Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate

Jimin Mun
jmun@andrew.cmu.edu
Language Technologies Institute,
Carnegie Mellon University
Pittsburgh, PA, USA

Cathy Buerger*
Dangerous Speech Project
Washington, D.C., USA

Jenny T. Liang*
Human-Computer Interaction
Institute Carnegie Mellon University
Pittsburgh, PA, USA

Joshua Garland
Arizona State University
Tempe, AZ, USA

Maarten Sap
Language Technologies Institute,
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Counterspeech, i.e., direct responses against hate speech, has become an important tool to address the increasing amount of hate online while avoiding censorship. Although AI has been proposed to help scale up counterspeech efforts, this raises questions of how exactly AI could assist in this process, since counterspeech is a deeply empathetic and agentic process for those involved. In this work, we aim to answer this question, by conducting in-depth interviews with 10 extensively experienced counterspeakers and a large scale public survey with 342 everyday social media users. In participant responses, we identified four main types of barriers and AI needs related to resources, training, impact, and personal harms. However, our results also revealed overarching concerns of authenticity, agency, and functionality in using AI tools for counterspeech. To conclude, we discuss considerations for designing AI assistants that lower counterspeaking barriers without jeopardizing its meaning and purpose.

ACM Reference Format:
Jimin Mun, Cathy Buerger, Jenny T. Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate . In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3613904.3642025

1 INTRODUCTION

Counterspeech, i.e., direct responses to mitigate the harms of hateful or dangerous speech [5], has emerged as a more positive, community-oriented alternative to deletion-based content moderation that avoids censorship concerns [96]. Its goals are multifaceted; counterspeech aims to not only minimize harms of hateful speech, but also to promote positive changes in online communities through open dialogue among users [14] and by fostering a sense of community [12]. Furthermore, how counterspeech is done can be highly varied and complex, as it can range from individual replies to a hateful post to coordinated mass responses via organized groups and hashtags (e.g., #iamhere, #BlackLivesMatter, #stopasianhate) [12, 42, 43, 145, 150]. As hate speech prevalence grows in online spaces [3, 83, 107], coordinating and responding with counterspeech has become increasingly challenging [26, 27].

AI has recently emerged as a potential tool or solution [2, 34, 149] to assist with this increased demand for counterspeech. However, designing an AI system for counterspeech is a unique challenge that requires an understanding of larger context of its impact on those who do it [8]. On one hand, it is important to reduce the burden on counterspeakers [67], who are sometimes victims of hate speech themselves [20]. On the other hand, naive AI solutions that simply generate counterspeech without human oversight could significantly detract from the authentic, empowering, and emotionally connecting experience that counterspeech provides to users [12, 42], despite being a growing area of research in AI [24, 108, 149].

Given these challenges and tensions, a more deliberate approach that involves stakeholders with varying degrees of counterspeaking experience is required, to design tools that can empower counterspeakers and increase participation in counterspeech. To enable this, we conduct the first set of studies to collect various perspectives from users to understand their counterspeech experiences and to inform the possible role of AI technology in counterspeech. We partnered with an NGO specialized in responses to hate speech to ask the following research questions:

- **RQ1** What are the barriers, if any, for users to engage in counterspeech?
- **RQ2** What AI tools, if any, can assist in removing or lowering these barriers?
- **RQ3** What are user concerns of using AI tools for counterspeech?

We conduct our studies with two populations of participants with differing levels of counterspeech experience to understand both the needs of current, extensively experienced, counterspeakers as well as less-experienced everyday social media users. We conducted semi-structured long-form interviews with activists who regularly

*Equal contribution.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642025

respond to hatred online (*N=10*), and also surveyed with *342* everyday social media users who may have never participated in counterspeech.

Through qualitative analysis, such as grounded theory, we found several emerging themes, corroborated by our quantitative results. Our participant responses surfaced tension between barriers and motivations to counterspeech. At a high level, participants identified *limited resources*, *lack of training*, *unclear impact*, and *fear of personal harms* as deterrence to engage, but were motivated by a sense of *moral duty* and *positive impact*. We characterized AI tools that were envisioned by participants, functional requirements that connected to the barriers and the characteristics of such tools (e.g., emotional, factual, empowering), toward designing participatory AI systems for counterspeech. Furthermore, we found that while participants thought that AI tools could help address these barriers, they expressed *functional doubts* and strong concerns of *authenticity* and *agency* in using AI counterspeech tools.

These findings highlight the gap between current research directions in counterspeech AI assistance and designing an empowering tool for counterspeakers in addressing the concerns raised by our participants. As a step towards closing this gap, we outline design considerations for *authenticity of counterspeech*, *moral agency*, and *mental health* by connecting our findings to previous works. More specifically, we make recommendations towards ensuring transparency in online communication, avoiding moral passivity and disengagement, and promoting mindfulness and intentionality in interactions and call for future studies to explore the concerns of AI counterspeech assistance and their solutions.

In summary, our work seeks to understand from both experienced activist counterspeakers and lay-users their perspectives on online counterspeech through learning about their counterspeech experiences, envisioned AI tools, and concerns about such tools. In so doing, we contribute to the discussion on the fight against online hate especially on counterspeech and role of AI-driven counterspeech tools. Specifically, our contributions are in **(1)** synthesizing the experience of two different populations of users to describe a theory of counterspeech engagement and barriers to counterspeech, **(2)** collaboratively envisioning potential usage of AI, **(3)** surfacing concerns of AI tools, and **(4)** broadening the vision of AI tools for counterspeech and offering design considerations.

2 RELATED WORK

To explore counterspeaker needs and AI tools for counterspeech, we first situate counterspeech more broadly in relation to other hate speech responses (Section 2.1). We then investigate related works to understand counterspeech and counterspeakers (Section 2.2). Finally, we provide an overview of prior works in AI and counterspeech (Section 2.3) to investigate the direction of AI assistance in counterspeech and the research gaps that we aim to address in our work.

2.1 Hate Speech Responses

Hate speech is a commonly seen problem in online communication, especially on social media platforms [36, 94]. Definition of hate speech varies across countries, institutions, and communities, which leads to disagreement over appropriate responses [79]. In the most extreme cases, hate speech is defined as a weaponized speech that causes direct harms to individuals such as inciting violence or hostility [79]; however, speech that causes indirect harms including devaluation of community norms such as inclusivity can also be considered hate speech [79, 99]. Thus the many forms and degrees of harm in hate speech require flexible countering responses [73]. One aspect of hate speech most commonly shared by various definitions is that it has a specific target, a group of people distinguished by different identities such as religion, race, gender, and sexual identity [41]. With the scale and pervasiveness of cyberspaces, online hate speech can be especially damaging to the targeted social groups as it fosters hostility and perpetuates stereotypes and can even lead to off-line violence [13, 17].

One of the most prevalent solutions to the growing scale of hate speech has been content moderation through algorithmic censorship [41, 103]. However, algorithmic censorship has been shown to have many shortcomings [100] causing growing concerns over freedom of speech, especially with the algorithms tailored to serve the platform often excluding the users from its decisions [1, 29, 75]. Moreover, many automated hate speech detection has been shown to be inaccurate, unable to detect subtle hate speech [57, 62] or biased against certain communities causing further marginalization [110, 111]. Some new proposed methods of online governance include user-based and community-based moderation models [122] such as collective-decision moderation [98], which require more involvement from users but offer a more democratic way to make moderation decisions. Counterspeech, along with these moderation methods, provides users and platforms with a more flexible solution that can be used as an alternative or complement to deletion-based methods to counter hate while influencing a positive cultural shift through dialogue.

2.2 Counterspeech and Counterspeakers

Counterspeech is a complex phenomenon with many potential benefits to correct stereotypes [84], prevent the spread of misinformation [81], reinforce correct information [126], and to expand responses to even more covert forms of hate speech [4]. Its beneficial outcomes can help alleviate online hate in many contexts: as part of a moderation decision to increase transparency by communicating moderation intent [113], as an intervention against cyberbullying that also shows support towards the victims [105], and more broadly, as a remedy for harmful speech (e.g., misinformation, propaganda, hateful speech) [97]. Research on online counterspeech has been focused on social media interactions [93, 97] to measure its patterns, effectiveness, and role against hate speech. Many of these studies have been focused on finding effective content for belief or sentiment change measured through subsequent behaviors of the poster (e.g., deletion of the hateful post) or those engaged in the discourse (e.g., sentiment of the comments) and have shown empathy [59] and civility [56] to have some positive impact on these measures. However, it requires effort and engagement from the users doing it, which becomes challenging.

Counterspeakers' goals are often multifaceted: to change the view of the bystanders, to recruit more counterspeakers, or to strengthen community norms [14]. Moreover, counterspeaking can create connection and solidarity against hate among users as they respond to online hate speech collectively as a part of an organized movements such as #iamhere [12, 42]. While content moderators sometimes use counterspeech to de-escalate conversations and algorithmic support for such proactive moderations have been investigated (e.g., flagging conversations that are likely to devolve into toxicity [116]), most counterspeakers who are everyday users have different positionality and power compared to the moderators [140]. Therefore, in our work, we focus on counterspeakers and the public to understand the barriers to counterspeaking and tools to support counterspeech. More specifically, we look at hurdles and reasoning behind decisions to not engage and how AI tools can lower these barriers, which have not yet been studied and require further investigation.

2.3 AI Assistance in Counterspeech

As counterspeech has become an increasingly important avenue for responding to hate speech online, the issue of scaling this effort has become an active research topic. Full automation of counterspeech is thoughts of as artificial agents such as bots responding to speech that are detected to be hateful [34]. Following such framework, automation of hate speech detection could be considered a part of counterspeech assistance; however, the focus of these studies have been on detection only with limited annotations [41, 103] lacking in specification or adaptation for diverse solutions [100]. Counterspeech generation and detection have more recently become the focus of research community. Automatic detection [147] and computational analysis of large scale counterspeech [43] have been used to understand its characteristics and to inform effective content. Prior works on automatic generation of counterspeech [91, 149] relied on curated or scraped datasets [25, 63] and evaluation metrics based on correct countering claims [55] or emotion and politeness [108]. Some methods used limited response intent such as question, denouncing, and humor in dataset [25] and as part of the generation method [53]. In addition, counterspeech generation in dialogue systems (e.g., Alexa, Cortana, etc.) has come into focus as well [132]. These approaches, while offering interesting explorations, did not consider counterspeech and its full context or the complex user intentions of counterspeaking.

Moreover, ethical implication and value of automation have not been addressed and require further consideration [34, 46]. This is becoming an increasingly important topic to address as online hate incidents are increasing [82] and AI, which is not well defined even among experts [129], is becoming more and more prevalent in our society and in the public's daily activities [19]. Therefore, in this work, we seek to answer how AI can best help counterspeakers by bringing attention to users, their intentions and barriers, and collaboratively identifying possibilities and concerns of AI for counterspeech.

3 METHODOLOGY

To understand counterspeech with diverse perspectives of both activists and everyday users, we used a mixed methods approach [119]. The interview participants, who are counterspeech activists, are characterized by their extensive involvement in counterspeaking through systematic efforts (e.g., repetitive, over an extended period of time) and their understanding of counterspeech. These participants were recruited through pre-established relationships with the partnered NGO. On the other hand, the survey participants were recruited on Amazon Mechanical Turk (MTurk) without particular pre-screening based on their counterspeech experiences, to capture perspectives more representative of everday social media users. Our approach focuses on understanding both the depth and breadth of counterpseech experiences by mixing qualitative and quantitative studies of the two populations. To this end, we developed semi-structured interview guidelines and a survey informed by the responses using an exploratory sequential design [33]. Moreover, to gather diverse applications of AI from both groups of participants, we first ask questions about an envisioned application without defining the type of AI (e.g., generative AI, classification models) as to ensure creative and even future-oriented answers. We then ask more specific questions about some common types of AI tools envisioned by experts in relevant literature and by the research team, such as tools that use generative AI (e.g., counterspeech bots, revision support through rewriting) or uses retrieval based support (e.g., tool that suggests facts). We used grounded theory to analyze the qualitative results and triangulation to support our findings discussed in Section 4.

Participant ID	Country of Residence	Years Counterspeaking
1	Ethiopia	2
2	USA	4
3	Cameroon	5
4	Australia	7
5	France	6
6	Canada	5
7	USA	10
8	USA	3
9	USA	2
10	USA	4

Table 1: Interview participant demographics: their country of residence and number of years being involved in systematic counterspeaking.

3.1 Interview Study

We first conducted semi-structured interviews with experienced counterspeakers to understand counterspeech from participants with diverse experiences countering hate and a more developed identity as counterspeakers. We asked questions to understand their counterspeech experiences and thoughts on AI tools, possible benefits and drawbacks, to provide insights into each of our research questions.

3.1.1 Interview Procedure. To gain an in-depth understanding of the challenges counterspeakers face, the strategies they use, and their thoughts about using AI to improve their counterspeech, we employed a qualitative research design, utilizing semi-structured interviews as the primary data collection method. We chose semi-structured interviews as they allow participants to express their thoughts, experiences, and perspectives, while also providing the flexibility to probe for deeper insights as the conversation unfolds. To ensure consistency, we developed an interview guide which consisted of open-ended questions designed to explore participants' experiences doing counterspeech (e.g., methods of finding hate speech, frequency and audience of counterspeech, and most rewarding experiences) and their insights into how AI tools could aid or complicate their work (e.g., their thoughts on counterspeech bots, open questions about envisioned AI tools). One of the authors conducted all interviews over online video calls in English between May and July of 2023.

3.1.2 Participant Recruitment. Ten participants were purposefully selected based on their long-term experience of counterspeaking online in a systematic way. Recruitment was carried out by direct invitations through the partnered NGO. The participants came from a variety of backgrounds in different contexts (from Europe, Africa, and North America). Some did counterspeaking collectively, while others responded individually. This sampling approach aimed to ensure a diverse range of perspectives and rich data for analysis. All participants provided informed consent prior to their participation, and confidentiality and anonymity were maintained throughout the study, with pseudonyms used in reporting findings. We report their country of residence and years of experience counterspeaking in Table 1.

3.2 Survey Study

To understand a broader user experience with hate speech and counterspeech, we conducted a survey study with participants from Amazon Mechanical Turk (MTurk). To ensure the quality of our results [76], we pre-qualified workers using the process detailed in Appendix A.1 and excluded survey results with nonsensical qualitative answers (e.g., repetitive answers, discussing irrelevant technology, or using copy-pasted answers from other internet sources). The study was approved by our institution's IRB, and all survey responses were collected in August 2023. The survey took median 8 minutes, and we compensated workers at a rate of $12/hr.

3.2.1 Survey Questions. To get an overview of counterspeech experiences of a wider population, we designed a survey asking participants questions shown in Table 2. The three main parts of the survey were hate speech experience, counterspeech experience, and AI tools for counterspeech, followed by questions on demographics. In the first part, we asked the participants questions about social media usage and their experiences with hate speech online such as types of hate and details about their experience (e.g., platform and type of online space - public or private) (5 questions). The second part of the survey consisted of questions about counterspeech experience as previous responses to hate speech and barriers to responding (7 questions). The third part of the survey focused on questions about AI tools (e.g., openness to using an AI tool, preferences towards specific types of AI assistance) (5 questions). To avoid priming the participants toward specific type of responses on envisioned AI tools, we showed open-ended questions (SQ14 and 15) on separate screen to the survey questions that mention tools suggested by the research team. Additionally, questions asking about hate speech and counterspeech experiences (SQ3-SQ14) were skipped for those who responded that they have never seen hate speech online (SQ2, option "Never", *N=12*).

3.2.2 Participant Demographics. Since we recruited mainly from U.S. and Canada, majority of the participants were from the U.S. (98%) and many of them identified as white or Caucasian (83%). On question about gender identity, 56% of participants identified as male and 41% female. When asked about sexuality, 85% reported as being straight (heterosexual) followed by bisexual (6%) and asexual (2%). On the political spectrum, participants were liberal-leaning with strongly liberal 22%, liberal 32%, moderate 18%, conservative 17%, and strongly conservative 8%. As shown in Figure 1a, 94% of the respondents had experience encountering hate speech online and 70% had experience responding to hate speech (e.g., writing a comment, sending a private message, or adding a negative reaction) but only 8% did so frequently or all the time even though 22% encountered hate speech frequently or all the time. Out of those who had experience countering hate speech, 72% had previously responded by adding a comment or reply under the post. Moreover, the most commonly seen type of hate speech was race-based or ethnicity-based (56%) (Figure 1b), and most participants primarily responded either equally as an ally and targeted group (45%) or as an ally (36%). Additional analysis of demographics of participants are shown in Appendix A.2

3.3 Analysis Methods

To conduct a comprehensive analysis, two separate groups of authors first performed grounded theory and open coding on the free text interview and survey data, respectively. The authors then consolidated and synthesized the findings to build a cohesive theoretical framework. Additionally, the quantitative responses were analyzed and integrated with the qualitative results.

3.3.1 Data Preparation. Interview data was transcribed by the author who conducted the interview to ensure participant confidentiality. Moreover, the qualitative responses of the survey data were aggregated using quantitative results to answer specific research questions. For barriers to counterspeech (RQ1, Section 4.1), we looked at the open-ended responses to the survey question about barriers (SQ12, option "other"). For the possibilities of AI tools in counterspeech (RQ2, Section 4.2), we analyzed the reason why participants were willing to use AI tools in counterspeech (SQ14) from participants who were willing to adopt AI tools in counterspeech (SQ12, options "likely" or "neutral"). We also analyzed participants responses on envisioned AI tools in counterspeech (SQ15). For the concerns of AI involvement in counterspeech (RQ3, Section 4.3), we analyzed the reason why participants did not want to use AI tools in counterspeech (SQ14) from participants who were unwilling to use these tools (SQ12, options "neutral" and "unlikely").

3.3.2 Qualitative Analysis. Grounded theory methodology [21, 47] was employed to analyze the interview data. This iterative and systematic approach allowed for the discovery and development of theories directly from the data. The analysis process involved three key coding stages: open coding (i.e., generating initial codes by breaking down the data into meaningful units), axial coding (i.e., identifying broader categories and subcategories to establish relationships between codes), and selective coding (i.e., developing a theoretical framework by refining and integrating categories). Survey data was also analyzed using open coding but did not employ the last two stages of the grounded theory.

Question Topic	QId	Question
Hate Speech	SQ1	Which social media platforms or online spaces do you use at least once a week
	SQ2	How often, if ever, do you encounter speech online that you consider to be hateful?
	SQ3	On which social media platforms or online spaces do you feel like you see hateful speech most often (choose up to three)
	SQ4	Do you see more hateful speech in private online spaces (e.g., direct messages, private Facebook group) or public online spaces?
	SQ5	Which of the following categories of hateful speech do you see most frequently? (select all that apply)
Counterspeech	SQ6	How frequently, if ever, have you responded to hateful speech in a way that tried to counter the speech? (e.g., writing a denouncing or disagreeing comment, sending a private message, adding a negative reaction)
	SQ7	How do you usually try to counter hateful speech? (choose up to three)
	SQ8	Which statement do you agree with most?
	SQ9	Who do you primarily respond to?
	SQ10	If you have written a reply to hateful speech online before, which of the following tactics have you used (check all that apply)
	SQ11	If you had to choose your most used tactic from that list, which would you choose (same question as before, but this time just tell us your most used tactic):
	SQ12	If you have seen hateful speech online before and chosen NOT to respond (react or write a reply), which of the following do you agree with? (choose all that apply)
AI Tools	SQ13	If there was an AI tool to help you respond to hateful speech by specifically addressing the concerns you selected previously, how likely would you be to consider using it?
	SQ14*	What are the reasons that you are likely or unlikely to use it?
	SQ15*	What type of AI assistance do you think would make you more likely to write a reply to hate speech?
	SQ16	Select the following possible AI tools that would be useful for you to post a reply to hate speech. (select all that apply)
	SQ17	How do you feel about the following bots that automatically engage with hate speech?

Table 2: Questions listed in the public survey with participants on MTurk. * denotes an open-ended survey question.

Interviews. The first two stages of coding were conducted by the same author who conducted the interviews, as was required by our IRB to protect participant confidentiality. During open coding, the interview transcript was broken down into meaningful units through line-by-line coding. Then, stage three coding was conducted by the research team, integrating the findings of both the survey research and the interview study.

Survey. For qualitative survey data, two authors inductively generated codes for 25% of the data by developing individual codebooks and labeling each instance with one or more codes. Each code was given a name and a short description. Next, the authors convened to merge their codebooks by combining codes with similar themes into a single code and unanimously agreeing to add, merge, or delete the codes through discussion in case of disagreement. Finally, the authors performed a second round of coding by deductively applying the shared codebook to the entire dataset individually. The authors then reconvened and for each instance, applied the codes upon unanimous agreement based on discussion. Disagreements largely occurred due to differing scopes of codes and at times different interpretations of the statements, which were resolved through discussion.

3.3.3 Quantitative Analysis. We report the quantitative responses with percentage of participant responses. For questions that allowed multiple choices (e.g., select all that apply, choose up to three), we calculated the percentage over the number of participants who responded. Following best practices for opinion surveys [77], we also aggregate similar responses together (e.g., extremely likely, likely).

4 FINDINGS

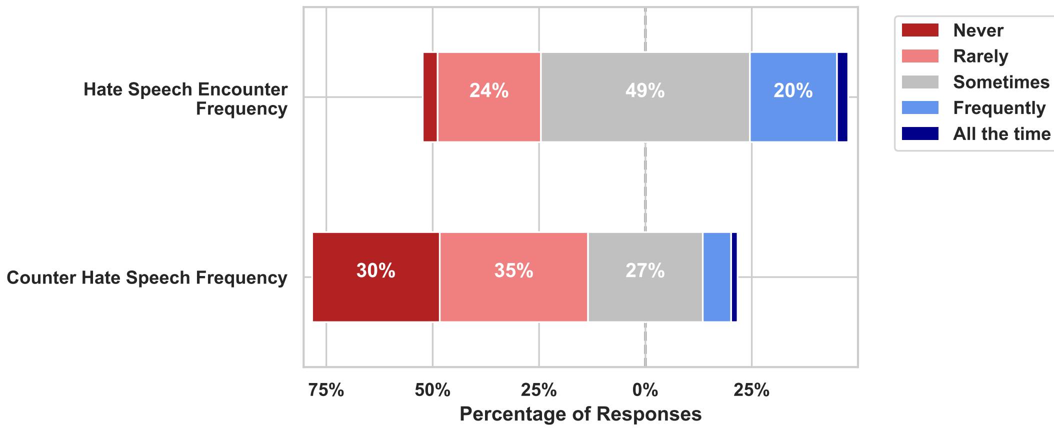
To understand counterspeech and potential impact of AI involvement from both activist- and lay-user perspective, we present findings from the analysis of both studies together in this section. As shown in Figure 2, we found that AI usage benefits could be mapped to the barrier each addressed. Moreover, the concerns of AI tools could also be linked to the themes discussed in other section in detracting from intrinsic and extrinsic motivations, questioning the usage benefits, and negatively impacting counterspeech as a whole.

We first describe our theory of the counterspeech process, and then findings related to each research question: barriers to counterspeech (RQ1), possibilities of AI tools (RQ2), and concerns of AI involvement in counterspeech (RQ3). We denote interview participants as IP and survey participants as SP throughout the section for clarity. We highlight the names of codes by using an underlined font style throughout this section for visibility and clarity of connection between the findings and the code. Codes developed for each section are listed in Appendix B.

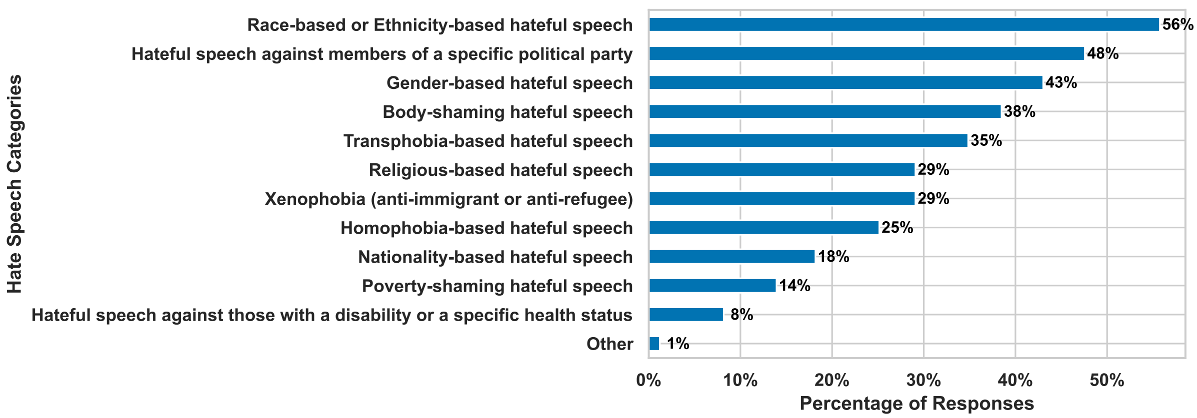
An Inductive Theory of the Counterspeaking Process and Motivation

In our analyses, we found that many counterspeakers shared a similar process for counterspeaking, and had similar themes for their motivation to engage.

Three-step Counterspeaking Process. Our analyses highlighted that these three steps were commonly discussed in participants' counterspeaking process: (1) deciding to respond, (2) formulating a response, and (3) posting as well as engaging with the reactions of the audience. As a first step, counterspeakers either came across or actively looked for hate speech, assessed its harms, and made decisions based on effort and impact trade-offs on whether *"it would be effective enough to be worth my time and effort"* (SP135). After deciding to respond, participants formulated a response, usually comments or posts, to counter hate. Some words used by participants characterizing the responses they would create were *"thoughtful"* (SP331), *"impactful"* (SP81), and *"mindful"* (SP289). The last step of counterspeaking was posting and dealing with the reactions including positive reactions such as *"liking a comment or responding to a comment"* (IP9), negative *"push back"* (IP4), or no reaction. Each step of counterspeech required effort and time, although varying in amount, and a set of barriers existed, which are discussed further in Section 4.1.



(a) Hate speech (SQ2, *N=342*) and counterspeech (SQ6, *N=330*) experiences of participants shown through responses to likert scale question on frequency of hate speech encounter and counter hate speech frequency.



(b) Most commonly seen hate speech categories (SQ5, *N=330*) identified by the participants shown in percentages of participants who selected the option. Participants were asked to select all that apply.

Figure 1: Participant responses on hate speech and counterspeech experience questions and most commonly seen type of hate.

Counterspeaker Motivations. Experienced counterspeakers expressed both intrinsic and extrinsic motivations, *moral duty* and *positive impact*, that encouraged them to take the above steps to engage against hate.

Counterspeakers we interviewed largely saw counterspeaking as a moral duty. Many believed that it was the *"right thing to do"* (IP1). Many found meaning in counterspeech in shared values as IP7 noted:

> *"I believe in it. I didn't do it just because I got a kick out of it."*

The sense of moral duty towards counterspeech was also shared by survey participants as SP107 wrote:

> *"I would really like to be able to make the internet a safer and more healthy place to spend their time, so if I could reduce the amount of hatred and misinformation on there, I would do it."*

However, some survey participants disagreed with this perspective sharing that *"I did not respond because for the most part people should be able to say what they want. It's up to us whether we choose to be hurt by someone's comments or not"* (SP190).

Another extrinsic motivation discussed by many participants was the positive impact of counterspeech. The types of impact participants found most rewarding varied, ranging from influencing the conversation to causing a view change or consoling those targeted by hateful speech. For example, participants mentioned the following rewarding experiences: seeing a positive change in the comment section, learning that their counterspeech *"changed the mind of (even) one person"* (IP1), and knowing that they have *"helped someone who had maybe been reading the comments and had been upset by them"* (IP4). Participants emphasized that they felt rewarded even when the scale of their impact was small.

4.1 Counterspeech Barriers (RQ1)

To answer our first research question (RQ1), we analyzed participant responses around pain points of counterspeech and reasons behind not engaging in counterspeech. As shown in Figure 2, there were

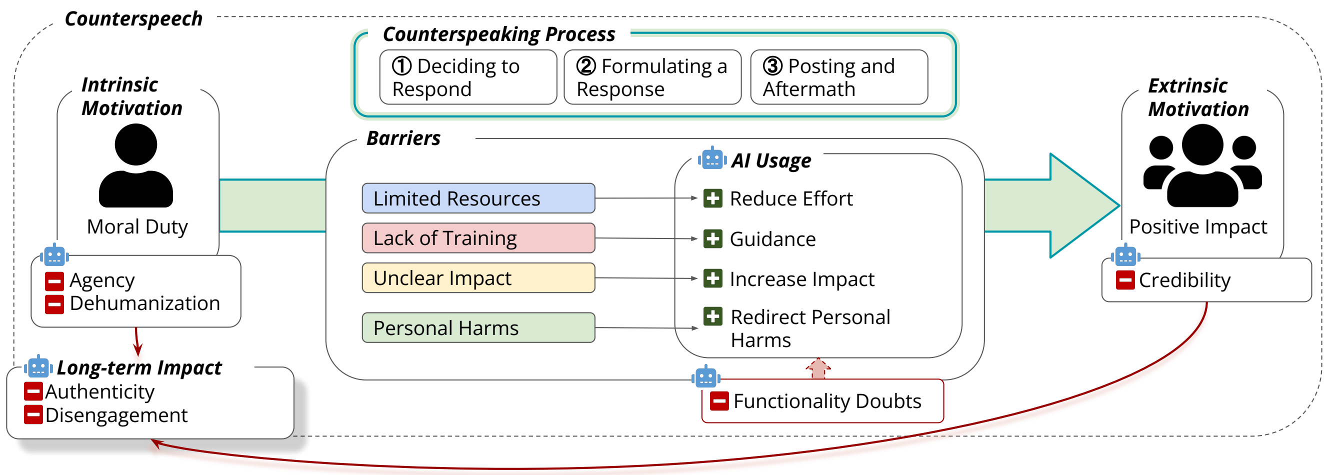


Figure 2: An overview of interactions between the themes surfaced in Section 4. The figure, from left to right, show an overall counterspeech experience surfaced from participant responses. Participants' intrinsic and extrinsic motivation encouraged participants to engage in the counterspeaking process broken down into three steps. Themes found in beneficial AI usage could be rooted in each barrier, shown by the arrows in the figure, and themes found in AI concerns were linked to different aspects of counterspeech including motivations, functionality of AI for counterspeech, and counterspeech as a whole.

four high-level themes in counterspeech barriers identified by the participants: *limited resources*, *lack of training*, *unclear impact*, and *personal harms.*

The barriers that were discussed commonly across all stages were required resources such as time and energy and personal harm on mental health. Overall, participants discussed that counterspeech can *"take a lot of resources"* (IP2) and can be overwhelming as *"it sometimes gets to be too much"* (IP5). Since most interview participants were volunteer counterspeakers, there was no compensation for their time or harm to their mental health. As IP2 shared, *"No one pays me - it's the time investment"*. Similarly, the resource barrier was reflected in the survey responses as 21% of the participants answered that they did not write a reply because they did not have enough time (Table 3). The concern for mental health was also shared. As SP12 said, *"I would get too upset about it and I have enough stress already in my life."* For the remaining discussion of counter speech barriers, we organize our discussion of how these barriers occur with respect to the three steps of the counterspeaking process.

Deciding to Respond. Our findings showed that the most influential consideration to the decision of counterspeech was in the resource and impact trade-offs. As a way of strategizing for impact, activists spent time finding hate speech looking for interactions where it would be *"worth it"* (IP1) for them to respond. They often looked at not only the hateful post but also the interactions around it (e.g., activity level of the comment thread), which could take additional time as reflected in IP10's experience of spending *"up to two hours looking for good actions"*. Similar effort to assess resource and impact trade-offs had to be made in discerning whether the hate speech was coming from *"trolls"* (SP115) or *"bots"* (SP341). For example, SP232 shared, *"I didn't know if the person who stated is in [sic] actor paid to say it so my response would be meaningless or that the writer is actually just a bot"*. Correct assessment of the impact was further complicated by algorithms on the platform as noticed by SP227, *"I don't want to engage & help the comment have more impact"*. Supporting these decision considerations shared by the participants, the unclear reach of counterspeech at this stage was the most influential reason to not engage (73%; see Table 3).

Formulating a Response. Participants shared that this stage can be one of the most time-intensive as IP6 reflected, *"It takes me forever to craft something that would make sense."* Moreover, lack of training was also noted as a barrier at this stage as people *"don't always know what to say"* (IP3) to reach their audience to effectively counter hate. This can be especially true for beginners. As IP7 recalled, when he first started counterspeaking, he was *"not as well-versed in the subject or in framing the argument and being persuasive."* Survey results also showed that lack of training was a barrier to counterspeech, however, not all participants responded by formulating their own message. Out of those who had responded to hate speech before, 72% of the survey participants said they had previously responded by commenting or posting a reply, but some chose to rather use existing response methods such as downvoting or disliking the post (70%) or reporting the hateful post (49%). While not all respondents had experience writing counterspeech, when asked about reasons for not engaging, 22% of participants did not know what to say and 17% did not know how to express what they wanted to say (see Table 3). Survey participants used various tactics in their responses. Most commonly, they had "tried to correct misinformation or fact-check inaccuracies" (70%) or "tried to shame the person who has posted hateful speech" (36%) but also posted links to other sources (34%), tried to be funny (21%) or emotionally connect (21%).

Posting and Aftermath. At this step of the counterspeaking process, the lack of reactions and negative reactions are barriers to

Counterspeech Barrier	***Response, % (N=330)***
I did not respond because I didn't think responding would have an impact	73
I did not respond because I didn't want people to be mad at me or send me hateful or threatening messages	31
I did not write a reply because I did not know what to say	22
I did not write a reply because I didn't have enough time	21
I did not write a reply because I didn't know how to say what I wanted to say	17
Other	9

Table 3: Options and participant responses to question about counterspeech barriers (SQ12 in Table 2). Participant responses are shown in percentages of participants who agreed with the listed reason for not responding. Participants were asked to choose all that apply.

future action. Counterspeaker activists reported being discouraged or demotivated by the limiting reach as one participant highlighted, *"When you feel unheard and it's like I'm doing this for nothing - it's not really getting the word out - it's frustrating"* (IP1). Counterspeech sometimes caused participants to become the target of hate as well, negatively impacting their mental health. For example, IP4 recalled that *"For a while, the push back was getting to me"*. Therefore, counterspeakers opened themselves up to varying amounts of risk when posting, and it became a significant barrier as shared by IP4, *"The less I say, the less room I give people to attack."* This is similarly reflected in the public survey opinion, as 31% of participants saying they did not respond because they did not want to upset others or receive hateful messages (see Table 3). Additionally, survey participants saw more hate speech in public online spaces (84%) and were primarily responding to audiences that included people that they did not know (94%). As IP1 reflected, *"it's risky when you put yourself out there"*, highlighting that responding to hate speech can open up confrontation in public spaces with strangers. This risks could be amplified when counterspeakers commented about highly controversial topics or did their work while living under authoritarian regimes or in conflict zones.

4.2 Possibilities of AI Tools (RQ2)

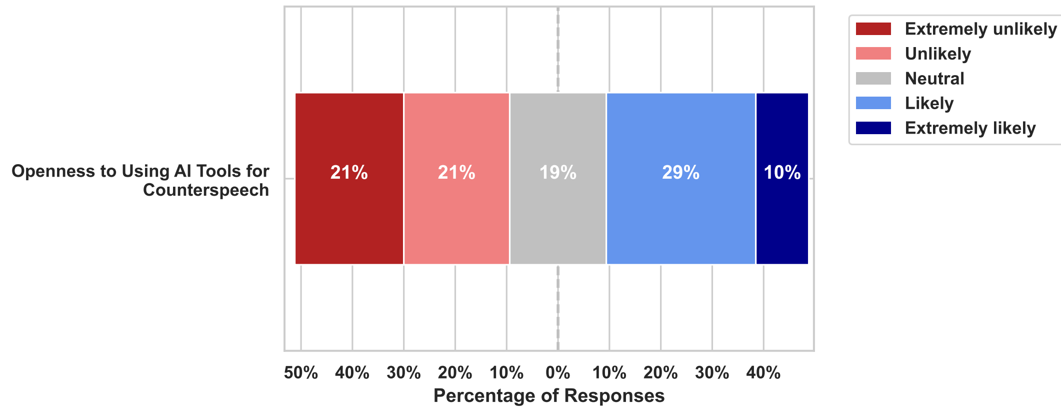
To answer our second research question (RQ2) on possible AI tools, we discuss our findings from the analysis on participants' responses about the benefits of AI tool usage and their description of tools that would encourage them to write a response. Overall, the biggest difference seen between the two populations was in level of AI involvement; active counterspeakers saw possibilities for using AI to augment their work to make the process more efficient and amplify their voices, whereas lay-participants expressed a more diverse set of needs and preferences. The usage goals and characteristics could be mapped to the themes of barriers they addressed as seen in Figure 2. We break these down into a description of the types of support, empowerment, emotional, and factual support and the types of tools, level of involvement in the counterspeaking process and the definition of usability.

Empowerment. Some participants also characterized their preferred tool as empowering and aligned, with themselves or cultural societal norms, especially towards creating a positive impact. For example, SP97 was interested in using AI tools *"So that I can help people being attack(ed) by these hateful speech."* Similarly, SP139 showed positive interest towards AI tools *"Because it might increase the chance to actually make an impact"*. SP28 was interested in the support of AI tools *"Because it would help me speak up more."* Participants also wanted tools that were aligned with them, which could *"write a response EXACTLY LIKE I WOULD"* (SP30). Some acknowledged that tools also need to be aligned with cultural and societal norms to create *"unbiased"* (SP126) responses.

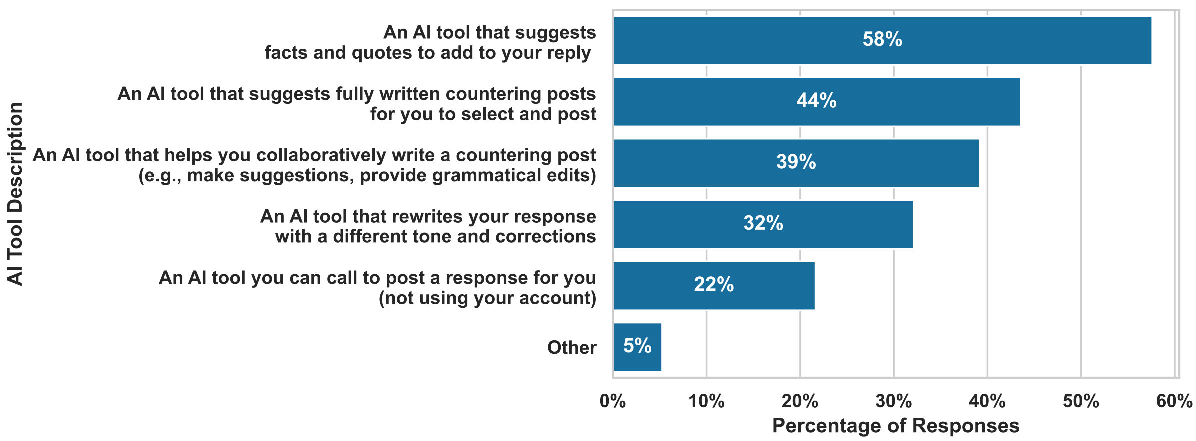
Emotional Support. Participants brought up the need for emotionally supportive AI tools in various ways. On one hand, some participants thought that AI tools could help them with regulating their own emotions as they believed that AI's *"emotional detachment could help me remain composed"* (SP160). On the other hand, others thought that AI tools could help connect people to *"think in a more empathetic way"* (SP262) focusing on different perceived AI capabilities and stylistic preferences. Emotional support was also a commonly brought up need to help formulate an effective communication. Most notably, participants wanted tools to help communicate or understand emotions that shows emotional intelligence. For example, SP107 proposed a tool with an *"AI that could delve into the psychology of why a person posted what they did, and then give me a response that could really tap into that person's mind and give me a response that they will actually care about."*

Factual Support. Many participants were also interested in AI tools that could gather or show relevant information. These participants wanted support in crafting factual and logical responses, either via resource recommendation or through fact- or argument-checking their own responses, to correct misinformation in the hateful speech. An example of such tool, discussed by SP9, would *"pull in articles to prove a point/correct misinformation"*. In corroboration of these desired needs, fact-based support was the most preferred in our quantitative survey responses as evidenced by relevant facts bot (66%) and an AI tool that suggests facts (58%) being the most selected or liked AI usages (Figure 3c and 3b).

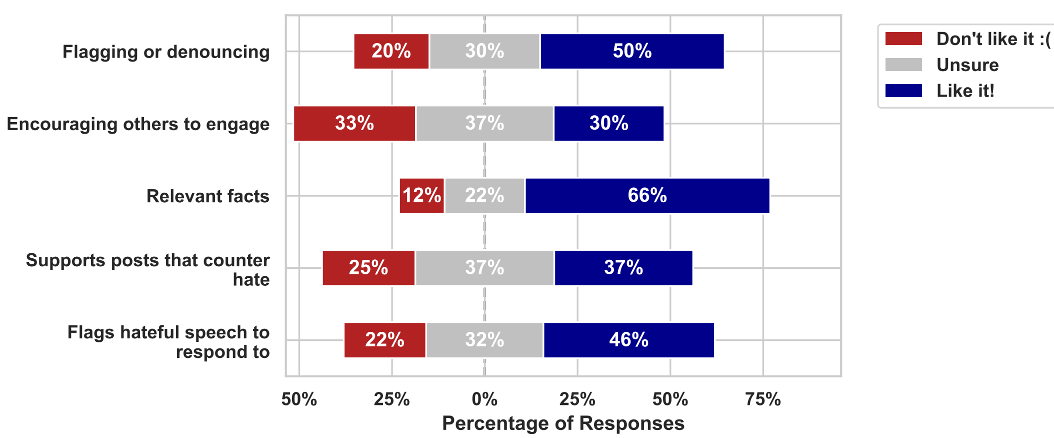
Level of AI Involvement. Participant responses showed preferences towards varying amount of AI involvement at different stages of the counterspeaking process (Section An Inductive Theory of the Counterspeaking Process and Motivation). In the first step of the process (*deciding to respond*), some survey participants showed preference for an AI tool that automates this step as a part of a full automation, i.e., that detects hate speech and automatically reports and/or *"blocks or deletes"* (SP148). Some interview participants also expressed interest towards a detection tool to *"identify hate speech more efficiently"* (IP5), however, unlike the survey participants, wanted full control of the subsequent counterspeaking steps.



(a) AI tool perception (SQ13, *N=330*) shown in percentages of responses to each likert scale option.



(b) Collaborative tool preferences (SQ16, *N=342*) showing the percentage of participants who selected the tool as potentially useful. Participants were asked to select all that apply



(c) Bot preferences (SQ17, *N=342*) shown in percentage of responses to each likert scale option

Figure 3: Survey participant responses on quantitative questions about AI tools.

In the second step (*formulating a response*), the varying levels of involvement ranged from collaborative tools to provide guidance *"to find arguments to counter hate speech more effectively"* (SP104) to response support tools like *"one(s) that recommends possible replies"* (SP132) . The collaborative tools described by participants were characterized by less AI involvement with more input from users that would provide correction of *"mistakes like grammatical mistakes, factual mistakes, etc."* (SP256), *"customization options"* to preserve communication style (SP160), or *"help with idea generation"* (SP231).

At the last step of the counterspeaking process (*posting and aftermath*), participants often wanted heavier AI involvement to reduce personal harms. One type of support identified was to provide an AI proxy, often in anonymity, to counterspeak *"without worrying about potential blowback"* (SP48). Moreover, participants wanted tools to help deal with possible negative reactions with a protective AI tool, for example, *"An AI that would respond to the hateful things that people said back or messaged to me."* (SP110).

Overall, participants wanted AI tool's involvement to also help reduce harm to mental health by making counterspeaking *"less stressful"* (SP35). Additionally, IP3 proposed an educational tool that could be helpful to address the barrier in lack of training, *"It would train people and walk them through the process of how to design a campaign."* Notably, the two populations showed different levels of preferred AI involvement, as lay-participants endorsed full automation and showed more interest towards heavier involvement.

Defining Usability. The most prevalent preference surfaced was towards a *usable* tool. While this result is unsurprising, understanding aspects of usability could be helpful as to avoid building tools that are confusing, difficult to learn, or even useless. Participants described usability in two dimensions, ease of adoption and functional. Especially to address the barriers related to limited resources, participants wanted efficiency and to save time, as specified by SP9 as *"something that makes it as quick and easy as possible to reply"*. Some participants further detailed a possible tool that would be *"provided by the site itself so I would not have to waste time getting a reply from another site"* (SP95) highlighting that it should cause a minimal amount of disruption or overhead. There were varying perceptions of the functionality of AI systems, as some survey participants believed that AI would be better than them in being *"more effective"* (SP169) by having *"more experience"* (SP201), whereas others were more skeptical and wanted to *"try it to see if it is beneficial for me"* (SP6). Thus, many participants echoed curiosity around usability of the systems, which depended on their functional benefit.

4.3 Concerns of AI Involvement in Counterspeech (RQ3)

Our final research question (RQ3) asked about concerns regarding AI involvement in counterspeech. To answer this question, we present our findings from analysis of interview and survey responses that presented reasons against using AI tools or showed concerns about AI involvement. Our analysis surfaced four themes: short-term concerns of *credibility*, long-term negative impact on *authenticity* leading to *disengagement* in counterspeech, loss of *agency* over and *dehumanization* of counterspeech, and *functionality doubts* about AI tools. As illustrated in Figure 2, concerns about AI involvement could be connected to their negative impact on motivations discussed in Section An Inductive Theory of the Counterspeaking Process and Motivation, benefits of AI tools mentioned in Section 4.2, or counterspeaking as a whole in long-term.

4.3.1 Loss of Authenticity and Agency. Authenticity and agency were overarching concerns across the two different populations of participants. In the short term, AI involvement could make counterspeech seem insincere and reduce speaker credibility to the audience (e.g., original poster, bystanders). For more experienced counterspeakers who have built an audience and a rapport, using AI could pose a risk of losing this trust. For example IP7 said, *"If people figure out it's a bot (an automated response), then it loses all credibility."* Similarly, sincerity of care about standing up against hate, especially in the choice to stand up for specific topics, made counterspeech meaningful as a moral action as expressed by SP294:

> *"If I'm going to respond to hate speech, I want it to come from me, because it's something I stand up for. I wouldn't want an AI to be apart [sic] of something so important to my personality and morality."*

Authentic care and intentions were also central to connection as shared by IP8:

> *"The process itself I find very satisfying. Having a sense of not being alone. You have all these people from all over the world, and you consider them friends, which is crazy, because you don't know them. For me, what's really touching is that someone out there is just there to support me. I'm not the only one who thinks this."*

Thus, participants emphasized the value of credibility and sincerity in their work as shared by IP2 that *"If you are trying to be genuine, then you have to, you know, be a genuine human."*

Participants also shared concerns over agency in the counterspeaking process that might arise with the use AI tools. For example, SP6 said, *"I like using my words not being tied to what AI says"*. Similarly, interview participants also communicated concerns of agency, reflecting that their identity as counterspeakers and more broadly as moral agents were reinforced through counterspeaking. For example, IP8 described how empowering it is when you make a visible impact after responding to online hatred:

> *"There is something incredibly magical about turning something really hateful in the other direction. You feel you aren't hopeless or helpless."*

Experienced counterspeakers also discussed their concerns about making moral compromises by using similar tools as those using hate speech bots making them *"no better than what is being used"* (IP6). Additionally, SP277 discussed possible moral degradation of counterspeech that would be made permissible by AI tools:

> *"I would like to have the support of the tool, and to be honest it sort of makes me feel like I have some plausible deniability if an issue arises. In a worst case scenario I would be able to "blame" it on the AI."*

Participants worried that over time these concerns could develop into long-term negative impacts on online communication. At a larger scale, excessive use of AI tools, especially without meaningful human oversight, could make joining real activism and finding genuine connection more difficult. IP5 warned against participation fatigue, comparing AI counterspeech to petitions:

> *"Look at petitions. There was a moment when petitions were kind of rare. Now you are harassed with petitions, and they don't have any purpose anymore. They are creating a false, passive attitude that 'I've already done my bit.' Why would we want to replicate that?"*

Moreover, without authentic intentions behind counterspeech, possible AI automation could reduce the meaning of counterspeech as IP6 emphasized through a comparison to robot fights:

> *"Where is the human component to that? Yeah, like, it's like watching those robot fights where it's just the robots. It's like, I don't know, then it becomes a game, right? And it almost, I don't know if the word is dehumanizes, but it desensitizes people from what is actually going on."*

Thus, careless AI involvement could exacerbate an already prevalent attitude towards disengagement shared by lay-participants who believed that engaging was not helpful and would rather avoid hate speech because they did not care enough. This sentiment was related in the response by SP137 who was not likely to use an AI tool because:

> *"[It] just seems like a waste of time that will create the prospect of them using an AI tool to respond to me. This will result in both of our AI tools going back and forth indefinitely and won't solve anything overall."*

4.3.2 Doubts in AI Capabilities. Functionality doubts of whether AI tools can actually address these barriers were also shared between the two populations of participants. Despite some survey participants calling for emotionally aware AI tools (Section 4.2), many participants believed that AI did not have the emotional intelligence to adequately counterspeak. Some noted that they are not *"funny or clever"* (IP2), and highlighting the lack of empathy in these systems, IP1 said, *"It couldn't cover the human aspect of empathy"*. Moreover, participants distrusted AI, sharing their perception that *"AI tools are often wrong and I wouldn't want its bias to affect what I am posting"* (SP143). The limitations of its training data and cultural bias were noted by IP3 who shared the concern that *"These are trained on Western data, so I immediately found a problem with that... Hate speech in Cameroon is definitely not [the same as] hate speech in the United States."* AI assistance was also seen as limited in solving the problem of personal harm such as becoming the target of retaliation, and SP156 noted that AI involvement would not solve the platforms' algorithmic problems to counterspeech as it would still *"drive(s) traffic to it (hate speech) which makes it a bigger problem"* (SP156).

5 DISCUSSION

To answer how AI assistance should be developed to improve the process of counterspeaking and not to detract from its meaning, we conducted interviews and surveys with both experienced counterspeakers and everday social media users to investigate three research questions: what are barriers to counterspeaking (RQ1), how could AI tools assist in counterspeech (RQ2), and what are some concerns about AI involvement in counterspeech (RQ3).

Our analyses surfaced barriers and AI needs with four different themes to inform functional needs of AI assistance and found that many participants thought AI tool could be empowering. Moreover, we discovered themes of counterspeaker motivations, especially in connection to self and others, highlighting the human components of counterspeech. Through understanding participants' concerns of AI involvement, we identified that without careful considerations, AI tools could do more harm than good, detracting from counterspeaker motivations and reducing meaning in communication. Based on our findings we build a set of recommendations and considerations for designing beneficent AI tools for counterspeech.

5.1 Possible Tools to Address Barriers to Counterspeech

Our study participants described many different barriers to counterspeech at various stages of engagement as discussed in Section 4.1. However, previous works in AI have focused on a narrow set of challenges for assistance and automation: automatic counterspeech generation [53, 55, 108, 149] and analysis and detection of both hate speech and counterspeech [43, 63]. Our findings paint a broader picture, especially through the theory of counterspeech engagement, and highlight where tools and resources would encourage bystander intervention towards countering hate. More specifically, participants described needs for further research and assistance in education, bot detection, and online safety. Education and training in all three stages (e.g., learning when to engage, how to engage, and how to handle backlash) could simplify and encourage counterspeech. For example, classrooms can play an essential role in counterspeech education as teachers already play an important role in educating students in handling digital risks and experiences [88]. However, as Castellví et al. has shown, training and resources are needed to equip them with diverse strategies to construct effective counterspeech [18]. Therefore, resources in counterspeech education, which could be enhanced by AI [10], and continued work on understanding different strategies and effective responses [59, 95] are highly necessary.

Moreover, participants highlighted the importance of *impact* in their decision to respond. In their consideration, knowing whether their comments would have an impact to their recipients or others mattered. Therefore, efforts in counterspeech research to understand and model the impact of counterspeech could be valuable to users to understand the impact of their actions, which could be aided by AI and algorithmic tools [44, 90]. Many participants also noted that they did not want to respond to a bot, and as previous works have shown, social bots (i.e., algorithms that produce content and interact with users on social media) are an ongoing problem in many social media platforms [40]. AI bot detection using neural methods have been explored to help social media platforms in filtering bots [32, 146], but with increasing capabilities of generative AI, detecting non-human agents is becoming harder [45]. Therefore, both legal efforts to deter bot accounts [130] and research on bot and AI detection would have to progress together to reduce the uncertainty and noise in online communication.

Another area of support highlighted by the participants was in online safety to overcome the challenges in posting and aftermath. In a survey conducted in 2017 and again in 2022, 41% of Americans reported experiencing online harassment [137] supporting the fear of retaliation experienced by bystanders [143] as related by our participants. The risk of intervening against online hate speech can also be exacerbated by gender roles and other identities of the intervener [142, 144]. While AI and algorithmic tools for moderation and flagging have been developed to alleviate online harassment and burden of moderators [6], they have shown to have varying levels of effectiveness [111] especially for marginalized communities who perceive greater harms associated with online harassment [68]. Along with improving existing methods and tools, further support in harassment reporting (e.g., collecting evidence or adding context) and recovery support could be assisted by AI-enhanced tools [50]. Moreover, policies from social media platforms and legal authorities have varying and unclear definition and responses to online harassment [101] which are often not reflective of cultural contexts [118]. This could cause interventions from users and AI tools to be confusing and ineffective. Therefore, continued investigation of online harassment through both research and policy efforts is necessary to create safer online spaces, which would not only encourage intervention from bystanders by minimizing retaliation harms to counterspeakers but also help define the role of counterspeech in this effort.

5.2 Preserving Authenticity of Counterspeech

One of the overarching concerns of AI tools that participants raised was authenticity of counterspeech and transparency of online communication. Participants raised concerns that AI involvement would cause counterspeech to be viewed as insincere or less credible. Moreover, many participants pointed out that AI tools or bots are frequently used to generate hate speech (Section 4.3.1), thus leading to the idea that automated AI counterspeech would be pitting bots against bots (e.g., *protesting bots*). Previous work supports this need for authenticity when speaking up against hate, for example in creating solidarity in activist movements and in calls to bystanders to organize and participate [66, 120]. This concern is also echoed by literature on AI which has shown that use of AI in communication can negatively affect trustworthiness of the speaker and authenticity of the message [48, 71].

Authenticity and trust is, therefore, a key consideration for designing AI tools that can help counterspeakers without damaging their message. To help promote authenticity in online communication and reduce information overload due to unclear sources [80], any system that generates automated counterspeech responses (bots) should clearly disclose that it is not a human [38]. Bot and AI generated text detection [32, 106], and disclosure policies [138] will also help ensure transparency in online communication. For responses generated through human-AI collaboration, there is less clarity on how much disclosure is necessary to aid authentic communication and connection. Other AI-mediated communication tools have been shown to create different perception of the writers and their intentions and led to feelings of deception [58, 65]. AI counterspeech tool, especially because of its closeness to morality (Section An Inductive Theory of the Counterspeaking Process and Motivation), may further exacerbate these negative effects. Therefore, it is an important research direction to understand how to frame and present communication that is collaboratively created with AI.

5.3 Cultivating Moral Agency and Engagement

Many participants saw counterspeech as a moral imperative believing that it was the right thing to do and chose to speak up against hate in ways that reflected their values (Section An Inductive Theory of the Counterspeaking Process and Motivation). This is consistent with the argument by Howard [67] that counterspeech is a moral duty for all citizens. However, with AI involvement, some participants showed concerns about creating passive moral attitudes leading to disengagement or shifting responsibilities to AI, echoing the concerns of moral passivity caused by AI [35] and misattribution of responsibility in AI mediated communication using AI as a scapegoat when conversations go awry [64].

Therefore, cultivating moral agency is an important design consideration to build AI tools that empowers users to engage and take responsibility towards moral duties. Future research on AI assisted counterspeech should focus on guiding users to be deliberate in their choices to engage and to not overrely on AI systems to make moral choices. Design methods to reduce overreliance such as cognitive forcing function such as checklists [11] or explanations about its outputs [135] should be explored and integrated into design of such AI systems. Furthermore, transparency about failures of AI systems, notably encoded biases [39, 139] and limited cultural context [109], would also be an important feature as AI generated text can instill values and norms [70, 78]. In addition to building unbiased and culturally aware AI systems [92], customization could help promote user agency and correct these shortcomings [74, 134]. However, customization can lead to more cognitive effort, so future AI counterspeech tools should consider effort-agency tradeoffs and explore distinctions between meaningful and non-meaningful effort [69].

5.4 Protecting Mental Health

The stress of responding and not caring enough were frequently discussed as reasons behind participants' choices to not engage. Experienced counterspeakers also noted that constant exposure to hate speech can lead to feelings of being overwhelmed and it becoming too much (Section 4.1). This is consistent with literature on the experience of cyberbullying victimization, becoming the target of cyberharassment or hate speech, and its link to mental health, especially in relation to depression, anxiety, and social media fatigue [16, 72, 117]. Exposure to hate speech can especially be damaging for the minority groups being targeted and can lead to depression, suicidal thoughts, and stress [151]. Some recipients of counterspeech (i.e., speaker behind hate speech) are not intentionally saying harmful things and could be willing to change. Therefore, call-out based [104] online shaming [131] and domination [141] which could escalate conflict or lead to harm [85] should be avoided.

Thus, any AI counterspeech tool should take into consideration mental health of those involved to empower and not harm users. One recommendation would be to design tools focusing on mental health such as human-AI collaboration for empathetic conversations [125] and focus on guiding systems to reflect a call-in culture that encourages relating to others in affirming ways rather than shaming [104], however, in excess, this could lead to reducing the opportunity for people to bring more authentic emotions, including negative ones, into conversations [65]. Therefore, further research is needed to understand how to best support users in generating authentic yet empathetic and emotionally-aware responses to counter hate speech. Moreover, while ease of use is important for counterspeech tools as discussed in Section 4.2, the ease of counterspeaking may lead to more exposures to content that might be harmful, especially if these tools are finding and encouraging users to respond or increasing the speed of hate and counter hate interactions. To mitigate this issue, design methods to encourage mindfulness such as design frictions that intentionally slows down interaction [30, 89] and reflective designs to encourage reflection on intentions [121, 123] should be further studied to find balance between efficiency and mindfulness in context of counterspeech assistance and be integrated to help users be more mindful.

5.5 Bias in AI Against Marginalized Groups

Our participants also highlighted biases of AI against minorities as one of the concerns (Section 4.3.2), which numerous studies have pointed out as a key issue in AI systems [7, 9, 15]. AI systems, especially language models, are often trained on Western-centric data [109], and often exclude and filter out marginalized identities during preprocessing [37]. These opaque design decisions can thus negatively impact marginalized communities, for example, through disparate performance [51, 110, 148] and operationalized stereotypes [7, 23].

As hate speech often targets marginalized communities [133], mitigating AI biases is a key consideration for designing a system to alleviate its impact. Therefore, methods to address this critical challenge should be further explored when building AI counterspeech tools. One strategy could involve amplifying marginalized voices and reflecting more diverse values through different aggregation or careful data collection [49, 112, 128]. Furthermore, ongoing efforts to audit biases in AI systems [7, 23, 87] can surface different biases and enhance their mitigation. Additionally, it's crucial to delve deeper into the multidimensionality of identities [60] and their representation [22] concerning AI systems and downstream tasks.

6 LIMITATIONS AND FUTURE WORK

Survey Participant Demographics. Both our survey and interview studies and their analysis are limited to the answers from our participants. Our survey study was conducted with North American (U.S. and Canada) participants and were overwhelmingly from United States. Moreover, due to our choice of platform (MTurk), our results might not be representative of populations that are less familiar with technology. The demographics of our survey participants was also skewed especially in racial and sexual identities as more than 80% identified as white or Caucasian and heterosexual. While our interview participants were from more diverse geographic regions, the interviews were all conducted in English, limiting our results to English-speaking countries. Therefore, our results may not generalize to populations outside the ones listed. These demographic limitations are especially important to be explored by future works given the disproportionate effect of technology on minority groups [114, 115] and frequent targeting of minority groups in hate speech [102, 107].

Definition of AI. To focus on collecting diverse ideas for AI usages, we did not restrict the definition of AI when asking participants to envision AI tools. However, this could have lead to different understanding of what constitutes AI, especially based on participants' familiarity with AI [86]. While collecting levels of experience with AI systems from the participants could have provided more insight into their answers, our work did not cover the interaction between different levels of experience with AI and differences in participant answers. This intersection of AI experiences and envisioning of AI tools would be an interesting future work.

Impact of AI and Counterspeech on Minority Groups. In addition to expanding our study to more diverse demographics, future works should consider studying the specific impact of AI-driven counterspeech systems on marginalized communities. AI systems in general are known to exacerbate societal biases [124] and lead to further marginalization and amplification of existing structural inequalities [31, 110]. This risk is particularly salient in AI counterspeech settings which aim to help and not further victimize hate-targeted communities, and there has already been evidence of backfiring of AI systems (e.g., refusing to talk about race [115] or misgendering the user [114]). Therefore, continued effort in developing participatory methods to ethically engage different marginalized communities and intersections of such communities [54, 61, 127] are essential.

Limited Context of Counterspeech. Future works could also explore counterspeech with a more global lens to understand a wider set of barriers and their interaction with AI assistance. Especially under authoritarian contexts, counterspeech, typically thought of as combating hate, could be used to suppress speech or dissidents [52]. Therefore, understanding the positionality of counterspeakers and their cultural contexts would be an important area of future work and an important ethical consideration toward understanding dual use of AI-driven counterspeech tool. Additionally, our work focused on text-based counterspeech largely agnostic to choice of community or online platform; future works could explore various modalities, platforms (e.g., TikTok), and communities for counterspeech. We also scoped our current work's focus on countering hate speech. However, other forms of online harm, such as fake news, could also be addressed by counterspeech. Future works could explore countering fake news and other forms of harm through AI tools. Additionally, in this paper, we lay out several design considerations that should be explored by future works. An iterative design process should take place to implement these considerations to co-design a counterspeech tool to empower and support users.

Limitation of Methods. While we used several measures to ensure quality of answers (Appendix A.1), due to decentralized nature of crowdsourcing-based studies, it is difficult to guarantee that data came from reliable and expected sources. Further, crowdworkers may use AI-based tools such as ChatGPT to perform annotation [136], which can be difficult to distinguish from human responses [28]. Therefore, in human annotation of bot-like responses, we could have allowed both false negatives and false positives, resulting in limitations in the internal validity of the data.

7 CONCLUSION

This work explored the experiences, needs, and concerns of activist counterspeakers and lay participants towards participatory AI for counterspeech. Our findings surfaced a theory of counterspeaking process, along with barriers at each step and motivations that drive this process. Our work highlighted the tension between the barriers (e.g., limited resources, lack of training, unclear impact, and personal harms) and motivations (e.g., moral duty and positive impact) and several ways that AI tool could help lower the barriers to counterspeech. Furthermore, we also surfaced concerns over the use of AI tools for counterspeech in authenticity, agency, and functional doubts.

Our findings reveal a considerable gap between current direction of research for AI assistance in counterspeech and an empowering assistive tool for users as the negative impact of AI involvement are not fully considered or addressed. To close this gap, we make several design recommendations connecting our findings to previous works to inform an empowering, user-focused, design of counterspeech AI tools. We provide recommendations focusing on transparency to build trust and authenticity in online communication, on design methods to encourage deliberation and moral agency, and on mindful designs to promote mental health. Our discussion also raises questions about how to best reduce effort and barriers of counterspeech without detracting from meaningful communication and connection. Thus, our work calls for further exploration and co-design of AI tools for counterspeech that addresses the participants' concerns to empower users in building safer and healthier online spaces.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their careful consideration and detailed feedback to improve our work. We thank the interview participants and MTurk workers for their contribution to counterspeech and our work. Additionally, we are thankful to Kathleen Fraser, Isar Nejadgholi, and Svetlana Kiritchenko for fruitful discussion.

REFERENCES

[1] Carolina Are. 2022. The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram. 22, 8 (2022), 2002–2019. https://doi.org/10.1080/14680777.2021.1928259

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

[3] Michael Baggs. 2021. Online hate speech rose 20% during pandemic: 'We've normalised it'. BBC (Nov. 2021).

[4] Fabienne Baider. 2023. Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech. 11, 2 (2023). https://www.cogitatiopress.com/politicsandgovernance/article/view/6465

[5] Susan Benesch, Derek Ruths, Haji Mohammad Saleem, Kelly P. Dillon, and Lucas Wright. 2016. Considerations for Successful Counterspeech. https://doi.org/10.15868/socialsector.34065

[6] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. Proceedings of the ACM on Human-Computer Interaction 1, CSCW (Dec. 2017), 24:1–24:19. https://doi.org/10.1145/3134659

[7] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

[8] Mark Blythe, Kristina Andersen, Rachel Clarke, and Peter Wright. 2016. Anti-Solutionist Strategies: Seriously Silly Design Fiction. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (San Jose California USA, 2016-05-07). ACM, 4968–4978. https://doi.org/10.1145/2858036.2858482

[9] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs.CL]

[10] Thomas Breideband, Jeffrey Bush, Chelsea Chandler, Michael Chang, Rachel Dickler, Peter Foltz, Ananya Ganesh, Rachel Lieber, William R. Penuel, Jason G. Reitman, John Weatherley, and Sidney D'Mello. 2023. The Community Builder (CoBi): Helping Students to Develop Better Small Group Collaborative Learning Skills. In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (Minneapolis, MN, USA) (CSCW '23 Companion). Association for Computing Machinery, New York, NY, USA, 376–380. https://doi.org/10.1145/3584931.3607498

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[12] Catherine Buerger. 2021. #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. 7, 4 (2021), 20563051211063843. https://doi.org/10.1177/20563051211063843

[13] Catherine Buerger. 2021. Speech as a Driver of Intergroup Violence: A Literature Review. https://doi.org/10.2139/ssrn.4066876

[14] Catherine Buerger. 2022. Why They Do It: Counterspeech Theories of Change. https://doi.org/10.2139/ssrn.4245211

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[16] Xiongfei Cao, Ali N. Khan, Ghulam H. K. Zaigham, and Naseer A. Khan. 2019. The Stimulators of Social Media Fatigue Among Students: Role of Moral Disengagement. Journal of Educational Computing Research 57, 5 (2019), 1083–1107. https://doi.org/10.1177/0735633118781907 arXiv:https://doi.org/10.1177/0735633118781907

[17] Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, Social Media and Online Hate Speech. Systematic Review. 58 (2021), 101608. https://doi.org/10.1016/j.avb.2021.101608

[18] Jordi Castellví, Mariona Massip Sabater, Gustavo A. González-Valencia, and Antoni Santisteban. 2022. Future teachers confronting extremism and hate speech. Humanities and Social Sciences Communications 9, 1 (15 Jun 2022), 201. https://doi.org/10.1057/s41599-022-01222-4

[19] Pew Research Center. 2023. Public Awareness of Artificial Intelligence in Everyday Activities.

[20] Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. Philosophy Compass 18, 1 (2023), e12890. https://doi.org/10.1111/phc3.12890

[21] K. Charmaz. 2006. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. SAGE Publications. https://books.google.com/books?id=v1qP1KbXz1AC

[22] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. arXiv:2101.03827 [cs.CY]

[23] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1504–1532. https://doi.org/10.18653/v1/2023.acl-long.84

[24] Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding Counterspeech for Online Harm Mitigation. arXiv:2307.04761 [cs.CL]

[25] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 2819–2829. https://doi.org/10.18653/v1/P19-1271

[26] Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021. Empowering NGOs in countering online hate messages. Online Social Networks and Media 24 (2021), 100150. https://doi.org/10.1016/j.osnem.2021.100150

[27] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. BUL Rev. 91 (2011), 1435.

[28] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[29] Jennifer Cobbe. 2021. Algorithmic Censorship by Social Platforms: Power and Resistance. 34, 4 (dec 2021), 739–766. https://doi.org/10.1007/s13347-020-00429-0

[30] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (San Jose, California, USA) (CHI EA '16). Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[31] KATE CRAWFORD. 2021. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press. http://www.jstor.org/stable/j.ctv1ghv45t

[32] Stefano Cresci. 2020. A Decade of Social Bot Detection. Commun. ACM 63, 10 (sep 2020), 72–83. https://doi.org/10.1145/3409116

[33] John W Creswell and Vicki L Plano Clark. 2017. Designing and conducting mixed methods research. Sage publications.

[34] Niklas Felix Cypris, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. 2022. Intervening against online hate speech: A case for automated Counterspeech. IEAI Research Brief (2022), 1–8.

[35] John Danaher. 2019. The rise of the robots and the crisis of moral patiency. AI & SOCIETY 34, 1 (01 Mar 2019), 129–136. https://doi.org/10.1007/s00146-017-0773-9

[36] Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate Speech in Online Social Media. SIGWEB Newsl. 2020, Autumn, Article 4 (nov 2020), 8 pages. https://doi.org/10.1145/3427478.3427482

[37] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

[38] Oren Etzioni. 2017. Opinion | How to Regulate Artificial Intelligence. (09 2017). https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html

[39] Mirko Farina and Andrea Lavazza. 2023. ChatGPT in society: emerging issues. Frontiers in Artificial Intelligence 6 (2023). https://doi.org/10.3389/frai.2023.1130913

[40] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. Commun. ACM 59, 7 (jun 2016), 96–104. https://doi.org/10.1145/2818717

[41] Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. Comput. Surveys 51, 4 (July 2018), 85:1–85:30. https://doi.org/10.1145/3232676

[42] Dennis Friess, Marc Ziegele, and Dominique Heinbach. 2021. Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. 38, 5 (2021), 624–646. https://doi.org/10.1080/10584609.2020.1830322

[43] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In Proceedings of the Fourth Workshop on Online Abuse and Harms. Association for Computational Linguistics, Online, 102–112. https://doi.org/10.18653/v1/2020.alw-1.13

[44] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. EPJ Data Science 11, 1 (24 Jan 2022), 3. https://doi.org/10.1140/epjds/s13688-021-00314-6

[45] Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey. arXiv:2310.15264 [cs.CL]

[46] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. Big Data & Society 7, 2 (July 2020), 2053951720943234. https://doi.org/10.1177/2053951720943234

[47] B.G. Glaser and A.L. Strauss. 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Transaction. https://books.google.com/books?id=oUxEAQAAIAAJ

[48] Ella Glikson and Omri Asscher. 2023. AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. Computers in Human Behavior 140 (2023), 107592. https://doi.org/10.1016/j.chb.2022.107592

[49] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In CHI Conference on Human Factors in Computing Systems (CHI '22). ACM. https://doi.org/10.1145/3491102.3502004

[50] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. "You Have to Prove the Threat is Real": Understanding the Needs of Female Journalists and Activists to Document and Report Online Harassment. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 242, 17 pages. https://doi.org/10.1145/3491102.3517517

[51] Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. arXiv:2010.02510 [cs.CL]

[52] Seva Gunitsky. 2015. Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability. Perspectives on Politics 13, 1 (2015), 42–54. https://doi.org/10.1017/S1537592714003120

[53] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5792–5809. https://doi.org/10.18653/v1/2023.acl-long.318

[54] Oliver L. Haimson, Kai Nham, Hibby Thach, and Aloe DeGuia. 2023. How Transgender People and Communities Were Involved in Trans Technology Design Processes. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 294, 16 pages. https://doi.org/10.1145/3544548.3580972

[55] Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. WokeGPT: Improving Counterspeech Generation Against Online Hate Speech by Intelligently Augmenting Datasets Using a Novel Metric. In 2023 International Joint Conference on Neural Networks (IJCNN) (2023-06). 1–10. https://doi.org/10.1109/IJCNN54540.2023.10191114

[56] Soo-Hye Han, LeAnn M. Brazeal, and Natalie Pennington. 2018. Is Civility Contagious? Examining the Impact of Modeling in Online Political Discussions. 4, 3 (2018), 2056305118793404. https://doi.org/10.1177/2056305118793404

[57] Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying Toxic Speech Detectors Against Veiled Toxicity. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 7732–7739. https://doi.org/10.18653/v1/2020.emnlp-main.622

[58] 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 (01 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf

[59] 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. https://doi.org/10.1073/pnas.2116310118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2116310118

[60] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[61] Christina N. Harrington, Aashaka Desai, Aaleyah Lewis, Sanika Moharana, Anne Spencer Ross, and Jennifer Mankoff. 2023. Working at the Intersection of Race, Disability and Accessibility. In Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (New York, NY, USA) (ASSETS '23). Association for Computing Machinery, New York, NY, USA, Article 26, 18 pages. https://doi.org/10.1145/3597638.3608389

[62] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: Controlling Language Models to Generate Implied and Adversarial Toxicity. In ACL. https://arxiv.org/abs/2203.09509

[63] Sabit Hassan and Malihe Alikhani. 2023. DisCGen: A Framework for Discourse-Informed Counterspeech Generation. arXiv:2311.18147 [cs.CL]

[64] Jess Hohenstein and Malte 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. https://doi.org/10.1016/j.chb.2019.106190

[65] Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. Scientific Reports 13, 1 (April 2023), 5487.

[66] Jonathan Horowitz. 2017. Who Is This "We" You Speak of? Grounding Activist Identity in Social Psychology. Socius 3 (2017), 2378023117717819. https://doi.org/10.1177/2378023117717819 arXiv:https://doi.org/10.1177/2378023117717819 PMID: 30221196.

[67] Jeffrey W Howard. 2021. Terror, Hate and the Demands of Counter-Speech. Br. J. Polit. Sci. 51, 3 (July 2021), 924–939.

[68] Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, and Mustafa Naseem. 2022. Women's Perspectives on Harm and Justice after Online Harassment. Proceedings of the ACM on Human-Computer Interaction 6, CSCW2 (Nov. 2022), 355:1–355:23. https://doi.org/10.1145/3555775

[69] Michael Inzlicht and Aidan V. Campbell. 2022. Effort feels meaningful. Trends in Cognitive Sciences 26, 12 (2022), 1035–1037. https://doi.org/10.1016/j.tics.2022.09.016

[70] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM. https://doi.org/10.1145/3544548.3581196

[71] Maurice Jakesch, Megan 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 (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[72] Cristina Jenaro, Noelia Flores, and Cinthia Patricia Frías. 2018. Systematic review of empirical studies on cyberbullying in adults: What we know and what we should investigate. Aggression and Violent Behavior 38 (2018), 113–122. https://doi.org/10.1016/j.avb.2017.12.003

[73] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3658–3666. https://doi.org/10.18653/v1/P19-1357

[74] Hyunjin Kang and Chen Lou. 2022. AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement. Journal of Computer-Mediated Communication 27, 5 (08 2022), zmac014. https://doi.org/10.1093/jcmc/zmac014 arXiv:https://academic.oup.com/jcmc/article-pdf/27/5/zmac014/45473652/zmac014.pdf

[75] David Kaye. 2019. Speech Police: The Global Struggle to Govern the Internet. Columbia Global Reports. http://www.jstor.org/stable/j.ctv1fx4h8v

[76] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. The shape of and solutions to the MTurk quality crisis. Political Science Research and Methods 8, 4 (2020), 614–629. https://doi.org/10.1017/psrm.2020.6

[77] Barbara A Kitchenham and Shari L Pfleeger. 2008. Personal opinion surveys. In Guide to advanced empirical software engineering. Springer, 63–92.

[78] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports 13, 1 (April 2023), 4569.

[79] Enes Kulenović. 2022. Should Democracies Ban Hate Speech? Hate Speech Laws and Counterspeech. Ethical Theory and Moral Practice (Nov. 2022). https://doi.org/10.1007/s10677-022-10336-2

[80] Samuli Laato, A. K. M. Najmul Islam, Muhammad Nazrul Islam, and Eoin Whelan. 2020. What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? European Journal of Information Systems 29, 3 (2020), 288–305. https://doi.org/10.1080/0960085X.2020.1770632 arXiv:https://doi.org/10.1080/0960085X.2020.1770632

[81] Rae Langton. 2018. 144Blocking as Counter-Speech. In New Work on Speech Acts. Oxford University Press. https://doi.org/10.1093/oso/9780198738831.003.0006 arXiv:https://academic.oup.com/book/0/chapter/155957982/chapter-ag-pdf/44951161/book_9256_section_155957982.ag.pdf

[82] Anti-Defamation League. 2023. Online Hate and Harassment: The American Experience. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023

[83] Kalev Leetaru. 2019. Online Toxicity Is As Old As The Web Itself But The Return To Communities May Help. Forbes Magazine (May 2019).

[84] Olivier Lemeire. 2021. Falsifying generic stereotypes. Philosophical Studies 178, 7 (01 Jul 2021), 2293–2312. https://doi.org/10.1007/s11098-020-01555-3

[85] Maxime Lepoutre. 2022. Hateful Counterspeech. Ethical Theory and Moral Practice (27 Oct 2022). https://doi.org/10.1007/s10677-022-10323-7

[86] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[87] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. https://openreview.net/forum?id=qVXYU3F017

[88] Sana Maqsood and Sonia Chiasson. 2021. "They Think It's Totally Fine to Talk to Somebody on the Internet They Don't Know": Teachers' Perceptions and Mitigation Strategies of Tweens' Online Risks. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 688, 17 pages. https://doi.org/10.1145/3411764.3445224

[89] Teale W. Masrani, Jack Jamieson, Naomi Yamashita, and Helen Ai He. 2023. Slowing It Down: Towards Facilitating Interpersonal Mindfulness in Online Polarizing Conversations Over Social Media. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 90 (apr 2023), 27 pages. https://doi.org/10.1145/3579523

[90] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate Begets Hate: A Temporal Study of Hate Speech. Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 92 (oct 2020), 24 pages. https://doi.org/10.1145/3415163

[91] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. ICWSM 13 (July 2019), 369–380. http://arxiv.org/abs/1808.04409

[92] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[93] Jozef Miškolci, Lucia Kováčová, and Edita Rigová. 2020. Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. 38, 2 (April 2020), 128–146. https://doi.org/10.1177/0894439318791786

[94] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (Prague, Czech Republic) (HT '17). Association for Computing Machinery, New York, NY, USA, 85–94. https://doi.org/10.1145/3078714.3078723

[95] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. arXiv:2311.00161 [cs.CL]

[96] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society 20, 11 (2018), 4366–4383.

[97] Dawn Carla Nunziato. 2021. The Varieties of Counterspeech and Censorship on Social Media Symposium: Cheap Speech Twenty-Five Years Later: Democracy & Public Discourse in the Digital Age. 54, 5 (2021), 2491–2552. https://heinonline.org/HOL/P?h=hein.journals/davlr54&i=2509

[98] Christina Pan, Sahil Yakhmi, Tara Iyer, Evan Strasnick, Amy Zhang, and Michael Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. Proc. ACM Hum.-Comput. Interact. CSCW (Oct. 2022).

[99] Bhikhu Parekh. 2012. Is There a Case for Banning Hate Speech? In The Content and Context of Hate Speech: Rethinking Regulation and Responses, Michael Herz and Peter Molnar (Eds.). Cambridge University Press, Cambridge, 37–56. https://doi.org/10.1017/CBO9781139042871.006

[100] Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? Proceedings of the National Academy of Sciences 120, 10 (2023), e2209384120. https://doi.org/10.1073/pnas.2209384120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2209384120

[101] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In Proceedings of the 2016 ACM International Conference on Supporting Group Work (GROUP '16). Association for Computing Machinery, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297

[102] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate Speech: A Systematized Review. SAGE Open 10, 4 (2020), 2158244020973022. https://doi.org/10.1177/2158244020973022 arXiv:https://doi.org/10.1177/2158244020973022

[103] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation 55, 2 (01 Jun 2021), 477–523. https://doi.org/10.1007/s10579-020-09502-8

[104] Loretta J Ross. 2019. Speaking up without tearing down. Teaching Tolerance 61 (2019), 19–22.

[105] Konrad Rudnicki, Heidi Vandebosch, Pierre Voué, and Karolien Poels. 2023. Systematic Review of Determinants and Consequences of Bystander Interventions in Online Hate and Cyberbullying among Adults. Behaviour & Information Technology 42, 5 (April 2023), 527–544. https://doi.org/10.1080/0144929X.2022.2027013

[106] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? arXiv:2303.11156 [cs.CL]

[107] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the Rise of Fear Speech in Online Social Media. 120, 11 (march 2023), e2212270120. https://doi.org/10.1073/pnas.2212270120

[108] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. (May 2022). arXiv:2205.04304 [cs.CL]

[109] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, 9080–9102. https://doi.org/10.18653/v1/2023.acl-long.505

[110] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In ACL. https://www.aclweb.org/anthology/P19-1163.pdf

[111] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

[112] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. arXiv:2111.07997 [cs.CL]

[113] Julia Sasse and Jens Grossklags. 2023. Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content. Proc. ACM Hum.-Comput. Interact. 7, CSCW2, Article 327 (oct 2023), 26 pages. https://doi.org/10.1145/3610176

[114] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 144 (nov 2019), 33 pages. https://doi.org/10.1145/3359246

[115] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173889

[116] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 370 (nov 2022), 27 pages. https://doi.org/10.1145/3555095

[117] Kaitlyn B. Schodt, Selena I. Quiroz, Brittany Wheeler, Deborah L. Hall, and Yasin N. Silva. 2021. Cyberbullying and Mental Health in Adults: The Moderating Role of Social Media Use and Gender. Frontiers in Psychiatry 12 (2021). https://doi.org/10.3389/fpsyt.2021.674298

[118] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 485, 16 pages. https://doi.org/10.1145/3544548.3581020

[119] Judith Schoonenboom and R. Burke Johnson. 2017. How to Construct a Mixed Methods Research Design. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 69, 2 (01 Oct 2017), 107–131. https://doi.org/10.1007/s11577-017-0454-1

[120] Maxie Schulte, Sebastian Bamberg, Jonas Rees, and Philipp Rollin. 2020. Social identity as a key concept for connecting transformative societal change with individual environmental activism. Journal of Environmental Psychology 72 (2020), 101525. https://doi.org/10.1016/j.jenvp.2020.101525

[121] Ava Elizabeth Scott. 2023. To Do or Not To Do? Managing Intentions with Technology. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 504, 7 pages. https://doi.org/10.1145/3544549.3577046

[122] Joseph Seering. 2020. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 107 (oct 2020), 28 pages. https://doi.org/10.1145/3415178

[123] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. 2005. Reflective Design. In Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility (Aarhus, Denmark) (CC '05). Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[124] Rifat Ara Shams, Didar Zowghi, and Muneera Bano. 2023. AI and the quest for diversity and inclusion: a systematic literature review. AI and Ethics (13 Nov 2023). https://doi.org/10.1007/s43681-023-00362-w

[125] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence 5, 1 (01 Jan 2023), 46–57. https://doi.org/10.1038/s42256-022-00593-2

[126] Jana Siebert and Johannes Ulrich Siebert. 2023. Effective mitigation of the belief perseverance bias after the retraction of misinformation: Awareness training and counter-speech. PLOS ONE 18, 3 (03 2023), 1–22. https://doi.org/10.1371/journal.pone.0282202

[127] Angela D. R. Smith, Alex A. Ahmed, Adriana Alvarado Garcia, Bryan Dosono, Ihudiya Ogbonnaya-Ogburu, Yolanda Rankin, Alexandra To, and Kentaro Toyama. 2020. What's Race Got To Do With It? Engaging in Race in HCI. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375156

[128] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070 [cs.AI]

[129] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. 2016. Artificial Intelligence and Life in 2030. http://ai100.stanford.edu/2016-report

[130] Barry Stricke. 2019. People v. Robots: A Roadmap for Enforcing California's New Online Bot Disclosure Act. Vanderbilt Journal of Entertainment & Technology Law 22, 4 (2019), 839–894.

[131] Krista Thomason. 2021. The Moral Risks of Online Shaming. In Oxford Handbook of Digital Ethics. Oxford University Press.

[132] Stefanie Ullmann and Marcus Tomalin. 2023. Counterspeech: Multidisciplinary Perspectives on Countering Dangerous Speech. Taylor & Francis.

[133] United Nations, Human Rights Council. 2021. Recommendations made by the Forum on Minority Issues at its thirteenth session on the theme "Hate speech, social media and minorities". Human Rights Council, Forty-sixth session, Agenda item 5. Available from https://undocs.org/A/HRC/46/58.

[134] Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. 2023. Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations. In 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). 1–7. https://doi.org/10.1109/HORA58378.2023.10156761

[135] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[136] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv preprint arXiv:2306.07899 (2023).

[137] Emily A Vogels. 2021. The state of online harassment. Pew Research Center 13 (2021), 625.

[138] John Frank Weaver. 2018. We Need the California Bot Bill, but We Need It to Be Better Everything Is Not Terminator. RAIL: The Journal of Robotics, Artificial Intelligence & Law 1, 6 (2018), [vi]–438. https://heinonline.org/HOL/P?h=hein.journals/rail1&i=444

[139] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. CoRR abs/2112.04359 (2021). arXiv:2112.04359 https://arxiv.org/abs/2112.04359

[140] Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. Proceedings of the International AAAI Conference on Web and Social Media 16, 1 (May 2022), 1121–1132. https://doi.org/10.1609/icwsm.v16i1.19363

[141] Suzanne Whitten. 2023. A Republican Conception of Counterspeech. Ethical Theory and Moral Practice (28 Jul 2023). https://doi.org/10.1007/s10677-023-10409-w
[142] Claudia Wilhelm and Sven Joeckel. 2019. Gendered Morality and Backlash Effects in Online Discussions: An Experimental Study on How Users Respond to Hate Speech Comments Against Women and Sexual Minorities. Sex Roles 80, 7 (April 2019), 381–392. https://doi.org/10.1007/s11199-018-0941-5
[143] Mickie Wong-Lo and Lyndal M. Bullock. 2014. Digital Metamorphosis: Examination of the Bystander Culture in Cyberbullying. Aggression and Violent Behavior 19, 4 (July 2014), 418–422. https://doi.org/10.1016/j.avb.2014.06.007
[144] Tianyi Xie and Renee V. Galliher. 2023. Responding to Microaggressions: Social Cost of Bystander Intervention Strategies. The Counseling Psychologist 51, 2 (2023), 242–269. https://doi.org/10.1177/00110000221140482 arXiv:https://doi.org/10.1177/00110000221140482
[145] Guobin Yang. [n. d.]. Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter. 4, 472 ([n. d.]), 13–17. https://repository.upenn.edu/handle/20.500.14332/2135
[146] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2019. Scalable and Generalizable Social Bot Detection through Data Selection. In AAAI Conference on Artificial Intelligence. https://api.semanticscholar.org/CorpusID:208202136
[147] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Seattle, United States, 2022-07). Association for Computational Linguistics, 5918–5930. https://doi.org/10.18653/v1/2022.naacl-main.433
[148] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, and Emily Alsentzer. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health 6, 1 (2024), e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X
[149] Wanzheng Zhu and Suma Bhat. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. (June 2021). arXiv:2106.01625 [cs.CL]
[150] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. CoRR abs/2005.12423 (2020). arXiv:2005.12423 https://arxiv.org/abs/2005.12423
[151] Kamila Zochniak, Oliwia Lewicka, Zuzanna Wybrańska, and Michał Bilewicz. 2023. Homophobic Hate Speech Affects Well-Being of Highly Identified LGBT People. Journal of Language and Social Psychology 42, 4 (2023), 453–463. https://doi.org/10.1177/0261927X231174569 arXiv:https://doi.org/10.1177/0261927X231174569

A SURVEY

A.1 Response Quality Checks

Pre-qualification Process. To ensure the quality of our results, we used a pre-qualification process to prevent fraudulent responses. The pre-qualification process included three questions:

(1) Humans are mammals.
How true do you think is the above statement?
(2) People are right handed.
What percentage of people do you think are right handed?
(3) Penguins can't fly.
What percentage of penguins do you think can't fly?

The workers answer using a slider with percentage ranging 0 to 100 or 11 point likert scale. We accept the answer to correct for each question if they answer 1) 10, 2) greater than or equal to 50%, and 3) 10. We consider a worker qualified if they score 3 on this task. The workers were paid 0.22 USD for the qualification task.

Bot Detection. We used Google's reCaptcha V2[1] and V3[2]. Moreover, one of the authors manually annotated answers for bot-like behaviors looking for responses that were repetitive, off-topic, or non-sensical.

[1]https://developers.google.com/recaptcha/docs/display
[2]https://developers.google.com/recaptcha/docs/v3

A.2 Participant Demographics

Participant demographics are shown in Table 4. We asked participant's age, race, transgender identity, gender identity, sexuality, religion, political leaning, education, and country of residence. Our participants were largely from the U.S. and white with many having bachelor's degree or some college experience. As we filtered for north American (U.S. and Canada) residents on MTurk, other countries of residence might indicate erroneous reporting.

B ANALYSIS RESULTS

The codes developed from participant responses following the methods in Sections 3.3.2 and 3.3.3 are listed in Tables 5, 6, and 7. The high level themes as discussed in An Inductive Theory of the Counterspeaking Process and Motivation, Section 4.1 and 4.3 are indicated using specific icons for visibility. Counterspeech barriers could be categorized to high level themes of *limited resources*, *lack of training*, *unclear impact*, and *personal harms*. Moreover, motivations are mapped to intrinsic motivation of *moral duty* and extrinsic, *positive impact*. Some concerns about AI were related to motivations and are indicated using the same icon. However, some additional themes emerged in long-term concerns and functional doubts.

(a) Age

Option	Response (%)
18-24 years old	02.9240
25-34 years old	34.7953
35-44 years old	34.5029
45-54 years old	14.3275
55-64 years old	09.9415
65+ years old	02.9240
Prefer not to disclose	00.5848

(b) Race

Option	Response (%)
White or Caucasian	82.6979
White or Caucasian,Asian,Native Hawaiian or Other Pacific Islander	00.2933
White or Caucasian,Asian	01.7595
White or Caucasian,Other	00.5865
White or Caucasian,American Indian/Native American or Alaska Native	01.1730
White or Caucasian,Black or African American	00.2933
White or Caucasian,Black or African American,Other	00.2933
White or Caucasian,Native Hawaiian or Other Pacific Islander	00.2933
American Indian/Native American or Alaska Native	00.5865
Asian	05.8651
Black or African American	04.6921
Prefer not to say	00.8798
Other	00.5865

(c) Transgender

Option	Response (%)
Yes	02.3460
No	96.4809
Prefer not to disclose	01.1730

(d) Gender

Option	Response (%)
Man	56.4327
Woman	41.2281
Two-spirit,Woman	00.2924
Genderqueer or gender fluid	00.2924
Additional gender category/identity	00.2924
Prefer not to disclose	01.4620

(e) Sexuality

Option	Response (%)
Straight (heterosexual)	85.0877
Bisexual	06.4327
Bisexual,Pansexual	00.5848
Asexual	02.0468
Asexual,Straight (heterosexual)	00.2924
Lesbian	01.1696
Gay	01.1696
Pansexual	01.1696
Questioning or unsure	00.2924
Prefer not to disclose	01.7544

(f) Religion

Option	Response (%)
Atheist	16.9591
Christian	39.1813
Agnostic	16.3743
Catholic	15.4971
Jewish	01.1696
Buddhist	01.1696
Hindu	00.5848
Muslim	00.2924
Nothing in particular	04.3860
Prefer not to disclose	02.3392
Something else, Specify:	02.0468

(g) Political Leaning

Option	Response (%)
Strongly liberal	22.2222
Liberal	32.4561
Moderate	18.7135
Conservative	16.6667
Strongly conservative	08.4795
Prefer not to disclose	01.4620

(h) Education

Option	Response (%)
Bachelor's degree	57.6023
Some college, but no degree	13.1579
Graduate or professional degree	07.6023
High school diploma or GED	10.8187
Associates or technical degree	09.3567
Some high school or less	00.8772
Prefer not to say	00.5848

(i) Country of Residence

Option	Response (%)
United States of America	97.9472
Namibia	00.2933
Canada	00.5865
United Kingdom of Great Britain and Northern Ireland	00.2933
India	00.2933
Albania	00.2933
Argentina	00.2933

Table 4: Demographics of survey participants. All the heuristics are reported as percentages.

Code	Description	Representative Quote
Barriers to Counterspeech		
Resources	Financial or people resources are limited, especially time it takes to do counterspeech	*"[The biggest challenge is] time. We're doing a max already but for security reasons we cannot be too big. This is not our job."*
Finding hate speech	Takes time to find hate speech	*"It's so time consuming looking for articles."*
Training	People don't have the training	*"People don't always know what to say."*
Reach	It's hard to reach people, which is discouraging	*"When you feel unheard and it's like I'm doing this for nothing - it's not really getting the word out - it's frustrating"*
Risk	There are risks of online or offline attacks	*"And the risk – the risk is real. I don't use anonymous posting. In our community, there is fear and so it's risky when you put yourself out there."*
Mental health	It is too hard on mental health (stress or boredom)	*"You are just overwhelmed with what you are seeing."*
Counterspeaker Motivations		
Right	It's the right thing to do / civic responsibility	*"I think it's because it's the right thing to do – I feel that at least I tried. "*
Impact	Seeing evidence of successful impact	*"When you can see the comment section change. When you can see other non-members speaking out. We take screen shots and save our successes."*
Scale - small	Quotes about the individual-level impact	*"Probably when we get in real time that we've helped someone, helped someone who had maybe been reading the comments and had been upset by them, they say something like, 'oh thank goodness, I was in a pit of dispair before seeing your comment."'*

Table 5: Codes developed from analysis of interview responses discussing counterspeech barriers and motivations.

Code	Description	Representative Quote
Activist Counterspeakers		
Time	References to time or efficiency	*"Anything that could help us be more efficient - help us produce more."*
Finding - AI	AI would help by locating hateful speech	*"I've spent up to two hours looking for good actions, so that would be super helpful. "*
Scale - big	AI would help scale the work of counterspeakers	*"A tool that would amplify voice against hate speech. That would assist in amplifying counterspeech and helping it reach the target audience."*
Lay-users		
Efficiency	The user wants to use the tool to save time.	*"It would be make responding to hatred so much easier if I could just click a few boxes and let an AI do the work. Even though it won't change the hater's heart, it would provide a counter to their hate speach."*
AI-better	The user thinks that AI would be better thant they are.	*"[An AI] could give me a constructive framework for a much more impactful response than I could otherwise generate on my own."*
Capability-dependent	The user wants to explore the capabilities of the tool before deciding.	*"I would be willing to see the suggestion that the AI offered and decide whether or not to use it."*
Information	The user thinks a tool would be helpful to compile information to counter hate.	*"if it was fact based i sure would use it, since i feel we all can have our own oppinions"*
Guidance	The user would use the tool to get guidance on how to respond effectively: formulating response and creating more diverse response in a more collaborative way, or to help them understand the hate or detection of hate.	*"I would potentially use it because it could give me a constructive framework for a much more impactful response than I could otherwise generate on my own"*
Emotions	The user wants help with regulating their emotions to communicate clearly or with effectively communicating user's emotions.	*"... It would help me to stay calm and collected. When I am faced with hateful speech, I can sometimes get emotional. This can make it difficult for me to respond effectively. The AI tool would help me to stay calm and collected, so that I could focus on responding to the hateful speech in a thoughtful and reasoned way. It would help me to feel more confident in my responses..."*
Empowerment	The user feels empowered by being able to speak up in addressing hate speech to create a positive impact.	*"Because it would help me speak up more."*
Reduce-stress	The user feels that having the tool could help reduce stress while responding to hate speech.	*"It might be less stressful to use than making a more personal comment."*
AI-proxy	The user would rather have the AI get involved, rather themselves (often under the guise of anonymity) either in responding or reporting.	*"To be honest it sort of makes me feel like I have some plausible deniability if an issue arises. In a worst case scenario I would be able to 'blame' it on the AI."*
Lay-users - AI Tools		""
Existing-technology	The user refers to existing technology a specific AI tool that current exists in the market.	*"I use most of the time ChatGPT."*
Report	The user would like a system that can automatically or with minimal input report hate speech.	*"[An AI] that identifies the hate speech and remove or block the comment."*
Response-support	The user wants an AI tool that suggests or automatically replies with a response to hate speech, which could also be refined by the user with minimal input that is well-written and thoughtful. The user usually wants efficient and time saving support with minimal engagement and is easy to use.	*"The one which suggest the reply in very decent manner."*
Factual	The user expresses that it would be beneficial to have an assistive technology that can gather factual information to formulate arguments or fact-check hate speech. The users also want help in creating responses that rational, intellectual, and logical arguments.	*"I would use it if it gave out information that was correct and if it was reliable."*
Collaborative	The user wishes to have more collaborative interaction to improve their responses. Some examples of interactions include correction to their written response such as grammar, emotional, or factual and checking their own biases.	*"Something I could be "checked" on, making sure MY post wasn't also toxic."*
Effective-communication	The user would like to use an AI tool that is sensitive to human emotions while addressing hate speech, and is capable of expressing nuanced emotions. The user wants support communicating clearly with the understanding of emotional, human factors focusing on meaning and impact.	*"An AI assistance that is nice and helps alleviate the situation."*
Aligned	The user would like an AI tool that is personally and/or culturally aligned and provide responses just like how they would or in an unbiased way.	*"One that is trained off my data and personality that I approve of."*
Protective	The user would like an AI tool that will protect them from retaliation often through anonimity.	*"I would like an AI tool that could prepare a message...[avoids] making myself a target."*

Table 6: Codes developed from analysis of responses showing openness to adopt AI tools in SQ14 and in the interview studie as well as responses to SQ15. The icons indicate the theme of barriers relevant to the code. The codes are separated into three sections: benefits identified by activist counterspeakers, benefits identified by lay-users, and AI tools discussed by lay-users. The colors of icons were chosen to match Figure 2.

Code		Description	Representative Quote
Activist CounterSpeakers			
Authenticity - strategy		Worries that inauthentic counterspeech would not be credible	*"The moment we deploy this online, a lot of people who share hateful content and know a lot about tech will recognize it."*
Real		Quotes comparing AI to what is "real"	*"We do not need, neither for us nor for the haters, the possibility to create a fake sentiment and take away our voices. It will boil down to who has the money to pay it more."*
Long-term		Concerns that AI counterspeech has long-term negative consequences	*"Really using the bot at all is tricky. You aren't inspiring real people to participate. If we are actually going to make change, we need those people to be engaged. We need people to get involved in their communities."*
Counterspeaker		Counterspeakers are aided by doing counterspeech	*"And there is something incredibly magic about turning something really hateful in the other direction. You feel you aren't hopeless or helpless."*
Becoming the monster		Troll farms are bad. Would we become just as bad by using a counter-speech bot?	*"I can see the appeal to that for sure, but I think that it takes out the human component. We're kind of no better than what is being used."*
Emotional Intelligence		Emotional intelligence, empathy, you need a human, authenticity	*"The process itself I find very satisfying. Having a sense of not being alone... For me, what's really touching is that someone out there is just there to support me. I'm not the only one who thinks this."*
Functionality - technical		Skepticism that AI counterspeech would work	*"I'm not sure about it getting the facts right. It's not good for fact checking."*
Lay-users			
Authenticity		The user expresses concern that what AI communicates is not their own words and would not represent what they are thinking especially their intentions (alignment) or would be considered inauthentic focusing on "who" is behind counterspeech.	*"Using AI is too impersonal and it sounds very generic."*
Engaging-not-helpful		The user believes engaging with people who espouse hate speech is not helpful in reducing that behavior or do not want hate speech getting more attention. These users sometimes express that they would engage if they knew that it would make an impact.	*"I don't think it matters if I get help with what I want to say if it's just falling on deaf ears."*
Avoidance		The user rather wishes to avoid hate speech rather than engage and availability of an AI tool will not change this.	*"I don't engage in any online hate/drama. I just scroll right through."*
Agency		The user does not want AI help especially because they are able to perform the task themselves. The user prefers humans to respond focusing on "how" counterspeech is generated.	*"I believe in stating things that I feel not what an AI tells me to feel."*
Capability-doubts		The user expresses that they do not think that AI response will have an impact or it will contain other functionality problems.	*"AI tools are often wrong and I wouldn't want it's bias' to affect what I am posting."*
Become-target		The user does not want to become the target of hate.	*"Having an AI help me write a response would not keep people from sending me hateful replies. I cannot handle that."*

Table 7: Codes from analysis of responses resistant to adopting AI tools from SQ14 and concerns raised by interview participants. The icons indicate relevant themes of motivations that are negatively affected and new themes of functionality doubts and long-term impact. The codes are presented in two sections: responses from activist counterspeakers and from lay-users.