

Protesting Robots? Designing Participatory AI for Counterspeech

JIMIN MUN, Language Technologies Institute, Carnegie Mellon University, USA

CATHY BUERGER, Dangerous Speech Project, USA

JENNY T. LIANG, Human-Computer Interaction Institute Carnegie Mellon University, USA

JOSHUA GARDLAND, Arizona State University, USA

MAARTEN SAP, Language Technologies Institute, Carnegie Mellon University, USA

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 take a participatory approach towards answering this question, by conducting in-depth interviews with 10 experienced counterspeakers and a large scale public survey with 342 inexperienced counterspeakers. 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 Gardland, and Maarten Sap. 2018. Protesting Robots? Designing Participatory AI for Counterspeech. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 25 pages. <https://doi.org/XXXXXXX.XXXXXXX>

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 [61]. Its goal are multifaceted; counterspeech aims to not only minimize harms of hateful speech, but also to promote positive change in online communities through open dialogue among users [10] and by fostering a sense of community [8]. 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) [8, 29, 30, 88, 91]. As hate speech prevalence grows in online spaces [3, 53, 67], coordinating and responding with counterspeech has become increasingly challenging [17, 18].

AI has recently emerged as a potential tool or solution [2, 24, 90] 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 [6]. On one hand, it is important to reduce the burden on counterspeakers [42], who are sometimes victims of hate speech themselves [13]. On the other hand, naive AI solutions that simply generate counterspeech without human oversight (e.g., “*protesting robots*”) could significantly detract from the authentic,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

empowering, and emotionally connecting experience that counterspeech provides to users [8, 29], despite being a growing area of research in AI [15, 68, 90].

We argue that a more deliberate approach with user participation is required to designing AI tools that can increase participation in counterspeech as well as empower counterspeakers. Thus, in this paper, we take the a participatory approach to understand users' 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 question:

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 counterspeakers as well as inexperienced bystanders. We conducted semi-structured long-form interviews with activists who regularly 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 their characteristics (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.

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 [26, 60]. With the scale and pervasiveness of cyberspaces, hate speech can be especially damaging to social groups as it fosters hostility and perpetuates stereotypes and can even lead to off-line violence [9, 12]. One of the most prevalent solution to the growing scale of hate speech has been content moderation through algorithmic censorship. Algorithmic censorship

has been shown to have many shortcomings [64] causing growing concerns over freedom of speech, especially with the algorithms tailored to serve the platform often excluding the users from its decisions [1, 20, 48]. Moreover, many automated hate speech detection has been shown to be inaccurate, unable to detect subtle hate speech [36] or biased against certain communities causing further marginalization [71, 72]. In addition to the new proposed methods of online governance, such as community moderation [77] and collective-decision moderation [63], counterspeech offers a way to counter hatred that preserves freedom of speech 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 [54], prevent the spread of misinformation [52], reinforce correct information [80], and to expand responses to even more covert hate speech [4]. However, it requires effort and engagement from the users, which raises the question of scale as hate speech increases.

Research on online counterspeech has been focused on social media interactions [59, 62] 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., deleting the hateful post) or those engaged in the discourse (e.g. sentiment in comments) and have shown empathy [38] and civility [35] to have some positive impact on these measures. However, positive effects of counterspeech is more diverse than given focus. The goals of engagement identified by counterspeakers were often multifaceted: to change the view of the bystanders, to recruit more counterspeakers, or to strengthen community norms [10]. Moreover, counterspeakers, who have not been the focus of many studies, sometimes respond to online hate speech collectively as a part of a more organized movements such as #iamhere [8, 29] creating connection and solidarity against hate. However, barriers to counterspeaking, more specifically, hurdles and reasoning behind decisions to not engage, have not yet been studied and require further investigation.

2.3 AI Assistance in Counterspeech

To address the issue of scale, focus of counterspeech assistance in AI has been on automatic generation of text countering hate. However, automatic detection [89] and computational analysis of large scale counterspeech [30] have been used to understand its characteristics and to inform effective content. Prior works on automatic generation of counterspeech [57, 90] relied on curated or scraped datasets [16] and evaluation metrics based on correct countering claims [33] or emotion and politeness [68]. These approaches, while offering interesting explorations, did not consider counterspeech and its full context or the user intentions of counterspeaking. 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 counterpseech with divers perspectives of both activists and everyday users, we used a mixed methods approach [74]. 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 [23]. We used grounded theory to analyze the qualitative results and triangulation to support our findings discussed in Section 4.

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 doing counterspeech online in a systematic way. One of the authors had conducted a previous study with the participants, so recruitment was carried out through direct invitations. 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.

3.1.3 Analysis Method. Grounded Theory methodology [14, 31] 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, axial coding, and selective coding.

- (1) Open Coding: In this initial stage, each interview transcript was thoroughly reviewed, and initial codes were generated by breaking down the data into meaningful units. This process involved line-by-line coding, enabling the emergence of concepts and patterns without preconceived notions.
- (2) Axial Coding: Building upon the initial codes, axial coding aimed to establish relationships between codes and identify broader categories and subcategories. This stage helped uncover the connections and interactions within the data.
- (3) Selective Coding: The final stage involved refining and integrating categories to develop a coherent theoretical framework. Codes were examined for commonalities and variations, leading to the identification of core concepts and their interrelationships.

Stage one and stage two coding was conducted by the same author who conducted the interviews, as was required by our IRB to protect participant confidentiality. The research team conducted stage three coding together, integrating the findings of both the survey research and the interview study.

3.2 Survey Study

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

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 [49], 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 1. 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

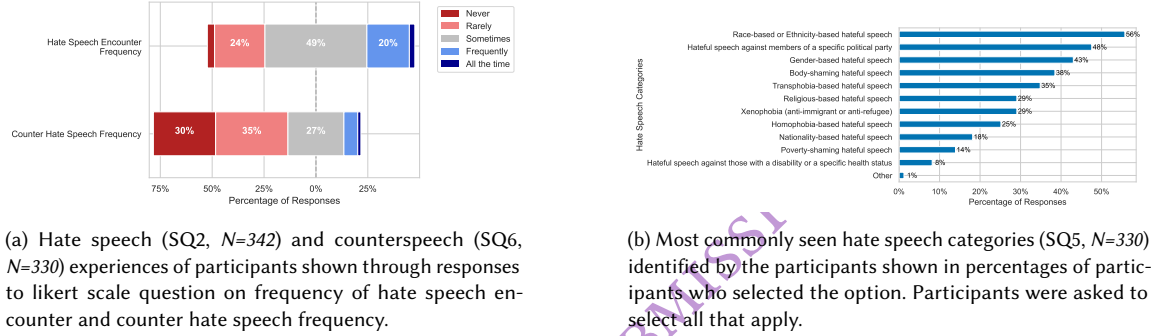


Fig. 1. Participant responses on hate speech and counterspeech experience questions and most commonly seen type of hate.

3.2.3 Analysis Methods.

Quantitative Responses. 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. We also aggregate similar responses together (e.g., extremely likely, likely) for simplicity.

Qualitative Responses. To analyze the qualitative survey responses, we performed open coding on the data. In this analysis, we interpreted generated codes as itemized claims about the data to be investigated in other work and surfaced coding disagreements, following recent best practices in qualitative analysis [34].

To open code the data, separate documents containing the survey responses to each question were created [69]. For barriers to counterspeech (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 (Section 4.2), we analyze the reasons 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 analyze participants responses on envisioned AI tools in counterspeech (SQ15). For the concerns of AI involvement in counterspeech (Section 4.3), we analyze the reasons 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").

Next, two authors inductively generated codes for 25% of the data by developing individual code books 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 into a shared codebook. The authors merged codes with similar themes into a single code. For the remaining codes, the authors discussed instances of disagreement and unanimously agreed to add, merge, or delete the code from the codebook.

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 they both agreed upon and discussed the disagreements on others. They added or removed the code from the instance upon unanimous agreement. Disagreements largely occurred due to differing scopes of codes and at times different interpretations of the statements, which were resolved through discussion.

3.3 Limitations

Both our survey and interview studies and their analysis contain limitations as our results can only include 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 less technologically literate populations. 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.

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 [84], which can be difficult to distinguish from human responses [19]. 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.

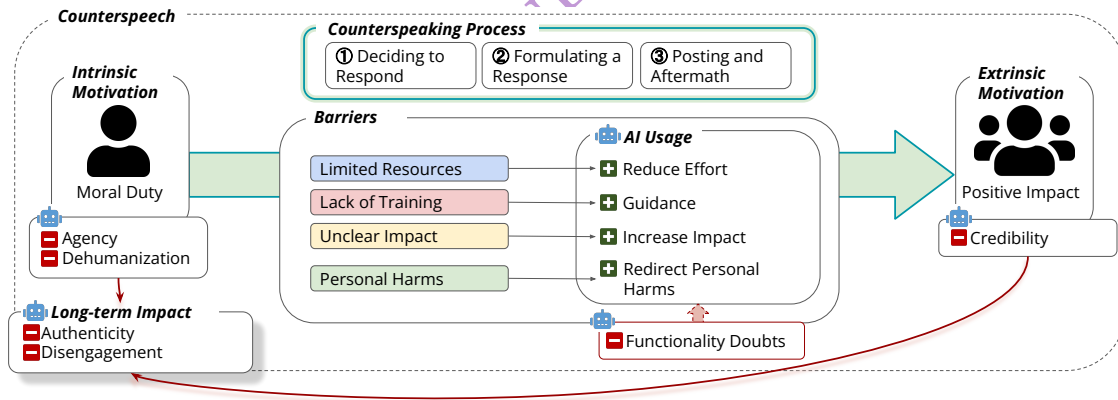


Fig. 2. An overview of interactions between the themes surfaced in Section 4. The figure, from left to right, show an overall counterspeech experience surfaced from participant responses. Participants’ intrinsic and extrinsic motivation encouraged participants to engage in the counterspeaking process 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.

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 a different 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.

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 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 2). 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.

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 2. Options and participant responses to question about counterspeech barriers (SQ12 in Table 1). 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.

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 2).

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 2). In their responses, survey participants used various tactics. 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 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 2). 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 types of tools, the usability and level of involvement in the counterspeaking process, and the types of support, empowerment and emotional and functional support.

Usability. The most prevalent preference surfaced was towards a usable tool, which participants described in two dimensions, easy to use and functional. Especially in addressing 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*

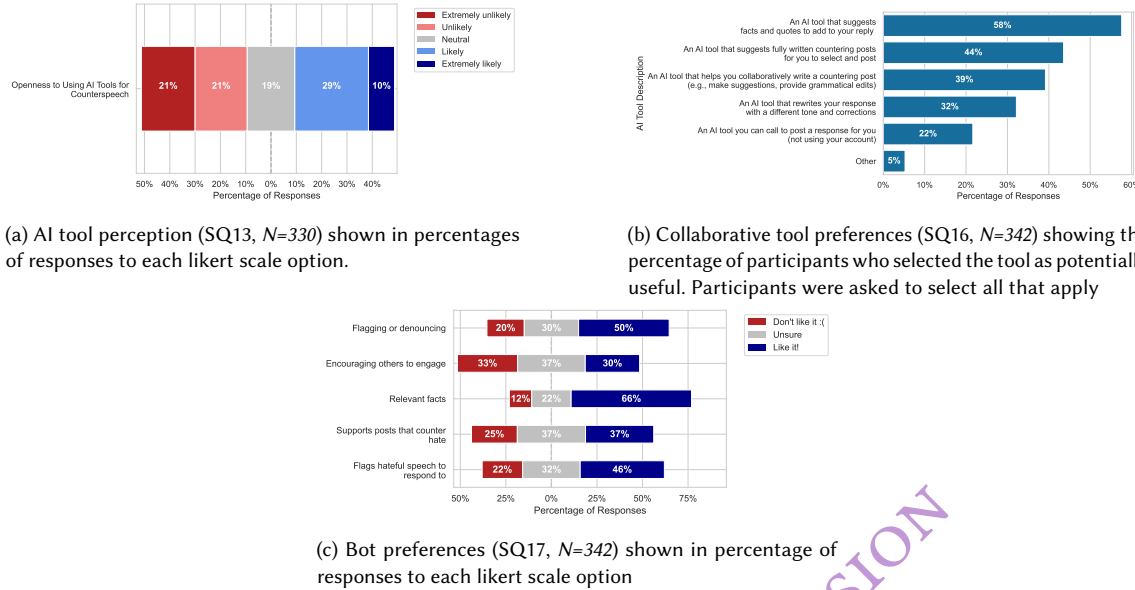


Fig. 3. Survey participant responses on quantitative questions about AI tools.

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, participants identified usability as a high level requirement of the system.

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.

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.

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 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).

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 and inexperienced counterspeakers 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 and describe potential avenues for future work.

5.1 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 [41, 75]. 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 [32, 45].

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 [51], any system that generates automated counterspeech responses (bots) should clearly disclose that it is not a human [27]. Bot and AI generated text detection [22, 66], and disclosure policies [85] 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 [37, 40]. 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 communicate and present communication that is collaboratively created with AI.

5.2 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 [42] 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 [25] and misattribution of responsibility in AI mediated communication using AI as a scapegoat when conversations go awry [39].

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 [7] or explanations about its outputs [83] should be explored and integrated into design of such AI systems. Furthermore, transparency about failures of AI systems, notably encoded biases [28, 86] and limited cultural context [70], would also be an important feature as AI generated text can instill values and norms [44, 50]. In addition to building unbiased and culturally aware AI systems [58], customization could help promote user agency and correct these shortcomings [47, 82]. 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 [43].

5.3 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 [11, 46, 73]. Mental health of those behind hate speech (i.e., the recipients of counterspeech) should also be taken into consideration as counterspeech can sometimes involve call-out based [65] online shaming behaviors [81] and domination [87] which could escalate conflict or lead to harm [55].

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 [79] and focus on guiding systems to reflect a call-in culture that encourages relating to others in affirming ways rather than shaming [65], however, in excess, this could lead to reducing the opportunity for people to bring more authentic emotions, including negative ones, into conversations [40]. 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 [21, 56] and reflective designs to encourage reflection on intentions [76, 78] 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.

5.4 Future Work

Future work could explore counterspeech with a more global lens to understand a wider set of barriers and their interaction with AI assistance. Additionally, our work focused on text-based counterspeech and AI assistance focusing on existing online platforms. Therefore, future works could explore additional modalities and platforms for counterspeech. We also scoped our current work to be against hate speech. However, other forms of harm, such as fake news, could also be addressed by counterspeech. Future work could explore countering fake news and other forms of hate 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.

6 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

To Robert, for the bagels and explaining CMYK and color spaces.

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] 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>
- [7] Zana Bućinca, 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>
- [8] Catherine Buerger. 2021. #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. 7, 4 (2021), 20563051211063843. <https://doi.org/10.1177/20563051211063843>
- [9] Catherine Buerger. 2021. *Speech as a Driver of Intergroup Violence: A Literature Review*. <https://doi.org/10.2139/ssrn.4066876>
- [10] Catherine Buerger. 2022. *Why They Do It: Counterspeech Theories of Change*. <https://doi.org/10.2139/ssrn.4245211>
- [11] 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>
- [12] 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>
- [13] Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. *Philosophy Compass* 18, 1 (2023), e12890. <https://doi.org/10.1111/phc3.12890>
- [14] K. Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE Publications. <https://books.google.com/books?id=v1qP1KbXz1AC>
- [15] Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding Counterspeech for Online Harm Mitigation. arXiv:2307.04761 [cs.CL]
- [16] 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>
- [17] Yi-Ling Chung, Serra Sinem Tekiroglu, 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>
- [18] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.* 91 (2011), 1435.
- [19] 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>
- [20] 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>
- [21] 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>
- [22] Stefano Cresci. 2020. A Decade of Social Bot Detection. *Commun. ACM* 63, 10 (sep 2020), 72–83. <https://doi.org/10.1145/3409116>
- [23] John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research*. Sage publications.
- [24] 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.

- [25] 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>
- [26] 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>
- [27] Oren Etzioni. 2017. Opinion | How to Regulate Artificial Intelligence. (09 2017). <https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html>
- [28] 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>
- [29] 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>
- [30] 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>
- [31] 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=oUxEQAIAAJ>
- [32] 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>
- [33] 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>
- [34] David Hammer and Leema K Berland. 2014. Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences* 23, 1 (2014), 37–46.
- [35] 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>
- [36] 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>
- [37] 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>
- [38] 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>
- [39] 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>
- [40] 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.
- [41] 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.
- [42] Jeffrey W Howard. 2021. Terror, Hate and the Demands of Counter-Speech. *Br. J. Polit. Sci.* 51, 3 (July 2021), 924–939.
- [43] 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>
- [44] 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>
- [45] 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>
- [46] 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>
- [47] 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>
- [48] David Kaye. 2019. *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports. <http://www.jstor.org/stable/j.ctv1fx4h8v>

- [49] 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>
- [50] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports* 13, 1 (April 2023), 4569.
- [51] 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>
- [52] 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
- [53] Kaley Leetaru. 2019. Online Toxicity Is As Old As The Web Itself But The Return To Communities May Help. *Forbes Magazine* (May 2019).
- [54] Olivier Lemeire. 2021. Falsifying generic stereotypes. *Philosophical Studies* 178, 7 (01 Jul 2021), 2293–2312. <https://doi.org/10.1007/s11098-020-01555-3>
- [55] Maxime Lepoutre. 2022. Hateful Counterspeech. *Ethical Theory and Moral Practice* (27 Oct 2022). <https://doi.org/10.1007/s10677-022-10323-7>
- [56] 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>
- [57] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhanian, 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>
- [58] 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>
- [59] 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>
- [60] 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>
- [61] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
- [62] 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>
- [63] Christina Pan, Sahil Yakhmi, Tara Iyer, Evan Strasznick, 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).
- [64] 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>
- [65] Loretta J Ross. 2019. Speaking up without tearing down. *Teaching Tolerance* 61 (2019), 19–22.
- [66] 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]
- [67] 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>
- [68] 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]
- [69] Johnny Saldaña. 2009. *The Coding Manual for Qualitative Researchers*. SAGE Publications.
- [70] 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>
- [71] 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>
- [72] 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>
- [73] 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/fpsy.2021.674298>
- [74] 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>
- [75] 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>

- [76] 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>
- [77] 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>
- [78] 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>
- [79] 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>
- [80] 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>
- [81] Krista Thomason. 2021. The Moral Risks of Online Shaming. In Oxford Handbook of Digital Ethics. Oxford University Press.
- [82] 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>
- [83] 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>
- [84] 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).
- [85] 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>
- [86] 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) <https://arxiv.org/abs/2112.04359>
- [87] 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>
- [88] 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>
- [89] Xinchun 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>
- [90] Wanzheng Zhu and Suma Bhat. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. (June 2021). [arXiv:2106.01625](https://arxiv.org/abs/2106.01625) [cs.CL]
- [91] 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) <https://arxiv.org/abs/2005.12423>

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 V2¹ and V3². Moreover, one of the authors manually annotated answers for bot-like behaviors looking for responses that were repetitive, off-topic, or non-sensical.

A.2 Participant Demographics

Participant demographics are shown in Table 3. 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 3.1.3 and 3.2.3 are listed in Tables 4, 5, and 6. 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 🤔.

¹<https://developers.google.com/recaptcha/docs/display>

²<https://developers.google.com/recaptcha/docs/v3>

(a) Age		(b) Race	
Option	Response (%)	Option	Response (%)
18-24 years old	02.9240	White or Caucasian	82.6979
25-34 years old	34.7953	White or Caucasian,Asian,Native Hawaiian or Other Pacific Islander	00.2933
35-44 years old	34.5029	White or Caucasian,Asian	01.7595
45-54 years old	14.3275	White or Caucasian,Other	00.5865
55-64 years old	09.9415	White or Caucasian,American Indian/Native American or Alaska Native	01.1730
65+ years old	02.9240	White or Caucasian,Black or African American	00.2933
Prefer not to disclose	00.5848	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		(d) Gender	
Option	Response (%)	Option	Response (%)
Yes	02.3460	Man	56.4327
No	96.4809	Woman	41.2281
Prefer not to disclose	01.1730	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		(f) Religion	
Option	Response (%)	Option	Response (%)
Straight (heterosexual)	85.0877	Atheist	16.9591
Bisexual	06.4327	Christian	39.1813
Bisexual,Pansexual	00.5848	Agnostic	16.3743
Asexual	02.0468	Catholic	15.4971
Asexual,Straight (heterosexual)	00.2924	Jewish	01.1696
Lesbian	01.1696	Buddhist	01.1696
Gay	01.1696	Hindu	00.5848
Pansexual	01.1696	Muslim	00.2924
Questioning or unsure	00.2924	Nothing in particular	04.3860
Prefer not to disclose	01.7544	Prefer not to disclose	02.3392
		Something else, Specify:	02.0468
(g) Political Leaning		(h) Education	
Option	Response (%)	Option	Response (%)
Strongly liberal	22.2222	Bachelor's degree	57.6023
Liberal	32.4561	Some college, but no degree	13.1579
Moderate	18.7135	Graduate or professional degree	07.6023
Conservative	16.6667	High school diploma or GED	10.8187
Strongly conservative	08.4795	Associates or technical degree	09.3567
Prefer not to disclose	01.4620	Some high school or less	00.8772
		Prefer not to say	00.5848
(i) Country of Residence		(j) Country of Residence	
Option	Response (%)	Option	Response (%)
United States of America	97.9472	United States of America	97.9472
Namibia	00.2933	Namibia	00.2933
Canada	00.5865	Canada	00.5865
United Kingdom of Great Britain and Northern Ireland	00.2933	United Kingdom of Great Britain and Northern Ireland	00.2933
India	00.2933	India	00.2933
Albania	00.2933	Albania	00.2933
Argentina	00.2933	Argentina	00.2933

Table 3. Demographics of survey participants. All the heuristics are reported as percentages.










Code	Description	Representative Quote
Barriers to Counterspeech		
 <u>Resources</u>	Financial or people resources are limited, especially time it takes to do counterspeech	<i>"[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."</i>
 <u>Finding hate speech</u>	Takes time to find hate speech	<i>"It's so time consuming looking for articles."</i>
 <u>Training</u>	People don't have the training	<i>"People don't always know what to say."</i>
 <u>Reach</u>	It's hard to reach people, which is discouraging	<i>"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"</i>
 <u>Risk</u>	There are risks of online or offline attacks	<i>"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."</i>
 <u>Mental health</u>	It is too hard on mental health (stress or boredom)	<i>"You are just overwhelmed with what you are seeing."</i>
Counterspeaker Motivations		
 <u>Right</u>	It's the right thing to do / civic responsibility	<i>"I think it's because it's the right thing to do – I feel that at least I tried. "</i>
 <u>Impact</u>	Seeing evidence of successful impact	<i>"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."</i>
 <u>Scale - small</u>	Quotes about the individual-level impact	<i>"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 despair before seeing your comment.'"</i>

Table 4. Codes developed from analysis of interview responses discussing counterspeech barriers and motivations.

Code		Description	Representative Quote
Activist Counterspeakers			
<u>Time</u>	🕒	References to time or efficiency	"Anything that could help us be more efficient - help us produce more."
<u>Finding - AI</u>	🔍	AI would help by locating hateful speech	"I've spent up to two hours looking for good actions, so that would be super helpful."
<u>Scale - big</u>	💪	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			
<u>Efficiency</u>	🕒	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 speech."
<u>AI-better</u>	🔍	The user thinks that AI would be better than they are.	"[An AI] could give me a constructive framework for a much more impactful response than I could otherwise generate on my own."
<u>Capability-dependent</u>	🔍	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."
<u>Information</u>	📄	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"
<u>Guidance</u>	✍️	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"
<u>Emotions</u>	✍️	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..."
<u>Empowerment</u>	💪	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."
<u>Reduce-stress</u>	🧘	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."
<u>AI-proxy</u>	🧑	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			
<u>Existing-technology</u>		The user refers to existing technology a specific AI tool that current exists in the market.	"I use most of the time ChatGPT."
<u>Report</u>	🕒	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."
<u>Response-support</u>	🕒	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."
<u>Factual</u>	📄	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."
<u>Collaborative</u>	✍️	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."
<u>Effective-communication</u>	✍️	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."
<u>Aligned</u>	💪	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."
<u>Protective</u>	🧑	The user would like an AI tool that will protect them from retaliation often through anonymity.	"I would like an AI tool that could prepare a message...[avoids] making myself a target."

Table 5. 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		
<u>Authenticity - strategy</u>	👤 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."
<u>Real</u>	🔗 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."
<u>Long-term</u>	🔗 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."
<u>Counterspeaker</u>	👍 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."
<u>Becoming the monster</u>	👍 Troll farms are bad. Would we become just as bad by using a counterspeech 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."
<u>Emotional Intelligence</u>	🗣️ 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."
<u>Functionality - technical</u>	🗣️ 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		
<u>Authenticity</u>	👤 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."
<u>Engaging-not-helpful</u>	🔗 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."
<u>Avoidance</u>	🔗 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."
<u>Agency</u>	👍 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."
<u>Capability-doubts</u>	🗣️ 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."
<u>Become-target</u>	🗣️ 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 6. 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.