. P12 raised questions: *"Can it really think like a human during a conversation? Like, is there a chance it might actually get convinced by the other person?"* These concerns reflected broader skepticism about AI's capacity to uphold normative boundaries in high-conflict environments.

5.3.3 Overstressing toxicity can hinder engagement. Participants agreed that addressing harmful content is important, but cautioned that overemphasizing toxicity—especially through AI-driven counterspeech—could ultimately dull the vibrancy of online communities.

Several noted that not all provocative or critical comments are inherently harmful; playful teasing, cultural humor, and direct honesty are often central to the character of online discourse (P1, P8, P11). However, participants expressed concern that AI, lacking cultural nuance and contextual understanding, may misinterpret such expressions as toxic. P9 warned that overly strict moderation could suppress casual joking and humor, making the online environment feel *"cold and dull." "Our* *language has all kinds of things—sarcasm, humor... That's where the beauty of language comes from. You can't just label things as good or bad, black or white. That's the end of art, the end of humanity."* (P11)

Beyond misclassification, participants also raised concerns that excessive AI intervention might discourage authentic participation, leading to sanitized, overly positive exchanges. As P1 remarked, *"You end up only seeing the nice comments, but miss the ones that are actually helpful."*

These concerns extended to C3, where participants questioned whether users invited to counterspeak would be willing—or feel comfortable—engaging in argument. P7 pointed out the challenge of accurately identifying users who are both relevant and motivated to respond, while P4 emphasized that frequent exposure to conflict may lead to disengagement: *"It could have some negative effects... Some people just don't want to be dragged in. If there's always drama, they might avoid getting involved at all."*

Altogether, participants suggested that overstressing toxicity—particularly without sensitivity to context—risks undermining the openness, spontaneity, and authenticity that make online communities meaningful and engaging in the first place.

6 Discussion

6.1 Integrating Creators' Perspectives

Creators, as influential figures in digital spaces, face heightened exposure to toxic comments [83], making their experiences and moderation goals distinct from those of typical users. By examining the perspectives of 15 creators across various platforms, experience levels, and audience sizes, our study fills a gap in counterspeech research, which often overlooks the voices of creators as key stakeholders.

The objectives creators described for engaging in counterspeech align with those identified in prior work: to promote open dialogue, reduce prejudice, or clarify their stance to influence public perception and prevent misunderstandings [9, 66]. Some creators shared examples of successfully changing a commenter's mind by appealing to empathy, echoing findings from existing research on empathy-based strategies [8]. However, they also emphasized unique motivations tied to their public roles, particularly, safeguarding their reputation and maintaining audience trust.

Our findings reveal that creators avoid counterspeech not only due to time constraints but also because of their shifting identity as public figures. While smaller creators often behave like typical users, those with larger followings grow increasingly cautious. They described avoiding direct responses to minimize reputational risk, citing peers who were repeatedly targeted for old comments. Some feared their replies would be overanalyzed or taken out of context. As creators gained influence, many saw themselves less as individuals and more as facilitators, responsible for maintaining a welcoming environment. This shift led them to refrain from engaging even when bystanders defended them, as they feared "taking sides" might escalate conflict. These concerns are supported by prior research showing that individuals with perceived authority are more likely to influence others' behavior [79]. This dynamic may also help explain why volunteer counterspeakers often feel their efforts go unnoticed when the creator remains silent [54, 72].

We found that creators also adapted their engagement strategies based on platform norms. On platforms where heated discussions are penalized, they avoided counterspeech to protect their reach; on others that reward engagement, some even leveraged controversy to attract attention. Another common reason for avoiding counterspeech was the desire to prioritize positive interactions and reduce the visibility of hate. Across the board, creators highlighted the emotional burden of counterspeech. Crafting thoughtful responses is time-consuming, often ineffective, and may lead to prolonged conflict, echoed prior research [24, 72]. Without support from the platform or their audience, creators viewed counterspeech as emotionally taxing—if not invasive—particularly when weighed against their broader goals of content creation and community building.

Understanding creators' motivations and challenges covering both personal and professional aspects can inform future research and practices in counterspeech. Implementing counterspeech without considering creators' perspectives risks escalating conflict and undermining their goals, making such interventions feel intrusive rather than supportive.

6.2 AI-Mediated Counterspeech and Design Implications

In this section, we highlight how content creators assess AI-driven counterspeech tools through the lens of their role as communicators and public-facing figures. Creators valued AI's potential to improve the quality and efficiency of responses to online hate, but they also expressed strong concerns about preserving authenticity, agency, and emotional well-being. They emphasized the importance of maintaining control over their voice, ensuring transparency in AI involvement, and minimizing unintended harms to themselves and their communities.

To situate these findings within the broader CSCW community, we structure the discussion in two parts. First, we engage with the literature on AI-Mediated Communication (AI-MC) to explore how creators perceive AI's role in shaping social interaction across four dimensions: self-presentation, human agency, transparency and trust, and mitigating online harm with AI assistance. Second, we draw on research in content moderation to offer design considerations for future tools, focusing on complementing platform moderation, contextual and cultural nuance, and avoiding overmoderation.

6.2.1 AI-Mediated Communication (AI-MC) in the context of counterspeech.

Augmenting, not replacing, creator agency. Our study revealed a consistent preference among creators for counterspeech approaches that augment human agency rather than replace it with autonomous AI responses. This preference stems from the belief that effective counterspeech relies on authentic interpersonal connection, something AI alone cannot provide. This aligns with prior work showing that users seek genuine human communication when navigating online conflict [29].

However, creators expressed concern about how followers might interpret such AI-augmented communication. If audiences became aware of AI involvement, creators feared it could appear insincere, undermining the authentic human connection that followers value. This echoes with the claim that when users suspected AI authorship, perceived trustworthiness declined [39].

In the context of AI-mediated counterspeech, we found that audiences may see AI involvement as augmenting, rather than diminishing, the sender's agency, potentially assigning greater responsibility to creators. This finding contributes to ongoing discussions in AI-Mediated Communication about how agency is perceived when AI assists in interpersonal messaging [32]. However, these interpretations are based on creators' own assumptions, shaped by their professional identities and reputational concerns. Future research should examine audience perspectives directly to understand whether AI assistance in creator–audience interactions undermines or reinforces trust.

In addition to interpersonal concerns, creators also highlighted the need to preserve their creative voice and authorship. This expectation shaped their vision for AI-MC tools: suggestions should be helpful but non-intrusive, and never dictate how they respond. If AI oversteps by pushing fixed templates or overly assertive phrasing, creators feared it would erode their sense of control and long-term creative identity due to potential overreliance. These concerns are validated by prior research showing that AI systems can influence how people write and think. For example, one study found that Google Smart Reply systems shape users' linguistic behavior [37], while another demonstrated that opinionated AI outputs can steer user expression and beliefs [40]. In the context of content creation, such influence is particularly sensitive because creators depend on their distinct tone and presence to sustain their audience and brand. This desire to preserve authorship also reflects creators' broader concern with managing their public image, a theme we explore further in the next section.

Optimized self-presentation. Several participants noted that AI could help scale up audience engagement, allowing them to acknowledge more comments and show attentiveness to a wider following. This was seen as a way to strengthen relationships and convert casual viewers into followers, aligning with Tong and Walther's finding that visible communicative effort supports intimacy and relationship maintenance [85].

Content creators' careers rely heavily on strategic self-presentation, carefully shaping how they are perceived by their audience. In our study, participants evaluated AI-driven counterspeech tools primarily through the lens of how such tools might affect their image. Most viewed AI assistance as acceptable and even beneficial. Some saw it as a way to improve response quality and tone, optimizing their public persona. Others valued AI's role as an emotional buffer, helping them avoid impulsive replies that could damage their reputation. In both cases, AI was seen as supporting, not undermining, their career goals.

Participants also stressed the importance of language content and stylistic nuance in conveying intent and identity, verbal cues that people carefully rely on to manage impressions and build relationships in computer-mediated settings [85]. They felt that AI-assisted responses should allow for variation in tone and expression, enabling them to maintain their own communication style. One key concern was that AI might misrepresent them, either through tone mismatch or biased outputs, potentially offending others or causing unnecessary harm.

Transparency, manipulation, and trust. While participants did not explicitly state whether they would disclose the use of AI in crafting responses, many acknowledged already using AI to assist with content creation. In evaluating AI-assisted counterspeech tools, they appreciated features like tone customization and the ability to train AI to match their voice, suggesting that full transparency about AI involvement may not be preferred. Prior work shows that strategic self-disclosure can foster trust, such as sharing professional experiences [44] or selectively revealing personal details to build credibility [13]. Yet, how disclosing AI use affects trust in content creators remains underexplored.

Participants expressed concern that revealing AI involvement might undermine trust. They feared that audiences, once aware of AI's role, might question the authenticity of all responses, even those written personally. In other words, creators worried that AI-mediated communication could weaken the reliability of social signals [23], casting doubt on both the message and the messenger. Though speculative, this concern may discourage creators from disclosing their use of AI tools.

At the same time, creators demanded transparency when AI agents operate autonomously in public spaces. One reason was accountability. If AI-generated messages offend or misfire, audiences may assume the creator is responsible. Another concern was manipulation: participants suspected platforms might deploy AI under the guise of moderation to shape discourse or suppress dissent. This aligns with findings that the boundary between persuasion and manipulation in AI systems is often blurred and context-dependent [32]. While creators welcomed independent AI agents as a mediator when clearly framed, they warned that unclear attribution or platform misuse could erode trust and compromise their voice.

Mitigating online harm with AI assistance. Our findings suggest that creators see potential in AI agents to mitigate online harm by stepping into conflicts with reasonable responses. Participants believed that such interventions could prompt reflection, particularly in factual disputes, where AI might serve a fact-checking role. This aligns with the design intent behind tools like Mediation-Bot [61], which uses structured dialogue to support conflict resolution, and highlights a shared interest in using AI to facilitate civil discourse. It also resonates with Auntie Meiyu [49], a chatbot that counters misinformation in private chats. While effective in providing factual corrections, its detection errors often undermined user trust. Similarly, our participants noted that fact-checking responses are only effective when the source is perceived as trustworthy, highlighting the need for counterspeech tools that maintain both accuracy and credibility.

Beyond reactive counterspeech, participants also imagined AI-assisted writing tools nudging users toward more polite or less biased language during comment drafting. They speculated this could reduce harmful content at the source. This aligns with findings by Levy and Barocas [51], who observed that AI-mediated cues of trustworthiness can counteract interpersonal bias, and echoes the suggestion by Naaman et al. [32] that AI-Mediated Communication (AI-MC) might enable pragmatic interpersonal adaptation in social interactions. Still, creators acknowledged these are assumptions, not proven effects. While AI may offer scalable solutions for harm reduction, its actual impact requires further study.

6.2.2 *Design considerations for AI-driven counterspeech tools.*

Complementing platform moderation with AI-assisted counterspeech. Our findings suggest that creators favor a moderation strategy that integrates both proactive and reactive approaches, rather than relying on one type alone. While participants appreciated counterspeech as a more socially intelligent and human-centered response to online harm, they also expressed concern about treating it as a substitute for proactive moderation tools such as filtering, nudging, or pre-posting comment warnings. These tools were viewed as essential for preventing harm, especially when dealing with high comment volume or emotionally charged interactions.

Participants described AI-assisted counterspeech as a meaningful complement to existing platform moderation. It was seen as less punitive than automated enforcement, yet more intentional than ignoring harmful content. Creators felt it was particularly effective for addressing moral misconduct that falls outside of formal policy violations. This perspective contributes to and extends Seering et al.'s concept of layered moderation, which calls for a combination of automated systems, community engagement, and interpersonal strategies to manage harm more effectively [78].

At the same time, creators raised concerns about the emotional labor and time-related burden of engaging in counterspeech. Even with AI assistance, they noted that responding to harm can be draining and may reduce their energy for content creation. These concerns build on prior work on invisible labor in content moderation [70], highlighting how such burdens extend to public-facing users. By centering creators' experiences, this study underscores the need to design moderation systems that consider both scalability and the emotional cost of participation.

Toward contextually and culturally adaptive AI for moderation. Our findings highlight creators' concerns that current AI moderation systems lack sensitivity to context and culture, both in detecting harmful content and generating appropriate responses. Participants pointed out that hate often appears in subtle forms, such as sarcasm, indirect phrasing, or culturally specific references, which are difficult for general-purpose AI models to detect accurately. When moderation thresholds are too rigid, they risk over-policing benign content or failing to catch harmful expression, undermining both creator trust and community norms. This echoes prior work on moderation misalignment and detection failure in nuanced contexts [42].

Beyond detection, participants emphasized that counterspeech generation must also reflect cultural communication styles. Several creators noted that Eastern communication tends to be more indirect than Western norms, suggesting that counterspeech designed with a Western rhetorical style may be perceived as confrontational or inappropriate in other contexts. This reflects long-standing distinctions in high- and low-context communication [31], where high-context cultures (like many East Asian cultures) rely more on implicit cues and shared understanding, while low-context cultures (such as the U.S.) favor direct, explicit communication. These differences highlight the need for AI-generated responses to adapt to social expectations across cultures.

Altogether, our findings point to the need for moderation systems that are not only accurate, but socially and culturally intelligent, supporting both detection and intervention in ways that align with diverse communicative norms.

Risks of over-moderation and the loss of community vitality. Our findings show that creators are concerned about how over-moderation can suppress authentic interaction and weaken community dynamics. Participants noted that humor, sarcasm, and dissent may be misread by AI systems, leading to overly sanitized spaces that lack spontaneity. As one creator put it, platforms can become "too clean to be interesting." This aligns with research showing that heavy-handed moderation may discourage participation and disrupt social cohesion [24, 78].

Creators also cautioned against the overuse of counterspeech. While they saw it as a useful, non-punitive strategy, many worried that repeated or poorly contextualized interventions could backfire. One major concern was that counterspeech might surface comments that otherwise would have been ignored, giving harmful content greater visibility. This, in turn, could create a combative atmosphere, discourage positive contributions, and reduce user willingness to engage.

Our study suggests that moderation should not equate to constant correction. Designers must consider how and when to respond to harm without overwhelming the comment space. AI tools should not only support counterspeech but also recognize when inaction, or subtle intervention, may better preserve the tone and vitality of online communities.

6.3 Limitations

While this study offers valuable insights into creators' perspectives on counterspeech and AI-powered moderation tools, several limitations should be acknowledged. First, the sample size of 15 participants—though varied in platform use and experience—may not fully capture the breadth of creators' experiences across different social media environments. All participants were also recruited from Chinese platforms, which may limit the transferability of findings to other cultural contexts.

Additionally, our study focused solely on creators as the senders of counterspeech, without incorporating perspectives from those on the receiving end. Understanding how audiences interpret and respond to creator- or AI-generated counterspeech is equally important, and future research should explore this dynamic.

Finally, the concept testing was based on hypothetical tools, and participants engaged with speculative scenarios rather than functioning AI systems. Their responses might differ in real-world contexts involving live interactions. Future work could involve longitudinal studies with deployed tools, as well as cross-cultural comparisons, to provide a more holistic view of counterspeech in practice.

7 Conclusion

This study highlights the unique challenges creators face in responding to online hate and how these challenges shape their expectations for AI-mediated counterspeech tools. Unlike typical users, creators must navigate public scrutiny, manage audience relationships, and maintain a consistent personal brand, factors that influence how they interpret and use counterspeech. Our findings show that creators prefer AI tools that augment their voice, support strategic self-presentation, and respect the emotional and cultural dimensions of online interaction. By centering creators' perspectives, this research offers design implications for building AI moderation systems that are not only effective but also context-sensitive and creator-centered, ultimately contributing to more supportive and sustainable digital communities.

References

[1] N. Aggrawal. 2018. Detection of Offensive Tweets: A Comparative Study. *Computer Reviews Journal* 1, 1 (2018), 75–89. https://www.purkh.com/abstract/detection-of-offensive-tweets-a-comparative-study-61798.html

[2] H. Ai-zhong and Y. Zhang. 2022. Ai-powered touch points in the customer journey: a systematic literature review and research agenda. *Journal of Research in Interactive Marketing* 17 (2022), 620–639. Issue 4. doi:10.1108/jrim-03-2022-0082

[3] A. Alrayes, T. F. Henari, and D. A. Ahmed. 2024. Chatgpt in education – understanding the bahraini academics perspective. *Electronic Journal of E-Learning* 22 (2024), 112–134. Issue 2. doi:10.34190/ejel.22.2.3250

[4] A. Ancell. 2022. Corporate counterspeech. *Ethical Theory and Moral Practice* 26 (2022), 611–625. Issue 4. doi:10.1007/s10677-022-10332-6

[5] Dennys Antonialli. 2019. Drag Queen vs. David Duke: Whose Tweets Are More 'Toxic'? https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets/ Accessed: 2024-10-28.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. 2017. deep learning for hate speech detection in tweets. (2017), 759–760. doi:10.1145/3041021.3054223

[7] A. Banchik. 2020. disappearing acts: content moderation and emergent practices to preserve at-risk human rights–related content. *New Media Society* 23 (2020), 1527–1544. Issue 6. doi:10.1177/1461444820912724

[8] S. Benesch, H. M. S. Saleem, K. Dillon, L. W. Wright, and D. R. Ruths. 2016. Counterspeech on twitter: a field study. (2016). doi:10.15868/socialsector.34066

[9] H. K. Bhargava. 2022. The creator economy: managing ecosystem supply, revenue sharing, and platform design. *Management Science* 68 (2022), 5233–5251. Issue 7. doi:10.1287/mnsc.2021.4126

[10] M. Bilewicz and W. Soral. 2020. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *political Psychology* 41 (2020), 3–33. Issue S1. doi:10.1111/pops.12670

[11] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe. 2017. Classification and its consequences for online harassment. *Proceedings of the Acm on Human-Computer Interaction* 1 (2017), 1–19. Issue CSCW. doi:10.1145/3134659

[12] Block Together. 2019. A web app intended to help cope with harassment and abuse on Twitter. https://blocktogether.org/. Accessed: 2019.

[13] Ross Bonifacio, Jirassaya Uttarapong, Rae Jereza, and Donghee Yvette Wohn. 2025. Self-Promotion Practices and Context Collapse Management of Adult Content Creators on OnlyFans. *Proceedings of the ACM on Human-Computer Interaction* 9, GROUP (2025), 1–21. doi:10.1145/3701206

[14] C. Buerger. 2021. iamhere: collective counterspeech and the quest to improve online discourse. *Social Media + Society* 7 (2021). Issue 4. doi:10.1177/20563051211063843

[15] M. Celuch, R. Latikka, R. Oksa, and A. Oksanen. 2023. online harassment and hate among media professionals: reactions to one's own and others' victimization. *Journalism Mass Communication Quarterly* 100 (2023), 619–645. Issue 3. doi:10.1177/10776990221148987

[16] B. Cepollaro, M. Lepoutre, and R. M. Simpson. 2022. Counterspeech. *Philosophy Compass* 18 (2022). Issue 1. doi:10.1111/phc3.12890

[17] Mohit Chandra. 2021. *Towards A More Holistic Approach On Online Abuse and Antisemitism.* Master's thesis. International Institute of Information Technology, Hyderabad, Hyderabad, India. Advisor(s) Ponnurangam Kumaraguru and Manish Shrivastava. https://web2py.iiit.ac.in/research_centres/publications/download/mastersthesis.pdf.a06899c2e49b7fc7.5468657369735f4d6f6869745f66696e616c5f636f70792e706466.pdf Report no: IIIT/TH/2021/5.

[18] A. R. Chrismanto, A. Afiahayati, Y. Sari, A. K. Sari, and Y. Suyanto. 2022. Spam comments detection on instagram using machine learning and deep learning methods. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi* 13 (2022), 46. Issue 1. doi:10.24843/lkjiti.2022.v13.i01.p05

[19] A. Christin and R. Lewis. 2021. the drama of metrics: status, spectacle, and resistance among youtube drama creators. *Social Media + Society* 7 (2021). Issue 1. doi:10.1177/2056305121999660

[20] Data Society. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/ Accessed: 2024-10-28.

[21] N. Deshpande, N. Farris, and V. Kumar. 2022. highly generalizable models for multilingual hate speech detection. (2022). doi:10.48550/arxiv.2201.11294

[22] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. 2018. measuring and mitigating unintended bias in text classification. (2018). doi:10.1145/3278721.3278729

[23] Judith Donath. 2007. Signals in Social Supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251. doi:10.1111/j.1083-6101.2007.00394.x

[24] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300372

[25] A. Essamri, S. McKechnie, and H. Winklhofer. 2019. Co-creating corporate brand identity with online brand communities: a managerial perspective. *Journal of Business Research* 96 (2019), 366–375. doi:10.1016/j.jbusres.2018.07.015

[26] D. Frieß, M. Ziegele, and D. Heinbach. 2020. Collective civic moderation for deliberation? exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication* 38 (2020), 624–646. Issue 5. doi:10.1080/10584609.2020.1830322

[27] C. Fumagalli. 2020. Counterspeech and ordinary citizens: how? when? *Political Theory* 49 (2020), 1021–1047. Issue 6. doi:10.1177/0090591720984724

[28] H. Ghadirian. 2016. Peer moderation of asynchronous online discussions: an exploratory study of peer e-moderating behaviour. *Australasian Journal of Educational Technology* (2016). doi:10.14742/ajet.2882

[29] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, London.

[30] P. Goffredo, V. Basile, B. Cepollaro, and V. Patti. 2022. Counter-twit: an italian corpus for online counterspeech in ecological contexts. *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (2022), 57–66. doi:10.18653/v1/2022.woah-1.6

[31] Edward T. Hall. 1981. *Beyond Culture* (reprint ed.). Anchor Books, Garden City, NY.

[32] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100. doi:10.1093/jcmc/zmz022

[33] Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences* 118, 50 (2021), e2116310118. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2116310118 doi:10.1073/pnas.2116310118

[34] S. Hassan and M. Alikhani. 2023. Discgen: a framework for discourse-informed counterspeech generation. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacifi* (2023). doi:10.18653/v1/2023.ijcnlp-main.28

[35] Hatebase. 2019. The world's largest structured repository of regionalized, multilingual hate speech. https://hatebase.org/. Accessed: 2024.

[36] M. Hietanen and J. Eddebo. 2022. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry* 47 (2022), 440–458. Issue 4. doi:10.1177/01968599221124309

[37] Jess Hohenstein and Malte F. Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190. doi:10.1016/j.chb.2019.106190

[38] Y. Hua, M. Ribeiro, R. West, T. Ristenpart, and . . 2022. characterizing alternative monetization strategies on youtube. *Proceedings of the Acm on Human-Computer Interaction* 6 (2022), 1–30. Issue CSCW2. doi:10.1145/3555174

[39] Maurice Jakesch, Melissa French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300469

[40] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America* 120, 11 (2023), e2208839120. doi:10.1073/pnas.2208839120

[41] G. Jd and J. Bright. 2021. hate contagion: measuring the spread and trajectory of hate on social media. (2021). doi:10.31234/osf.io/b9qhd

[42] S. Jhaver. 2023. personalizing content moderation on social media: user perspectives on moderation choices, interface design, and labor. *Proceedings of the Acm on Human-Computer Interaction* 7 (2023), 1–33. Issue CSCW2. doi:10.1145/3610080

[43] S. Kopf. 2020. "rewarding good creators": corporate social media discourse on monetization schemes for content creators. *social media + Society* 6 (2020). Issue 4. doi:10.1177/2056305120969877

[44] Yubo Kou, Bryan Semaan, and Bonnie Nardi. 2017. A Confucian Look at Internet Censorship in China. In *Human-Computer Interaction – INTERACT 2017*. Springer, 377–398. doi:10.1007/978-3-319-67744-6_25

[45] R. Kowalski, A. Toth, and M. Morgan. 2017. bullying and cyberbullying in adulthood and the workplace. *the journal of social psychology* 158 (2017), 64–81. Issue 1. doi:10.1080/00224545.2017.1302402

[46] E. Kulenović. 2022. should democracies ban hate speech? hate speech laws and counterspeech. *Ethical Theory and Moral Practice* 26 (2022), 511–532. Issue 4. doi:10.1007/s10677-022-10336-2

[47] S. Laaksonen, J. Haapoja, T. Kinnunen, M. Nelimarkka, and R. Pöyhtäri. 2020. the datafication of hate: expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data* 3 (2020). doi:10.3389/fdata.2020.00003

[48] S. Lee, Y. Chang, O. D. Lee, S. Ryu, and Q. Yin. 2024. Exploring online social platform affordances for digital creators: a multi-method approach using qualitative and configurational analysis. *Industrial Management Amp; Data Systems* 124 (2024), 1501–1530. Issue 4. doi:10.1108/imds-12-2023-0951

[49] Tuan-He Lee and Susan R. Fussell. 2025. Countering Misinformation in Private Messaging Groups: Insights From a Fact-checking Chatbot. *Proceedings of the ACM on Human-Computer Interaction* 9, GROUP (2025), 1–30. doi:10.1145/3701189 Article GROUP10, Publication date: January 2025.

[50] P. Leerssen. 2022. an end to shadow banning? transparency rights in the digital services act between content moderation and curation. (2022). doi:10.31219/osf.io/7jg45

[51] Karen Levy and Solon Barocas. 2018. Refractive Surveillance: Monitoring Customers to Manage Workers. *International Journal of Communication* 12 (2018), 1166–1188. https://ijoc.org/index.php/ijoc/article/view/7041

[52] R. Lewis, A. Marwick, and W. Partin. 2020. "we dissect stupidity and respond to it": response videos and networked harassment on youtube. (2020). doi:10.31235/osf.io/veqyj

[53] Taylor Lorenz. 2021. For Creators, Everything Is for Sale. *The New York Times* (March 10 2021). https://www.nytimes.com/2021/03/10/style/creators-selling-selves.html Updated March 11, 2021.

[54] Kaitlin Mahar, David Karger, and Amy X. Zhang. 2018. Squadbox: A Tool To Combat Online Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. doi:10.1145/3173574.3174190

[55] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. 2019. spread of hate speech in online social media. (2019). doi:10.1145/3292522.3326034

[56] C. Meisner. 2023. the weaponization of platform governance: mass reporting and algorithmic punishments in the creator economy. *Policy Internet* 15 (2023), 466–477. Issue 4. doi:10.1002/poi3.359

[57] E. Mohamed. 2024. the impact of artificial intelligence on social media content. *journal of social sciences* 20 (2024), 12–16. Issue 1. doi:10.3844/jssp.2024.12.16

[58] M. Mudambi. 2024. fighting misinformation on social media: an empirical investigation of the impact of prominence reduction policies. *Production and Operations Management* (2024). doi:10.1177/10591478241283839

[59] J. Mun. 2024. counterspeakers' perspectives: unveiling barriers and ai needs in the fight against online hate. 91 (2024), 1–22. doi:10.1145/3613904.3642025

[60] Joanne Neale. 2016. Iterative categorization (IC): a systematic technique for analysing qualitative data. *Addiction* 111, 6 (Feb. 2016), 1096–1106. doi:10.1111/ADD.13314

[61] Kavous Salehzadeh Niksirat, Diana Korka, Hamza Harkous, Kévin Huguenin, and Mauro Cherubini. 2023. On the Potential of Mediation Chatbots for Mitigating Multiparty Privacy Conflicts - A Wizard-of-Oz Study. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 142:1–142:33. doi:10.1145/3579618 Article 142, Publication date: April 2023.

[62] M. Obermaier, U. Schmid, and D. Rieger. 2023. Too civil to care? how online hate speech against different social groups affects bystander intervention. *European Journal of Criminology* 20 (2023), 817–833. Issue 3. doi:10.1177/14773708231156328

[63] M. Ozanne, A. Bhandari, N. Bazarova, and D. DiFranzo. 2022. shall ai moderators be made visible? perception of accountability and trust in moderation systems on social media platforms. *Big Data Society* 9 (2022). Issue 2. doi:10.1177/20539517221115666

[64] H. Park. 2023. uncovering the root of hate speech: a dataset for identifying hate instigating speech. (2023). doi:10.18653/v1/2023.findings-emnlp.412

[65] J. Pater, M. Kim, E. Mynatt, and C. Fiesler. 2016. characterizations of online harassment. (2016). doi:10.1145/2957276.2957297

[66] Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia Rho. 2024. Behind the Counter: Exploring the Motivations and Barriers of Online Counterspeech Writing. arXiv:2403.17116 [cs.HC] https://arxiv.org/abs/2403.17116

[67] D. Rahmawan, J. Mahameruaji, and R. Anisa. 2019. YouTube channel "kokbisa" as platform for science and environmental communication. In *Proceedings of the 1st Workshop on Environmental Science, Society, and Technology, WESTECH 2018, December 8th*. doi:10.4108/eai.8-12-2018.2283930

[68] D. K. H Rahmi. 2024. empathy and hate speech in social media: the case of indonesia. *international journal of social science and human Research* 07 (2024). Issue 03. doi:10.47191/ijsshr/v7-i03-29

[69] B. Rieder, E. Borra, Ò. Coromina, and A. Matamoros-Fernández. 2023. making a living in the creator economy: a large-scale study of linking on youtube. *Social Media + Society* 9 (2023). Issue 2. doi:10.1177/20563051231180628

[70] Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven. doi:10.2307/j.ctvhrcz0v

[71] L. Rouse and A. Salter. 2021. cosplay on demand? instagram, onlyfans, and the gendered fantrepreneur. *Social Media + Society* 7 (2021). Issue 3. doi:10.1177/20563051211042397

[72] K. Rudnicki, H. Vandebosch, P. Voué, and K. Poels. 2022. systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour and Information Technology* 42 (2022), 527–544. Issue 5. doi:10.1080/0144929x.2022.2027013

[73] Stuart J. Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach* (3 ed.). Pearson, Upper Saddle River, NJ.

[74] Goldman Sachs. 2023. The creator economy could approach half-a-trillion dollars by 2027. https://www.goldmansachs.com/intelligence/pages/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html. Retrieved October 1, 2024.

[75] P. Saha, K. Singh, A. Kumar, B. Mathew, and A. Mukherjee. 2022. Countergedi: a controllable approach to generate polite, detoxified and emotional counterspeech. (2022). doi:10.48550/arxiv.2205.04304

[76] David E. M. Sappington. 1991. Incentives in Principal-Agent Relationships. *Journal of Economic Perspectives* 5, 2 (1991), 45–66.

[77] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). ACL, Valencia, Spain, 1–10. doi:10.18653/v1/W17-1101

[78] J. Seering, G. Kaufman, and S. Chancellor. 2020. Metaphors in moderation. *New Media Amp; Society* 24 (2020), 621–640. Issue 3. doi:10.1177/1461444820964968

[79] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 111–125. doi:10.1145/2998181.2998277

[80] H. M. Stegeman, C. Are, and T. Poell. 2024. Strategic invisibility: how creators manage the risks and constraints of online hyper(in)visibility. *Social Media + Society* 10 (2024). Issue 2. doi:10.1177/20563051241244674

[81] P. Tewari and S. Mehendale. 2023. Persisting misogyny: a gendered analysis of online harassment of indian content creators on instagram and its impact on mental health. *Cardiometry* (2023), 493–501. Issue 25. doi:10.18137/cardiometry.2022.25.493501

[82] H. Thach, S. Mayworm, D. Delmonaco, and O. L. Haimson. 2022. (in)visible moderation: a digital ethnography of marginalized users and content moderation on twitch and reddit. *New Media Amp; Society* 26 (2022), 4034–4055. Issue 7. doi:10.1177/14614448221109804

[83] K. Thomas, P. Kelley, S. Consolvo, P. Samermit, and E. Bursztein. 2022. "it's common and a part of being a content creator": understanding how creators experience and cope with hate and harassment online. (2022). doi:10.1145/3491102.3501879

[84] A. Tobi. 2024. towards an epistemic compass for online content moderation. *Philosophy Technology* 37 (2024). Issue 3. doi:10.1007/s13347-024-00791-3

[85] Stephanie T. Tong and Joseph B. Walther. 2011. Just Say "No Thanks": Romantic Rejection in Computer-Mediated Communication. *Journal of Social and Personal Relationships* 28, 4 (2011), 488–506. doi:10.1177/0265407510384895

[86] Twitter. 2020. Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hate. Retrieved September 9, 2020.

[87] J. Uyheng, D. Bellutta, and K. M. Carley. 2022. Bots amplify and redirect hate speech in online discourse about racism during the covid-19 pandemic. *Social Media + Society* 8 (2022). Issue 3. doi:10.1177/20563051221104749

[88] Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. Intent-Aware and Hate-Mitigating Counterspeech Generation via Dual-Discriminator Guided LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 9131–9142. https://aclanthology.org/2024.lrec-main.800

[89] S. M. West. 2018. Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media Amp; Society* 20 (2018), 4366–4383. Issue 11. doi:10.1177/1461444818773059

[90] S. Whitten. 2023. A republican conception of counterspeech. *Ethical Theory and Moral Practice* 26 (2023), 555–575. Issue 4. doi:10.1007/s10677-023-10409-w

[91] K. Wise, B. Hamman, and K. Thorson. 2006. Moderation, response rate, and message interactivity: features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication* 12 (2006), 24–41. Issue 1. doi:10.1111/j.1083-6101.2006.00313.x

[92] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300390

[93] YouTube. 2020. Hate speech policy. https://support.google.com/youtube/answer/2801939?hl=en. Retrieved September 9, 2020.

[94] V. Zhezha, B. Kola, and A. M. Melinceanu. 2023. Exploring the landscape of digital marketing in albania: insights from local companies. *Academic Journal of Interdisciplinary Studies* 12 (2023), 341. Issue 4. doi:10.36941/ajis-2023-0120

[95] W. Zhu and S. Bhat. 2021. Generate, prune, select: a pipeline for counterspeech generation against online hate speech. (2021). doi:10.48550/arxiv.2106.01625

[96] A. Álvarez and F. Winter. 2018. normative change and culture of hate: an experiment in online environments. *European Sociological Review* 34 (2018), 223–237. Issue 3. doi:10.1093/esr/jcy005

A Concept Test Script

This appendix includes the script used during the concept testing portion of our study. The original script was written in Chinese and later translated into English by one of our bilingual authors to ensure accuracy and contextual relevance.

Script

Next, we'll move into a concept testing section. We've designed three AI-powered counterspeech tools and would like to hear your thoughts on them. Please imagine that one of your usual posts has received a hateful comment and how you might feel and respond.

Scenario 1. Imagine you receive a harmful comment on one of your regular posts. In this concept, an AI steps in as a third-party participant in the comment section. It joins the conversation on its own, like another user, after detecting the harmful comment, or when someone invites it. The AI might post a message to encourage empathy, help others understand each other's views, or flag the comment as inappropriate. It acts like a chatbot trained in counterspeech strategies, aiming to calm things down and support healthier dialogue in your comment section.

Follow-up question: What do you like or dislike about this concept?

Scenario 2. Imagine you receive a harmful comment on one of your usual posts. This tool suggests replies or helps you revise your own, aiming to improve your response for counterspeech purposes. It's a co-writing assistant that supports you in crafting replies using different counterspeech strategies.

Follow-up question: What do you like or dislike about this concept?

Scenario 3. Imagine you receive a harmful comment on one of your usual posts. In this concept, the AI detects the hurtful comment and sends a prompt to your followers or other active users, inviting them to help you respond. Those who are notified can choose whether or not to reply, offering support or counterspeech voluntarily.

Follow-up question: What do you like or dislike about this concept?

Post-test Reflective Questions

- What do you think about AI participating in the comment section to address harmful comments?
- Do you have any concerns about using AI for counterspeech?
- How do you think these AI tools might affect your experience as a content creator?
- Do you think using AI for counterspeech could lead to more constructive online discussions and reduce online hate?
- Do you have any other questions or anything else you'd like to share with us?

Received October 2024; revised April 2025; accepted August 2025