Understanding Counterspeech for Online Harm Mitigation

Yi-Ling Chung[*], The Alan Turing Institute, UK yilingchung27@gmail.com
Gavin Abercrombie, The Interaction Lab, Heriot-Watt University, UK g.abercrombie@hw.ac.uk
Florence Enock, The Alan Turing Institute, UK fenock@turing.ac.uk
Jonathan Bright, The Alan Turing Institute, UK bright@turing.ac.uk
Verena Rieser[†], The Interaction Lab, Heriot-Watt University, UK v.t.rieser@hw.ac.uk

Abstract Counterspeech offers direct rebuttals to hateful speech by challenging perpetrators of hate and showing support to targets of abuse. It provides a promising alternative to more contentious measures, such as content moderation and deplatforming, by contributing a greater amount of positive online speech rather than attempting to mitigate harmful content through removal. Advances in the development of large language models mean that the process of producing counterspeech could be made more efficient by automating its generation, which would enable large-scale online campaigns. However, we currently lack a systematic understanding of several important factors relating to the efficacy of counterspeech for hate mitigation, such as which types of counterspeech are most effective, what are the optimal conditions for implementation, and which specific effects of hate it can best ameliorate. This paper aims to fill this gap by systematically reviewing counterspeech research in the social sciences and comparing methodologies and findings with natural language processing (NLP) and computer science efforts in automatic counterspeech generation. By taking this multi-disciplinary view, we identify promising future directions in both fields.

1 Introduction

The exposure of social media users to online hate and abuse continues to be a cause for public concern. Volumes of abuse on social media continue to be significant in absolute terms (Vidgen et al., 2019), and some claim they are rising on platforms such as Twitter where, at the same time, content moderation appears to be becoming less of a priority (Frenkel and Conger, 2022). Receiving abuse can have negative effects on the mental health of targets, and also on others witnessing it (Siegel, 2020; Saha et al., 2019). In the context of public figures, the impact on the witnesses (bystanders) is arguably even more important, as the abuse is potentially witnessed by a large volume of people. In addition, politicians and other prominent actors are driven out of the public sphere precisely because of the vitriol they receive on a daily basis (News, 2018), raising concerns for the overall health of democracy.

Within this context, research on mechanisms for combating online abuse is becoming ever more important. One such research angle is the area of "counterspeech" (or counter-narratives): content that is designed to resist or contradict abusive or hateful content (Benesch, 2014a; Saltman and Russell, 2014; Bartlett and Krasodomski-Jones, 2015), also see Figure 1. Such coun-

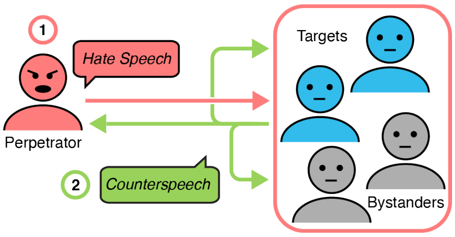

Figure 1: Counterspeech dynamics. (1) Perpetrator(s) generate Hate Speech. This may be witnessed by either targets and/or bystanders. (2) Counterspeaker(s) respond with counterspeech, which may be directed at the perpetrator(s), bystanders (e.g. to provide alternative perspectives), or other targets (e.g. in support). Counterspeakers may themselves be targets or bystanders, or could be members of organised counterspeech groups. They can have *in-* or *out*-group identities with respect to either the perpetrator(s) or the target(s). Counterspeech is directed at recipients, who can be one or more of (a) the perpetrator(s), (b) the target(s), or (c) other bystanders. Both counterspeakers and targets can be individual or multiple (one-to-one, one-to-many and so on).

[*] Now at Genaios Safe AI.
[†] Now at Google DeepMind.

terspeech (as we will elaborate more fully below) is an important potential tool in the fight against online hate and abuse as it does not require any interventions from the platform or from law enforcement, and may contribute to mitigating the effects of abuse (Munger, 2017; Buerger, 2021b; Hangartner et al., 2021; Bilewicz et al., 2021) without impinging on free speech. Several civil organisations have used counterspeech to directly challenge hate, and Facebook has launched campaigns with local communities and policymakers to promote accessibility to counterspeech tools.[1] Similarly, Moonshot and Jigsaw implemented The Redirect Method, presenting alternative counterspeech or counter videos when users search queries that may suggest an inclination towards extremist content or groups.[2]

The detection and generation of counterspeech is important because it underpins the promise of AI-powered assistive tools for hate mitigation. Identifying counterspeech is vital also for analytical research in the area: for instance, to disentangle the dynamics of perpetrators, victims and bystanders (Mathew et al., 2018; Garland et al., 2020, 2022), as well as determining which responses are most effective in combating hate speech (Mathew et al., 2018, 2019; Chung et al., 2021a).

Automatically producing counterspeech is a timely and important task for two reasons. First, composing counterspeech is time-consuming and requires considerable expertise to be effective (Chung et al., 2021b). Recently, large language models have been able to produce fluent and personalised arguments tailored to user expectations addressing various topics and tasks. Thus, developing counterspeech tools is feasible and can provide support to civil organisations, practitioners and stakeholders in hate intervention at scale. Second, by partially automating counterspeech writing, such assistive tools can lessen practitioners' psychological strain resulting from prolonged exposure to harmful content (Riedl et al., 2020; Chung et al., 2021b).

However, despite the potential for counterspeech, and the growing body of work in this area, the research agenda remains a relatively new one, which also suffers from the fact that it is divided into a number of disciplinary silos. In methodological terms, meanwhile, social scientists studying the dynamics and impacts of counterspeech (e.g. Munger, 2017; Buerger, 2021b; Hangartner et al., 2021; Bilewicz et al., 2021) often do not engage with computer scientists developing models to detect and generate such speech (e.g. Chung et al., 2021c; Saha et al., 2022) (or vice versa). This disconnection may increase the time and effort for tackling online harms.

The aim of this review article is to fill this gap, by providing a comprehensive, multi-disciplinary overview of the field of counterspeech covering computer science[3] and the social sciences over the last ten years. We make a number of contributions in particular. Firstly, we outline a definition of counterspeech and a framework for understanding its use and impact, as well as a detailed taxonomy. Visualised in Figure 1, such a framework helps delineate the interaction of hate speech and responses within people involved in the conversations (i.e. perpetrators, targets and bystanders). We review research on the effectiveness of counterspeech, bringing together perspectives on the impact it makes when it is experienced. Thus, computer scientists can adeptly approach counterspeech studies and develop effective tools based on our analysis. We also analyse technical work on counterspeech, looking specifically at the task of counterspeech generation, scalability, and the availability and methodology behind different datasets. Importantly, across all studies, we focus on commonalities and differences between computer science and the social sciences, including how the impact of counterspeech is evaluated and which specific effect of hate speech it best ameliorates.

We draw on our findings to discuss the challenges and directions of open science (and safe AI) for online hate mitigation. For computer scientists, we provide evidence-based recommendations for automatic approaches to counterspeech tools using Natural Language Processing (NLP). Similarly, for social scientists, we set out future perspectives on interdisciplinary collaborations with AI researchers on mitigating online harms, including conducting large-scale analyses and evaluating the impact of automated interventions. Taken together, our work offers researchers, policymakers and practitioners the tools to further understand the potentials of automated counterspeech for online hate mitigation.

2 Background

Interest in investigating the social and computational aspects of counterspeech has grown considerably in the past five years. However, while extant work reviews the impact of counterspeech on hate mitigation (Saltman and Russell, 2014; Carthy et al., 2020; Buerger, 2021a), none have systematically addressed this issue in combination with computational studies in order to synthesise social scientific insights and discuss the potential role of automated methods in reducing harms. Carthy et al. (2020) present a focused (2016-2018) systematic review of research into the impact of counter-narratives on prevention of violent radicalisation. They cate-

[1] https://counterspeech.fb.com/en/

[2] https://moonshotteam.com/the-redirect-method/

[3] While most studies on computational approaches to counterspeech included in this review adopt natural language processing techniques, we use 'computer science' to broadly cover the research field in which the studies are done.

gorise the techniques employed in counter-narratives into four groups: (1) counter-stereotypical exemplars (challenging stereotypes, social schema or moral exemplars), (2) persuasion (e.g., through role-playing and emotion inducement), (3) inoculation (proactively reinforcing resistance to attitude change or persuasion), and (4) alternative accounts (disrupting false beliefs by offering different perspectives of events). The measurements of counter-narrative interventions are based on (1) intent of violent behaviour, (2) perceived symbolic/realistic group threat (e.g., perception of an outgroup as dangerous), and (3) in-group favouritism/outgroup hostility (e.g., level of trust, confidence, discomfort and forgiveness towards out-groups). They argue that counter-narratives show promise in reducing violent radicalisation, while its effects vary across techniques, with counter-stereotypical exemplars, inoculation and alternative accounts demonstrating the most noticeable outcomes. Buerger (2021a) reviews the research into the effectiveness of counterspeech, attempting to categorise different forms of counterspeech, summarise the source of influences in abusive/positive behaviour change, and elucidate the reasons which drive strangers to intervene in cyberbullying. Here, the impact of counterspeech is mostly evaluated by the people involved in hateful discussions, including hateful speakers, audiences, and counterspeakers. In comparison, we focus on *what* makes counterspeech effective by comprehensively examining its use based on aspects such as strategies, audience and evaluation.

On the computational side, some work reviews the use of counterspeech in social media using natural language processing, including work outlining counterspeech datasets (Adak et al., 2022; Alsagheer et al., 2022), discussing automated approaches to counterspeech classification (Alsagheer et al., 2022) and generation (Chaudhary et al., 2021; Alsagheer et al., 2022), and work focusing on system evaluation (Alsagheer et al., 2022). However, NLP work from computer sciences is not typically informed by important insights from the social sciences, including the key roles of intergroup dynamics, the social context in which counterspeech is employed, and the mode of persuasion by which counterspeech operates. Taking an interdisciplinary approach, we join work from the computer and social sciences.

3 Review Methodology

Taking a multi-disciplinary perspective, we systematically review work on counterspeech from computer science and the social sciences published in the past ten years. To ensure broad coverage and to conduct a reproducible review, we follow the systematic methodology of Moher et al. (2009). The search and inclusion process is shown in Figure 2.

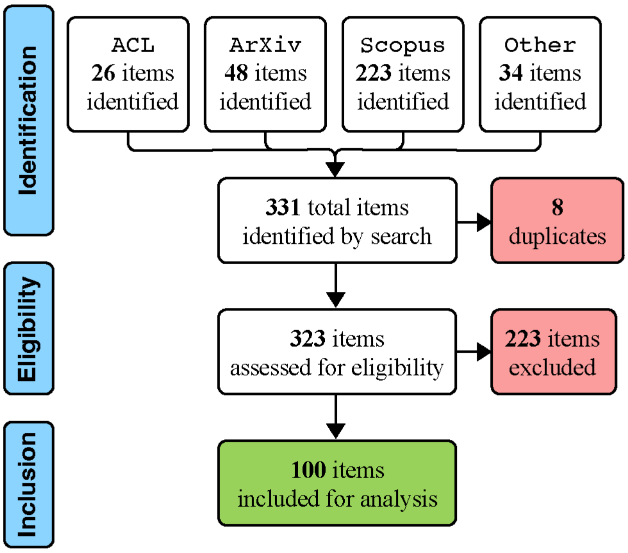

Figure 2: Flow diagram showing the identification, eligibility screening, and inclusion phases of the selection of items analysed in this review.

We used keyword terms related to counterspeech to search three key databases (ACL Anthology, ArXiv, and Scopus) that together offer a broad coverage of our target literature. We included the search terms 'counter-speech', 'counter-narratives', 'counter-terrorism', 'counter-aggression', 'counter-hate', 'counter speech', 'counter narrative', 'countering online hate speech', 'counter hate speech', and 'counter-hate speech'. We also included 34 publications that we had identified previously from other sources, but that were not returned by keyword search due to not including relevant keywords or not being indexed in the target search repositories. The search covers the data within the period between 2005 and 2023. Of the returned results, we include all publications that concern (1) analysis of the use and effectiveness of interventions against hateful or abusive language online, (2) characteristics of counterspeech users or recipients, or (3) data and/or implementation designed for counterspeech (e.g., counterspeech classification or generation). These inclusion criteria were applied by two of the authors. Following this process, we include 100 papers for analysis in this review. Each of the papers was read by at least one of the co-authors of the article.

Our review is divided into several sections (the results of which are presented sequentially below). First, we examine definitional characteristics of counterspeech, looking at how the term itself is defined, how different taxonomies have been created to classify different types of counterspeech, and the different potential purposes attributed to it. Based on the definitional characteristics, we examine studies that have looked at the impact of counterspeech, discussing the different analytical designs employed and analysing evidence of the results. Following this, we discuss computational approaches to counterspeech, focusing in particular on both detection and generation. Finally, we examine ethical issues in the domain of counterspeech, and also speculate about future perspectives and directions in the field.

4 Defining counterspeech

Counterspeech is multifaceted and can be characterised in several different ways. In Table 1 we outline a framework for describing and designing counterspeech, covering who (speaker) sends what kinds of messages (strategies) to whom (recipients), and for what purpose (purpose). Using this structure, we summarise how counterspeech has typically been categorised in past studies.

Most studies in the field use one of three main terms: *counterspeech*, *counter-narratives* (Reynolds and Tuck, 2016; Carthy and Sarma, 2021; Tuck and Silverman, 2016; Iqbal et al., 2019) and *hope speech* (Snyder et al., 2018). These three terms broadly refer to a similar concept: content that challenges and rebuts hateful discourse and propaganda (Saltman and Russell, 2014; Bartlett and Krasodomski-Jones, 2015; Benesch et al., 2016; Saltman et al., 2021; Garland et al., 2022) using non-aggressive speech (Benesch et al., 2016; Reynolds and Tuck, 2016; Schieb and Preuss, 2016). There are some differences between the terms. Ferguson (2016) considers counter-narratives as intentional strategic communication within a political, policy, or military context. Additionally, the term counter-narrative also refers to narratives that challenge a much broader view or category such as forms of education, propaganda, and public information (Benesch et al., 2016). Such counter-narratives are often discussed in the context of the prevention of violent extremism. Hope speech, meanwhile, could be seen as a particular type of counterspeech: it promotes positive engagement in online discourse to lessen the consequences of abuse, and places a particular emphasis on delivering optimism, resilience, and the values of equality, diversity and inclusion (Chakravarthi, 2022). In this paper, we review work that relates to all of these three concepts, and largely make use of the catch-all term counterspeech, while acknowledging the slight differences between the concepts.

4.1 Classifying counterspeech

Researchers have identified a variety of different types of counterspeech. Here, we outline four main ways in which counterspeech can vary, in terms of the identity of the counterspeaker, the strategies employed, the recipient of the counterspeech and the purpose of counterspeech.

Counterspeakers (who) Psychological studies show that the identity of a speaker plays a key role in how large an audience their message reaches and how persuasive the message is. Common crucial factors include group identity (such as race, religion, and nationality), level of influence, and socioeconomic status. For instance, counterspeech provided by users with large numbers of followers and from an in-group member is more likely to lead to changes in the behaviour of perpetrators of hate (Munger, 2017).

Some studies characterise individuals who use counterspeech and suggest that these users exhibit different characteristics and interests than users who spread hate (Mathew et al., 2018, 2019; Buerger, 2021b). Through lexical, linguistic and psycholinguistic analysis of users who generate hate speech or counterspeech on Twitter, Mathew et al. (2018) find that counterspeakers are higher in agreeableness, displaying traits such as altruism, modesty, and sympathy, and display higher levels of self-discipline and conscientiousness. Possibly driven by a motive to help combat hate speech, counterspeakers tend to use words related to government, law, leadership, pride, and religion. Regarding the impact of being a counterspeaker, in an ethnographic study, members of a counterspeech campaign reported feeling more courageous and keen to engage in challenging discussions after expressing opinions publicly (Buerger, 2021b).

Strategies (how) Counterspeech can take many forms. Benesch et al. (2016) first identify eight types of counterspeech used on Twitter: (1) *presentation of facts*, (2) *pointing out hypocrisy or contradiction*, (3) *warning of consequences*, (4) *affiliation* [i.e. establishing an emotional bond with the perpetrators or targets of hate], (5) *denouncing*, (6) *humour/sarcasm*, (7) *tone* [a tendency or style adopted for communication, e.g., empathetic and hostile], and (8) *use of media*. Based on this taxonomy, follow-up studies on counterspeech make minor modifications to cover strategies in a broader scope. Mathew et al. (2018) analyzed and classified counterspeech on Twitter, taking Benesch et al. (2016)'s taxonomy but dropping *the use of media* and adding *hostile language* and *positive tone*, which replaces general strategy *tone*. Similarly, Mathew et al. (2019) collected and annotated counterspeech comments from Youtube, adopting Benesch et al. (2016)'s taxonomy but excluding *tone* and adding *positive tone*, *hostile language* and *miscellaneous*. Chung et al. (2019) collaborated with NGOs to collect manually written counterspeech. For data annotation, they followed the taxonomies pro-

Aspects	Description
Speaker	Who is the counterspeaker? What is the social identity and status of the counterspeaker?
Strategy	Which linguistic and rhetorical methods are used in the counterspeech? Which emotions or attitudes are expressed towards the hateful content?
Recipient	Who is the target audience? Are they hate speakers, targets of hate, or bystanders?
Purpose	What is the aim of disseminating counterspeech?

Table 1: Framework for describing and designing counterspeech.

vided by Benesch et al. (2016) and Mathew et al. (2019), while adding *counter question* and discarding *the use of media*. Counterspeech examples for each strategy are provided in Table 2.

Counterspeech recipients (whom) Depending on the purpose of the counterspeech, the target audience may be perpetrators, victims or bystanders (see Figure 1). Identifying the appropriate target audience or 'Movable Middle' is crucial to maximise the efficacy of counterspeech. Movable middle refers to individuals who do not yet hold firm opinions on a topic and can hence be potentially open to persuasion. They are also receptive to arguments and more willing to listen. These individuals often serve as ideal recipients of messages addressing social issues such as vaccination hesitancy (Litaker et al., 2022). In the context of counterspeech, previous studies show that a small group of counterspeakers can shape online discussion when the audience holds moderate views (Schieb and Preuss, 2016; Buerger, 2021b).

Wright et al. (2017) group counterspeech acts into four categories based on the number of people involved in the discussion: *one-to-one*, *one-to-many*, *many-to-one*, or *many-to-many*. Some successful cases where counterspeech induces favourable changes in the discourse happen in a one-to-one discussion. This allows for dedicated opinion exchange over an ideology, which in some cases even yields long-lasting changes in beliefs. The use of hashtags is a good example of one-to-many and many-to-many interaction where conversations surge quickly (Benesch et al., 2016; Wright et al., 2017). For instance, Twitter users often include hashtags to express support (e.g., #BlackLivesMatter) or disagreement with haters (e.g., #StopHate) to demonstrate their perspective.

The purpose of counterspeech Hateful language online can serve to reinforce prejudice (Citron and Norton, 2011), encourage further division, promote power of the ingroup, sway political votes, provoke or justify offline violence, and psychologically damage targets of hate (Jay, 2009). Just as the effects of hate are wide-ranging, counterspeech may be used to fulfil a variety of purposes.

• ***Changing the attitudes and behaviours of perpetrators*** In directly challenging hateful language, one key aim of counterspeech can be to change the attitudes of the perpetrators of hate themselves. The strategy here is often to persuade the perpetrator that their attitudes are mistaken or unacceptable, and to deconstruct, discredit or delegitimise extremist narratives and propaganda (Reynolds and Tuck, 2016). Counterspeech aimed at changing the attitudes of spreaders of hate may address the hate speaker directly, countering claims with facts or by employing empathy and affiliation. Challenging attitudes is often seen as a stepping stone to altering behaviours (Stroebe, 2008). In attempting to change the minds of perpetrators, counterspeakers ultimately hope to discourage associated behaviours such as sharing such content again in the future or showing support for other hateful content (i.e., stopping the spread of hate). In changing the minds of perpetrators, counterspeakers may also hope to prevent them from engaging in more extreme behaviours such as offline violence.

• ***Changing the attitudes and behaviours of bystanders*** More commonly, counterspeech is initiated with the intention of reaching the wider audience of bystanders rather than perpetrators of hate themselves (Buerger, 2022). These bystanders are not (at least yet) generating hateful language themselves, but rather are people exposed to hateful content either incidentally or by active engagement. Here, counterspeakers hope to persuade bystanders that the hateful content is wrong or unacceptable, again by deconstructing and delegitimising the hateful narrative. The strategy here may be to offer facts, point out hypocrisy, denounce the content, or use humour to discredit the speaker. Additionally, counterspeakers will often invoke empathy for targets of hate. In preventing bystanders from forming attitudes and opinions in line with the hateful narrative, counterspeakers hope to mitigate further intergroup division and related behaviours such as support for or engagement with additional abuse or physical violence. Counterspeakers may also hope to encourage others to generate rebutals and rally support for victims (Benesch, 2014a), bringing positive changes in online discourse.

• ***Showing support for targets of hate*** A third key way in which counterspeech functions is to show

Strategy	Example
Facts	Actually, studies show that on the whole migrants contribute more to public finances than they take out, see this article for example.
Hypocrisy	Immigrants stealing British resources? A bit rich given how much was stolen from colonies by the British Empire.
Consequences	Spreading hateful content is illegal. Police will knock on your door.
Affiliation	As a British national, I know life is hard here right now. But I assure you that your unemployment is not the fault of immigrants.
Denouncing	Stop with the racist and derogatory slurs. It's unacceptable to talk this way.
Counter questions	Do you have a problem with all immigrants or only ones from lower income countries? Are you suggesting we have enough qualified and willing British born workers to fill all the jobs?
Humour	You should think about how the Spanish feel next time you go on holiday to Costa Del Sol (laughing emoji)?
Positive tone	Immigrants strengthen UK society in so many ways - greater diversity, skillsets and innovation to name a few! And no way our NHS could function without the immigrant workforce.

Table 2: Synthetic examples of different counterspeech strategies in response to an example of abuse against immigrants. Here the abuse example is: 'Immigrants are invading and stealing our resources'.

support directly to targets of hate. Online abuse can psychologically damage the wellbeing of targets and leave them feeling fearful, threatened, and even in doubt of their physical safety (Benesch, 2014b; Leader Maynard and Benesch, 2016; Saha et al., 2019; Siegel, 2020). By challenging such abuse, counterspeakers can offer support to targets and encourage bystanders to do the same (Buerger, 2021b). This support aims to alleviate negative emotion brought on by hate by demonstrating to targets that they are not alone and that many people do not hold the attitudes of the perpetrator. Here the particular strategies may be to denounce the hate and express positive sentiment towards the target group. Intergroup solidarity may in turn reduce retaliated antagonism.

5 The Impact of Counterspeech

While we have delineated the characteristics of counterspeech, its concrete effects on harm mitigation remain debated. The methods applied for evaluating the effectiveness of counterspeech vary considerably across studies in the field. In this section we provide an evidence-based analysis of counterspeech's efficacy, examining how it is used in real-life scenarios and its influence based on eight aspects.

Research design A wide range of methodologies have been adopted to assess the impact of counterspeech on hate mitigation, including observational studies (Ernst et al., 2017; Stroud and Cox, 2018; Garland et al., 2022), experimental (Munger, 2017; Obermaier et al., 2021; Hangartner et al., 2021) and quasi-experimental designs (Bilewicz et al., 2021). In observational studies, investigators typically assess the relationship between exposure to counterspeech and outcome variables of interest without any experimental manipulation. For instance, a longitudinal study of German political conversations on Twitter examined the interplay between organized hate and counterspeech groups (Garland et al., 2022). There is also an ethnographic study interviewing counterspeakers on Facebook to understand external and internal practices for collectively intervening in hateful comments, such as how to build effective counterspeech action and keep counterspeakers engaged (Buerger, 2021b). For experimental and quasi-experimental designs, both aim at estimating the causal effects of exposure to different kinds of counterspeech on outcome variables in comparison with controls (no exposure to counterspeech).

Languages and countries In the reviewed work, the impact of counterspeech is investigated in five different languages across nine countries. Notably, experiments are focused on counterspeech used in Indo-European languages such as English (USA, UK, Canada and Ireland), German (Germany), Urdu (Pakistan) and Swedish (Sweden). Only two studies are dedicated to Afro-Asiatic languages, Arabic (Egypt and Iraq). We did not find research dedicated to other language families, suggesting that the language coverage of counterspeech studies is still low.

Platforms Most experiments were conducted on text-based social media platforms, such as eight on Twitter (Benesch et al., 2016; Reynolds and Tuck, 2016; Silverman et al., 2016; Stroud and Cox, 2018; Munger, 2017; Hangartner et al., 2021; Poole et al., 2021; Garland et al., 2022), six on Facebook (Reynolds and Tuck, 2016; Silverman et al., 2016; Schieb and Preuss, 2016;

Leonhard et al., 2018; Saltman et al., 2021; Buerger, 2021b), and one on Reddit (Bilewicz et al., 2021), as well as image-based online spaces, such as three on Youtube (Reynolds and Tuck, 2016; Silverman et al., 2016; Ernst et al., 2017) and one on Instagram (Stroud and Cox, 2018). Often, the counterspeech interventions are directly monitored on such platforms, but in some cases, fictitious platforms are created in order to mimic online social activity under a controlled environment (Obermaier et al., 2021; Carthy and Sarma, 2021; Bélanger et al., 2020). There are three studies analysing the impact of counterspeech across multiple platforms (Reynolds and Tuck, 2016; Silverman et al., 2016; Stroud and Cox, 2018).

Twitter and Facebook are widely used for measuring the effects of counterspeech, with eight and six experiments respectively. For Twitter, this can be explained by its easily accessible API (even if at the time of writing continued research access to the API was in doubt). Similarly, because of difficulties in gathering data, Schieb and Preuss (2016) resort to developing an agent-based computational model for simulating hate mitigation with counterspeech on Facebook. It is worth highlighting that none of the studies we reviewed had investigated recently popular mainstream platforms, such as Tiktok, Weibo, Telegram, and Discord.

The target of hate speech Abusive speech can be addressed towards many different potential targets, and each individual hate phenomenon may require different response strategies for maximum effectiveness. Existing studies have evaluated the effectiveness of counterspeech on several hate phenomena, with Islamophobia, Islamic extremism, and racism being the most commonly addressed, while hate against LGBTQ+ community and immigrants being the least studied. In these studies, abusive content is typically identified based on two strategies - hateful keyword matches (Hangartner et al., 2021; Bilewicz et al., 2021), or user accounts (e.g., content produced by known hate speakers) (Garland et al., 2022).

Types of interventions A wide range of methods are exploited to design and surface counterspeech messages to a target audience. We broadly categorise these methods based on modality and approach to creation. Counter speech is generally conveyed in text (Bélanger et al., 2020; Hangartner et al., 2021; Poole et al., 2021) or video mode (Ernst et al., 2017; Saltman et al., 2021; Carthy and Sarma, 2021). In both cases, counterspeech materials can be created in three different ways: written by experimenters as stimuli (Obermaier et al., 2021; Carthy and Sarma, 2021), as well as written by individuals or campaigns that are collected from social media platforms (Benesch et al., 2016; Garland et al., 2022; Buerger, 2021b). We also found one study integrating counterspeech messages in media such as films, TV dramas and movies (Iqbal et al., 2019).

Counterspeech strategies Following the strategies summarised in Section 4.1, commonly used counterspeech strategies include facts (Buerger, 2021b; Obermaier et al., 2021), denouncing (Stroud and Cox, 2018; Saltman et al., 2021), counter-questions (Silverman et al., 2016; Reynolds and Tuck, 2016; Saltman et al., 2021), and a specific tone (humour or empathy) (Reynolds and Tuck, 2016; Munger, 2017; Hangartner et al., 2021; Saltman et al., 2021). There are more fine-grained tactics for designing counterspeech in social science experiments. According to psychological studies, the use of social norms can reduce aggression and is closely related to legal regulation in society (Bilewicz et al., 2021). This tactic was tested in an intervention study where participants were exposed to counterspeech with one of the inducements of empathy, descriptive norms (e.g., *Let's try to express our points without hurtful language*) and prescriptive norms (e.g., *Hey, this discussion could be more enjoyable for all if we would treat each other with respect.*) (Bilewicz et al., 2021). Bélanger et al. (2020) designed counterspeech based on substances rather than tactics, varying three different narratives: (1) social (seeking to establish a better society), (2) political (bringing a new world order through a global caliphate), and (3) religious (legitimising violence based on religious purposes). Considering broader counterspeech components, a few organisations further focus on challenging ideology (e.g., far-right and Islamist extremist recruitment narratives), rather than deradicalising individuals (Silverman et al., 2016; Saltman et al., 2021). Counterspeech drawing from personal stories in a reflective or sentimental tone is also considered as it can resonate better with target audiences (Silverman et al., 2016). In addition to neutral or positive counterspeech, radical approaches are taken by counter-objecting, degrading or shaming perpetrators in public for unsolicited harmful content (Stroud and Cox, 2018; Obermaier et al., 2021).

Types of evaluation metrics Based on Reynolds and Tuck (2016)'s counterspeech *Handbook*, we identified the following three types of metrics used by the authors of the papers to evaluate the effectiveness of counterspeech interventions: social impact, behavioural change, and attitude change measures.

• ***Social impact metrics*** are (usually automated) measurements of how subjects interact with counterspeech online. Such measures include, bounce rate, exit rate,[4]

[4] Bounce rate is the number of users who leave a website without clicking past the landing page; exit rate measures how many people leave the site from a given section (Reynolds and Tuck, 2016).

geo-location analysis and the numbers of likes, views, and shares that posts receive (Garland et al., 2020; Hangartner et al., 2021; Poole et al., 2021; Reynolds and Tuck, 2016; Leonhard et al., 2018; Saltman et al., 2021; Silverman et al., 2016). For example, for one of their experiments, Saltman et al. (2021) measure the 'click-through rates' of Facebook users redirected from hateful to counterspeech materials, while Hangartner et al. (2021) measure retweets and deletions (in addition to behavioural change measures).

Social impact measures are also applied to synthetic data by Schieb and Preuss (2016), who measure the 'likes' of their (simulated) participants as hate and counterspeech propagate through a network (as well as applying behavioural metrics). Taking a more distant, long-term view, Iqbal et al. (2019) cite Egypt's overall success at countering radicalisation with counterspeech campaigns by comparing its position on the Global Terrorism Index with that of Pakistan.

While the majority of these measurements are automated, Leonhard et al. (2018) use survey questions to examine participants willingness to intervene against hate speech depending on the severity of the hate, the number of bystanders, and the reactions of others. Unlike the survey-based approaches described below, they do not consider *changes* in attitude. In addition, Buerger (2021b) assess the success of the #jagärhär counterspeech campaign (#iamhere in English, a Sweden-based collective effort that has been applied in more than 16 countries) based on the extent to which it has facilitated the emergence of alternative perspectives.

- ***Behavioural change measures*** reveal whether subjects change their observable behaviour towards victims before and after exposure to counterspeech, for example in the tone of their language as measured with sentiment analysis.

For instance, Hangartner et al. (2021) conduct sentiment analysis to determine the behaviour of previously xenophobic accounts after treatment with counterspeech, Bilewicz et al. (2021) measure levels of verbal aggression before and after interventions, and Garland et al. (2020) assess the proportion of hate speech in online discourse before and after the intervention of an organised counterspeech group. Other such measures are those of Saltman et al. (2021), who compare the number of times users that violate Facebook policies before and after exposure to counterspeech, and Munger (2017), who examine the likelihood of Twitter users continuing to use racial slurs following sanctions by counterspeakers of varying status and demographics. And in a network simulation experiment, Schieb and Preuss (2016) measure the effect of positive or negative (synthetic) posts on (synthetic) user behaviour.

- ***Attitude change measures*** are used to assess whether people (hate/counter speakers or bystanders) change their underlying attitudes or intentions through non-automated methods such as interviews, surveys, focus groups, or qualitative content analysis.

For potential hate speech perpetrators, Carthy and Sarma (2021) use psychological testing to measure the extent to which participants legitimized violence after exposure to differing counterspeech strategies, Bélanger et al. (2020) compare support for ISIS and other factors using in participants exposed to differing counterspeech strategies and a control group, and Ernst et al. (2017) code user comments on hate and counterspeech videos to perform qualitative content analysis of users' attitudes.

For bystanders that may be potential counterspeakers, Obermaier et al. (2021) use a survey to examine whether counterspeech leads to increased intentions to intervene. And for those already engaged in counterspeech, Buerger (2021b) conduct interviews with members of an organised group to reveal their perceptions of the efficacy of their interventions.

Effectiveness Owing to the variation in experimental setups, aims, and evaluation methods of the counterspeech efforts we review, it is not straightforward to compare their levels of success. Indeed, several of the studies concern broad long-term goals that cannot be easily evaluated at all (e.g. Reynolds and Tuck, 2016; Silverman et al., 2016) or provide only anecdotal evidence (e.g. Benesch et al., 2016; Stroud and Cox, 2018; Buerger, 2021b).

Beyond this, evidence of successful counterspeech forms a complex picture. For example, Garland et al. (2022) show that organised counterspeech is effective, but can produce backfire effects and actually attract more hate speech in some circumstances. They also show that these dynamics can alter surrounding societal events—although they do not make causal claims for this. Similarly, Ernst et al. (2017) find mixed results, with counterspeech encouraging discussion about hate phenomena and targets in some cases, but also leading to increases in hateful comments. However, Silverman et al. (2016) suggest that even such confrontational exchanges can be viewed as positive signs of engagement.

There is some evidence for the comparative efficacy of different counterspeech strategies. Bilewicz et al. (2021) find that three of their intervention types ('disapproval', 'abstract norm', 'empathy') are effective in reducing verbal violence when compared with no intervention at all. Here, empathy had the weakest effect, which they put down to the empathetic messages being specific to particular behaviours, limiting their capacity to modify aggression towards wider targets. Hangartner et al. (2021) also found that empathy-based counterspeech can consistently reduce hate speech, although this effect is small. And Carthy and Sarma (2021) found that counterspeech that seeks to correct false information in the hate speech actually leads to higher levels of violence legitimisation, while having participants actively counter terrorist rhetoric themselves ('Tailored Counter-Narrative') was the most effective strategy to reduce this. They found counterspeech to be more effective on participants that are already predisposed to cognitive reflection. However, focusing on the effect of factual correction on the victims rather than perpetrators of hate speech, Obermaier et al. (2021) found it to be effective in providing support and preventing them from hating back and therefore widening the gap between groups.

There is also some evidence that the numbers of the different actors involved in a counterspeech exchange can affect an intervention's success. Schieb and Preuss (2016) find that counterspeech can impact the online behaviour of (simulated) bystanders, with the effectiveness strongly influenced by the proportions of hate and counter speakers and neutral bystanders. According to their model, a small number of counterspeakers can be effective against smaller numbers of hate speakers in the presence of larger numbers of people lacking strong opinions. Saltman et al. (2021) found their counterspeech strategies to be effective only for higher risk individuals within the target populations, although they did not see any of the potential negative effects of counterspeech (such as increased radicalisation) reported elsewhere.

Focusing on who in particular delivers counterspeech, Munger (2017) finds that success of counterspeech depends on the identity and status of the speaker. However, with only a small positive effect, Bélanger et al. (2020) found that the content of counterspeech was more important than the source. And Garland et al. (2022) found that, while organised counterspeech can be effective, the efforts of individuals can lead to increases in hate speech. In Buerger (2021b), members of #jagärhär claim that their counterspeech interventions were successful in making space for alternative viewpoints to hate speech.

6 Computational Approaches to Counterspeech

In this section, we switch the focus to look at NLP literature on counterspeech emerging from the field of computer science. We tackle three subjects in particular: the datasets being used in these studies, approaches to counterspeech detection, and approaches to counterspeech generation.

6.1 Counterspeech Datasets

Collection strategies Approaches for counterspeech collection focus on gathering two different kinds of datasets: spontaneously produced comments crawled from social media platforms, and deliberately created responses aiming to contrast hate speech. In the first case, content is retrieved based on keywords/hashtags related to targets of interest (Mathew et al., 2018; Vidgen et al., 2020; He et al., 2022; Vidgen et al., 2021) or from pre-defined counterspeech accounts (Garland et al., 2020). In principle, due to the easily accessible API required for data retrieval, the majority of datasets are collected from social media platforms including Twitter (Mathew et al., 2018; Procter et al., 2019; Garland et al., 2020; Kennedy et al., 2020; Vidgen et al., 2020; He et al., 2022; Goffredo et al., 2022; Toliyat et al., 2022; Lin et al., 2022), and only a few are retrieved from Youtube (Mathew et al., 2019; Kennedy et al., 2020; Priyadharshini et al., 2022) and Reddit (Kennedy et al., 2020; Vidgen et al., 2021; Lee et al., 2022; Yu et al., 2022), respectively (though again it is worth noting that at the time of writing the Twitter API was starting to become a lot less accessible). To find the best strategy for collecting online content, Möhle et al. (2023) compare the keywords-matching method with automated filtering using a multilingual model fine-tuned on English data for German counterspeech collection. They found neither strategy helped curate significantly more counterspeech compared to a random sampling baseline.

In the second category, counterspeech is written by crowd workers (Qian et al., 2019) or operators expert in counterspeech writing (Chung et al., 2019, 2021c). While such an approach is expected to offer relatively controlled and tailored responses, writing counterspeech from scratch is time-consuming and requires human effort. To address this issue, advanced generative language models are adopted to automatically produce counterspeech (Tekiroğlu et al., 2020; Fanton et al., 2021; Bonaldi et al., 2022), as we will discuss further below.

Granularity and languages Regarding granularity of taxonomies, most existing datasets provide binary annotation (counterspeech/non-counterspeech) (Garland et al., 2020; Vidgen et al., 2020; He et al., 2022; Vidgen et al., 2021), while three datasets feature annotations of the types of counterspeech (Mathew et al., 2018, 2019; Chung et al., 2019). Recently, Yu et al. (2023) propose a taxonomy that distinguishes the target of counterspeech (i.e. whether the counterspeech addresses the hateful content or the author of the hateful comment) and identifies the argument components in the counterspeech (i.e. logical arguments and appealing to emotion). In terms of hate incidents, datasets are available for several hate phenomena such as Islamophobia (Chung et al., 2019) and East Asian prejudice during the COVID-19 pandemic (Vidgen et al., 2020; He et al., 2022). The aforementioned datasets are mostly collected and analyzed at the level of individual text, not at discourse or conversations (e.g., multi-turn dialogues (Bonaldi et al., 2022)). Most of the datasets are in English, while only a few target multilinguality, including Italian (Chung et al., 2019; Goffredo et al., 2022), French (Chung et al., 2019), Spanish (Vallecillo-Rodríguez et al., 2023), German (Garland et al., 2020; Möhle et al., 2023), and Tamil (Priyadharshini et al., 2022).

6.2 Approaches to Counterspeech Detection

Previous work on counterspeech detection has focused on binary classification (i.e. whether a text is counterspeech or not) (Vidgen et al., 2020; Garland et al., 2022; He et al., 2022) or identifying the types of counterspeech as a multi-label task (Mathew et al., 2018; Garland et al., 2020; Chung et al., 2021a; Goffredo et al., 2022). Automated classifiers are developed to analyse large-scale social interactions of abuse and counterspeech addressing topics such as political discourse (Garland et al., 2022) and multi-hate targets (Mathew et al., 2018). Moving beyond monolingual study, Chung et al. (2021a) evaluate the performance of pre-trained language models for categorising counterspeech strategy for English, Italian and French in monolingual, multilingual and cross-lingual scenarios.

6.3 Approaches to Counterspeech Generation

Various methodologies have been put forward for the automation of counterspeech generation (Qian et al., 2019), addressing various aspects including the efficacy of a hate countering platform (Chung et al., 2021b), informativeness (Chung et al., 2021c), multilinguality (Chung et al., 2020), politeness (Saha et al., 2022), and grammaticality and diversity (Zhu and Bhat, 2021). These methods are generally centred on transformer-based large language models (e.g., GPT-2 (Radford et al., 2019)). By testing various decoding mechanisms using multiple language models, Tekiroğlu et al. (2022) find that autoregressive models combined with stochastic decoding yield the optimal counterspeech generation. In addition to tackling hate speech, there are studies investigating automatic counterspeech generation to respond to trolls (Lee et al., 2022) and microagressions (Ashida and Komachi, 2022).

Evaluation of counterspeech generation Assessing counter speech generation is complex and challenging due to the lack of clear evaluation criteria and robust evaluation techniques.

Previous work evaluates the performance of counterspeech systems via two aspects: automatic metrics and human evaluation. Automatic metrics, generally, evaluate the generation quality based on criteria such as linguistic surface (Papineni et al., 2002; Lin, 2004), novelty (Wang and Wan, 2018), and repetitiveness (Bertoldi et al., 2013; Cettolo et al., 2014). Despite being scalable, these metrics are uninterpretable and can only infer model performance according to references provided (e.g., dependent heavily on exact word usage and word order) and gathering an exhaustive list of all appropriate counterspeech is not feasible. For this reason, such metrics cannot properly capture model performance, particularly for open-ended tasks (Liu et al., 2016; Novikova et al., 2017) including counterspeech generation. As a result, human evaluation is heavily employed based on aspects such as suitableness, grammatical accuracy and relevance (Chung et al., 2021c; Zhu and Bhat, 2021). Despite being trusted and high-performing, human evaluation has inherent limitations such as being costly, difficult (e.g., evaluator biases and question formatting), and time-consuming (both in terms of evaluation and moderator training), and can be inconsistent and inflict psychological harm on the moderators.

The effectiveness of counterspeech generations should be also carefully investigated 'in-the-wild' to understand its social media impact, reach of content, and the dynamics of hateful content and counterspeech (see Section 5). This line of research is limited. The closest work to this research space is by Zheng et al. (2023) that identifies the characteristics of good counterspeech in terms of the quality and effectiveness and user preference for machine-generated counterspeech through a survey. Based on 29 subjects (i.e. bystanders) evaluating 60 pseudo-threads on Twitter (at the time of experiments), they conclude that clear and direct responses with thorough explanations are mostly preferred by users.

Potentials and limits of existing generative models We believe that in some circumstances counterspeech may be a more appropriate tool than content moderation in fighting hate speech as it can depolarise discourse and show support to victims. However, automatic counterspeech generation is a relatively new research area. Recent progress in natural language processing has made large language models a popular vehicle for generating fluent counterspeech. However, counterspeech generation currently faces several challenges that may constrain the development of efficient models and hinder the deployment of hate intervention tools. Similar to the use of machine translation and email writing tools, we advocate that counterspeech generation tools should be deployed as suggestion tools to assist in hate countering activity (Chung et al., 2021c,b).

- ***Faithfulness/Factuality in generation*** Language models are repeatedly reported to produce plausible and convincing but not necessarily faithful/factual statements (Solaiman et al., 2019; Zellers et al., 2019; Chung et al., 2021c). We refer to faithfulness as being consistent and truthful in adherence to the given source (i.e. model inputs) (Ji et al., 2023). Such unfaithful/non-factual generation is particularly intolerable for counterspeech generation as it can create unwanted consequences or elicit hatred. Many attempts have been made to mitigate this issue (Ji et al., 2023) such as correcting unfaithful data (Nie et al., 2019) and measuring faithfulness of generated outputs (Dušek and Kasner, 2020; Zhou et al., 2021). For the task of counterspeech generation, Chung et al. (2021c) present the first knowledge-bound generation pipeline consisting of a knowledge retrieval module that retrieves relevant knowledge to the context of hate speech and a generation module that generates a counterspeech response. Following this approach, Jiang et al. (2023) employ a retrieval-augmented unsupervised generation method that refines retrieved knowledge based on stance consistency and semantic overlap for hate speech and allows for generation without gold-standard data. In a similar vein, Furman et al. (2023) prompt large language models with argumentative information in hate speech to enhance the quality of counterspeech generation and show that this approach is especially beneficial for low-resource scenarios. To facilitate reliable counterspeech generation applications, we encourage reporting the faithfulness/factuality of models.

- ***Toxic degeneration and debiasing*** Language models can also induce unintendedly biased and/or toxic content, regardless of whether explicit prompts are used (Dinan et al., 2022). In the use case of counterspeech generation, this can result in harm to victims and bystanders as well as risking provoking perpetrators into further abusive behaviour. This issue has been mitigated by two approaches: data and modelling. The data approach aims at creating proper datasets for fairness by removing undesired and biased content (Blodgett et al., 2020; Raffel et al., 2020). The modelling approach focuses on controllable generation techniques that, for instance, employ humans for post-editing (Tekiroğlu et al., 2020) and detoxification techniques (Gehman et al., 2020). Another line of research emphasises that implicit stereotypical beliefs or biases from hateful content should be addressed in counterspeech generation (Mun et al., 2023; Akazawa et al., 2023). For instance, Akazawa et al. (2023) tune large language models to infer implicit biases from hate speech and found that such extra information helps improve generation quality.

- ***Diversity, Generalisation and Specialisation*** With the rise of online hate, models that can generalize across domains would help produce counterspeech involving new topics and events, while it may come with the cost of losing specificity. Generalisable methods can ameliorate the time and manual effort required for collecting and annotating data. However, as discussed in Section 5, counterspeech is multifaceted and contextualised. For instance, abuse against women can often be expressed in a more subtle form as microaggressions. Specific and diverse responses to hateful or prejudiced language are often preferred as they can provide coherent discourse relations and potential connection with personal events (Finnegan et al., 2015). In a user study comparing model-generated and human-written counterspeech, Mun et al. (2023) show that humans prefer and use more specific strategies targeting stereotypical statements when countering hate while models tend to produce less convincing arguments according to annotators. To produce more specific responses, Hassan and Alikhani (2023) show that grounding generation in context using discourse-augmented prompting strategies results in contextual, diverse and accurate counterspeech. Similarly, Gupta et al. (2023) propose to guide generation based on five intents (informative, question, denouncing, humour, and positive) for generating diverse counterspeech. To address the generalisation capabilities of large language models for counterspeech generation, Bonaldi et al. (2023) introduce attention-based regularisation techniques that help contextualise token representations (i.e. include broader hate speech context) and guide models to focus on specific attention distributions (e.g. use words related to minority targets). There may not be a one-size-fits-all solution. Overall, model generalisability is still challenging (Fortuna et al., 2021; Yin and Zubiaga, 2021), and can have potential limitations (Conneau et al., 2020; Berend, 2022). Finding the right trade-off between generalisation and specialisation is key.

7 Future Perspectives

Of the many promising abuse intervention experiments that we review, results are not always consistent, demonstrating weak claims or limited success (applicable only to certain settings). Possible reasons include short-term experiments, small sample sizes and non-standardised experimental designs. To improve this, effective interventions should come with the characteristics of scalability, durability, reliability, and specificity. In this section, we highlight key distinctions and overlaps across areas that have and have not been explored in social sciences and computer science, discuss ethical issues related to evaluating counterspeech in real-life settings and automating the task of counterspeech generation, and identify best practices for future research.

Distinctions and overlaps across areas By recognizing the commonalities and differences between social sciences and computer science, we pinpoint the unique contributions of each discipline and encourage interdisciplinary collaborations to address complex societal challenges and better understand human behaviour with the help of computational systems.

- ***Terminological clarity*** Throughout the counterspeech literature, terminology is used inconsistently. Terms such as counterspeech and counter-narratives are often used interchangeably or used to refer to similar concepts. In social science, counterspeech is used to refer to content that disagrees with abusive discourses and counter-narratives often entail criticism of an ideology with logical reasoning. As a result, counter-narrative stimuli designed in social experiments are generally long form (Bélanger et al., 2020). In computer science on the other hand, the distinctions between counterspeech and counter-narratives have been vague, and training data is generally short form (while this may be bound by character limit on social media platforms). For instance, short and generic responses such as '*How can you say that about a faith of 1.6 billion people?*' can be commonly found in counter-narrative datasets (Chung et al., 2019).

- ***The focus of evaluation*** Social scientists and counterspeech practitioners generally attempt to understand and assess the impact of counterspeech on reducing harms (e.g., which strategies are effective and public perception towards counterspeech), whereas computer scientists focus more on technical exploration of automated systems and testing their performance in producing counterspeech (e.g., comparing system outputs with a pre-established ground truth or supposedly ideal output). One commonality between the social science and computer science studies is that most findings are drawn from controlled and small-scale studies. Applying interventions to real-world scenarios is a critical next step.

- ***Datasets*** Dataset creation is an important component in computer science for developing machine learning models for generating counterspeech, while such contributions are less commonly considered in social sciences which rely on experiments using hand-crafted stimuli and one-time analyses of their effectiveness.

- ***Scope of research*** We observe that, while computer scientists have focused on responses to abusive language and hate speech, social science studies address a wider range of phenomena, in particular radicalisation and terrorist extremism. It can be difficult to measure the effectiveness of counterspeech in challenging these over the short term, leading to some of the differences in evaluation metrics across disciplines.

- ***Lack of standardised methodologies*** A variety of methodologies have been adopted in the literature, making comparisons across studies difficult. Without standardised evaluations, it is difficult to situate the results and draw robust findings.

Ethical Issues, Risks and Challenges of Conducting Counterspeech Studies Effective evaluation of counterspeech not only identifies users who may need help, but also safeguards human rights and reinforces a stronger sense of responsibility in the community. This discussion is based on the authors' opinion and not stemming from the review.

- ***Evaluating counterspeech in real-life settings*** Conducting the evaluation of counterspeech in real-world scenarios appears to provide a proactive and quick overview of its performance on hate mitigation. Nevertheless, the best ways to approach this remains an open question. For instance, one side argues about the morality of exposing participants to harm, while another points to the importance of internet safety. Exercising counterspeech can offer mitigation of online abuse in good faith and there are legal groundings that can potentially be applied to encourage such an action. As an example, **Good Samaritan laws** provide indemnity to people who assist others in danger (Smits, 2000). These safeguards aim to ensure that individuals are not hesitant to help others in distress due to the fear of facing legal consequences in case of unintentionally making errors in their efforts to provide support. In 2017 the EU Commission released a communication emphasizing the need to tackle illegal content online, stating that '*This Communication ... aims to provide clarifications to platforms on their liability when they take proactive steps to detect, remove or disable access to illegal content (the so-called "Good Samaritan" actions)*' (Commission, 2017). We argue that this statement can be extended to the scenario of applying counterspeech to online hate mitigation.

Responsible open-source research can facilitate reproducibility and transparency of science. Recently, reproducible research has been deemed critical in both social sciences (Stroebe et al., 2012; Derksen and Morawski, 2022) and computer science, and low replication success is found despite using materials provided in the original papers (Belz et al., 2023; Collaboration, 2015). To tackle this issue, a few initiatives for transparent research have been proposed, advocating researchers to state succinctly in papers how experiments are conducted (e.g., stimuli, mechanisms for data selection) and evaluated, including A 21 Word Solution (Simmons et al., 2012) and Open Science Framework.[5] Furthermore, practising data sharing encourages researchers to be responsible for fair and transparent experimental designs, and to avoid subtle selection biases that might affect substantive research questions under investigation (Dennis et al., 2019). At the same time, when handling sensitive or personal information, data sharing should adhere to research ethics and privacy standards (Dennis et al., 2019; de la Cueva and Méndez, 2022). For instance, in the case of hate speech, using synthetic examples or de-identification techniques is considered a good general practice for ensuring the safety of individuals (Kirk et al., 2022).

• ***Automating counterspeech generation*** There are several ethical challenges related to automating the task of counterspeech generation. First of all, there is the danger of dual-use: the same methodology could also be used to silence other voices.

Furthermore, effective and ethical counterspeech relies on the accuracy and robustness of detecting online hate speech: an innocent speaker may be publicly targeted and shamed if an utterance is falsely classified as hate speech – either directly or indirectly as in end-to-end response generation. For example, Google's Jigsaw API (Google Jigsaw, 2022), a widely used tool for detecting toxic language, makes predictions that are aligned with racist beliefs and biases—for example it is less likely to rate anti-Black language as toxic, but more likely to mark African American English as toxic (Sap et al., 2022). It is thus important to make sure that the underlying tool is not biased and well-calibrated to the likelihood that an utterance was indeed intended as hate speech. For example, the 'tone' of counterspeech could be used to reflect the model's confidence.

A related question is free speech: what counts as acceptable online behaviour, what sort of speech is deemed inappropriate, in which contexts, and should be targeted by counterspeech? A promising direction for answering this complex question is participatory design to empower the voices of those who are targeted (Birhane et al., 2022).

In sum, there is a trade-off between risks and benefits of counterspeech generation. Following the 'Good Samaritan' law: automating counterspeech provides timely help to victims in an emergency which is protected against prosecution (even if it goes wrong). Similar legislation is adopted by other countries, including the European Union, Australia and the UK. Under this interpretation, well-intentional counterspeech (by humans and machines) is better than doing nothing at all.

Best practices We provide best practices for developing successful intervention tools.

1. Bear in mind practical use cases and scenarios of hate-countering tools. A single intervention strategy is unlikely to diminish online harm and successful counterspeech interventions would benefit from personalisation. To design successful counterspeech tools, it is important to consider the purposes of counter messages (e.g., support victims and debunk stereotypes), the speakers (e.g., practitioners, authorities and high-profile people), recipients (e.g., ingroup/outgroup, political background and education level), the content (e.g., strategy, style, and tones), intensity (e.g., one message per week/month), and the communication medium (e.g., videos, text, and platforms).

2. Look beyond automated metrics and consider deployment settings for evaluating the performance of generation systems. Generation systems are generally evaluated on test sets in a controlled environment using accuracy-based metrics (e.g., ROUGE and BLEU) that cannot address social implications of a system. Drawn from social science studies, metrics assessing social impact (e.g., user engagement), behavioural change (e.g., measure abuse reduction in online discourse) and attitude change (e.g., through self-description questionnaires) can be considered. A good intervention system is expected to pertain long-lasting effects.

3. Be clear about the methodology employed in experiments, open-source experimental materials (e.g., stimuli, questionnaires and codebook), and describe the desirable criteria for evaluating counterspeech intervention. As standardised procedures are not yet established for the assessment of counterspeech interventions, examining the impact of interventions becomes difficult. A meaningful description of experimental design would therefore enhance reproducible research and help capture the limitation of existing research.

4. Establish interdisciplinary collaboration across areas such as counter-terrorism, political science, psychology and computer science. AI researchers can help guide policymakers and practitioners to, for instance, identify long-term interventions by performing large-scale data analysis using standardized procedures on representative and longitudinal samples. With expertise in theories of human behaviour change and experimental design, social science researchers can conduct qualitative

[5] https://osf.io/

evaluations of AI intervention tools in real-life scenarios to understand their social impact.

8 Conclusion

Online hate speech is a pressing global issue, prompting scientists and practitioners to examine potential solutions. Counterspeech, content that directly rebuts hateful content, is one promising avenue. While NLP researchers are already beginning to explore opportunities to automate the generation of counterspeech for the mitigation of hate at scale, research from the social sciences points to many nuances that need to be considered regarding the impact of counterspeech before this intervention is deployed. Taking an interdisciplinary approach, we have attempted to synthesize the growing body of work in the field. Through our analysis of extant work, we suggest that findings regarding the efficacy of counterspeech are highly dependent on several factors, including methodological ones such as study design and outcome measures, and features of counterspeech such as the speaker, target of hate, and strategy employed. While some work finds counterspeech to be effective in lowering further hate generation from the perpetrator and raising feelings of empowerment in bystanders and targets, others find that counterspeech can backfire and encourage more hate. To understand the advantages and disadvantages of counterspeech more deeply, we suggest that empirical research should focus on testing counterspeech interventions in real-world settings which are scalable, durable, reliable, and specific. Researchers should agree on key outcome variables of interest in order to understand the optimal social conditions for producing counterspeech at scale by automating its generation. We hope that this review helps make sense of the variety of types of counterspeech that have been studied to date and prompts future collaborations between social and computer scientists working to ameliorate the negative effects of online hate.

Acknowledgements

We thank Bertie Vidgen for the valuable feedback on the initial structure of this manuscript and Hannah Rose Kirk for her help with the collection of target literature.

Gavin Abercrombie and Verena Rieser were supported by the EPSRC projects ‘Equally Safe Online’ (EP/W025272/1 and EP/W025493/1) ‘Gender Bias in Conversational AI’ (EP/T023767/1). Yi-Ling Chung, Florence Enock and Jonathan Bright were supported by the Ecosystem Leadership Award under the EPSRC Grant EPX03870X1 and The Alan Turing Institute.

References

Adak, Sayantan, Souvic Chakraborty, Paramita Das, Mithun Das, Abhisek Dash, Rima Hazra, Binny Mathew, Punyajoy Saha, Soumya Sarkar, and Animesh Mukherjee. 2022. Mining the online infosphere: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1453.

Akazawa, Nami, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Distilling implied bias from hate speech for counter narrative selection. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 29–43, Prague, Czechia. Association for Computational Linguistics.

Alsagheer, Dana, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

Ashida, Mana 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)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Bartlett, Jamie and Alex Krasodomski-Jones. 2015. Counter-speech examining content that challenges extremism online. *DEMOS, October*.

Belz, Anya, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Benesch, Susan. 2014a. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: US Holocaust Memorial Museum*.

Benesch, Susan. 2014b. Defining and diminishing hate speech. *State of the world’s minorities and indigenous peoples*, 2014:18–25.

Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A field study. *Dangerous Speech Project*.

Berend, Gábor. 2022. Combating the curse of multilinguality in cross-lingual WSD by aligning sparse contextualized word representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2459–2471, Seattle, United States. Association for Computational Linguistics.

Bertoldi, Nicola, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Bilewicz, Michał, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, 47(3):260–266.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bonaldi, Helena, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.

Bonaldi, Helena, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. *arXiv preprint arXiv:2211.03433*.

Buerger, Catherine. 2021a. Counterspeech: A literature review. *Available at SSRN 4066882*.

Buerger, Catherine. 2021b. #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, 7(4):20563051211063843.

Buerger, Catherine. 2022. Why they do it: Counterspeech theories of change. *Available at SSRN 4245211*.

Bélanger, Jocelyn J., Claudia F. Nisa, Birga M. Schumpe, Tsion Gurmu, Michael J. Williams, and Idhamsyah Eka Putra. 2020. Do counter-narratives reduce support for isis? yes, but not for their target audience. *Frontiers in Psychology*, 11.

Carthy, S. L. and K. M. Sarma. 2021. Countering terrorist narratives: Assessing the efficacy and mechanisms of change in counter-narrative strategies. *Terrorism and Political Violence*, 0(0):1–25.

Carthy, Sarah L, Colm B Doody, Katie Cox, Denis O'Hora, and Kiran M Sarma. 2020. Counter-narratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3):e1106.

Cettolo, Mauro, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Chakravarthi, Bharathi Raja. 2022. Multilingual hope speech detection in english and dravidian languages. *International journal of data science and analytics*, 14(4):389—406.

Chaudhary, Mudit, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941*.

Chung, Yi-Ling, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chung, Yi-Ling, 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*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Chung, Yi-Ling, Serra S. Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Chung, Yi-Ling, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Online.

Chung, Yi-Ling, Serra Sinem Tekiroğlu, and Marco Guerini. 2021c. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

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

Collaboration, Open Science. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Commission, European. 2017. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

de la Cueva, Javier and Eva Méndez. 2022. Open science and intellectual property rights. how can they better interact? state of the art and reflections. report of study. european commission.

Dennis, Simon, Paul Garrett, Hyungwook Yim, Jihun Hamm, Adam F Osth, Vishnu Sreekumar, and Ben Stone. 2019. Privacy versus open science. *Behavior research methods*, 51:1839–1848.

Derksen, Maarten and Jill Morawski. 2022. Kinds of replication: Examining the meanings of “conceptual replication” and “direct replication”. *Perspectives on Psychological Science*, 17(5):1490–1505. PMID: 35245130.

Dinan, Emily, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Dušek, Ondřej and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ernst, Julian, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer, Gary Bente, and Hans-Joachim Roth. 2017. Hate beneath the counter speech? A qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.

Fanton, Margherita, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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)*, pages 3226–3240, Online. Association for Computational Linguistics.

Ferguson, Kate. 2016. Countering violent extremism through media and communication strategies: A review of the evidence.

Finnegan, Eimear, Jane Oakhill, and Alan Garnham. 2015. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Frontiers in Psychology*, 6.

Fortuna, Paula, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Frenkel, Sheera and Kate Conger. 2022. Hate Speech’s Rise on Twitter Is Unprecedented, Researchers Find. *The New York Times*.

Furman, Damián, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.

Garland, Joshua, 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*, pages 102–112, Online. Association for Computational Linguistics.

Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Goffredo, Pierpaolo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Google Jigsaw. 2022. Perspective API. Accessed: 26 May 2023.

Gupta, Rishabh, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Hangartner, Dominik, 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):e2116310118.

Hassan, Sabit and Malihe Alikhani. 2023. DisCGen: A framework for discourse-informed counterspeech generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–429, Nusa Dua, Bali. Association for Computational Linguistics.

He, Bing, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 90–94, New York, NY, USA. Association for Computing Machinery.

Iqbal, Khuram, Saad Kalim Zafar, and Zahid Mehmood. 2019. Critical evaluation of pakistan's counter-narrative efforts. *Journal of Policing, Intelligence and Counter Terrorism*, 14(2):147–163.

Jay, Timothy. 2009. Do offensive words harm people? *Psychology, public policy, and law*, 15(2):81.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Jiang, Shuyu, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Kennedy, Chris J, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Kirk, Hannah, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leader Maynard, Jonathan and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, 9(3).

Lee, Huije, Young Ju NA, Hoyun Song, Jisu Shin, and Jong C. Park. 2022. Elf22: A context-based counter trolling dataset to combat internet trolls. In *Proceedings of the 13th Language Resources and Evaluation, LREC 2022, Marseille, France, June 20-25, 2022*, pages 3530–3541. European Language Resources Association.

Leonhard, Larissa, Christina Rueß, Magdalena Obermaier, and Carsten Reinemann. 2018. Perceiving threat and feeling responsible. how severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on facebook. *Studies in Communication and Media*, 7(4):555–579.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, Hao, Pradeep Nalluri, Lantian Li, Yifan Sun, and Yongjun Zhang. 2022. Multiplex anti-Asian sentiment before and during the pandemic: Introducing new datasets from Twitter mining. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 16–24, Dublin, Ireland. Association for Computational Linguistics.

Litaker, John R., Carlos Lopez Bray, Naomi Tamez, Wesley Durkalski, and Richard Taylor. 2022. Covid-19 vaccine acceptors, refusers, and the moveable middle: A qualitative study from central texas. *Vaccines*, 10(10).

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Mathew, Binny, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv:1812.02712*.

Mathew, Binny, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4):264–269. PMID: 19622511.

Möhle, Pauline, Matthias Orlikowski, and Philipp Cimiano. 2023. Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 44–61, Prague, Czechia. Association for Computational Linguistics.

Mun, Jimin, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Munger, Kevin. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

News, BBC. 2018. MPs 'being advised to quit Twitter' to avoid online abuse. *BBC News*.

Nie, Feng, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Obermaier, Magdalena, Desirée Schmuck, and Muniba Saleem. 2021. I'll be there for you? effects of islamophobic online hate speech and counter speech on muslim in-group bystanders' intention to intervene. *New Media & Society*, 0(0):14614448211017527.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Poole, Elizabeth, Eva Haifa Giraud, and Ed de Quincey. 2021. Tactical interventions in online hate speech: The case of #stopislam. *New Media & Society*, 23(6):1415–1442.

Priyadharshini, Ruba, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Procter, Rob, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matt Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. *arXiv preprint arXiv:1908.11732*.

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Reynolds, Louis and Henry Tuck. 2016. The counter-narrative monitoring & evaluation handbook. *Institute for Strategic Dialogue*.

Riedl, Martin J., Gina M. Masullo, and Kelsey N. Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, 107:106262.

Saha, Koustuv, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 255–264, New York, NY, USA. Association for Computing Machinery.

Saha, Punyajoy, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech.

Saltman, Erin, Farshad Kooti, and Karly Vockery. 2021. New models for deploying counterspeech: Measuring behavioral change and sentiment analysis. *Studies in Conflict & Terrorism*, 0(0):1–24.

Saltman, Erin Marie and Jonathan Russell. 2014. White paper–the role of Prevent in countering online extremism. *Quilliam publication*.

Sap, Maarten, 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*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Schieb, Carla and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ICA annual conference, at Fukuoka, Japan*, pages 1–23.

Siegel, Alexandra A. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

Silverman, Tanya, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2012. A 21 word solution. *Available at SSRN 2160588*.

Smits, Jan M. 2000. The good samaritan in european private law; on the perils of principles without a programme and a programme for the future.

Snyder, C. R., Kevin L. Rand, and David R. Sigmon. 2018. Hope Theory: A Member of the Positive Psychology Family. In *The Oxford Handbook of Hope*, pages 257–276. Oxford University Press.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Stroebe, Wolfgang. 2008. Strategies of attitude and behaviour change.

Stroebe, Wolfgang, Tom Postmes, and Russell Spears. 2012. Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7(6):670–688.

Stroud, Scott R and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. *Mediating Misogyny: Gender, Technology, and Harassment*, pages 293–310.

Tekiroğlu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Tekiroğlu, Serra Sinem, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Toliyat, Amir, Sarah Ita Levitan, Zheng Peng, and Ronak Etemadpour. 2022. Asian hate speech detection on twitter during covid-19. *Frontiers in Artificial Intelligence*, 5.

Tuck, Henry and Tanya Silverman. 2016. *The counter-narrative handbook*. Institute for Strategic Dialogue.

Vallecillo-Rodríguez, Maria Estrella, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71:227–245.

Vidgen, Bertie, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Vidgen, Bertie, Helen Margetts, and Alex Harris. 2019. How much online abuse is there? A systematic review of evidence for the UK. *Alan Turing Institute Policy Briefing*.

Vidgen, Bertie, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Wang, Ke and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

Wright, Lucas, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Yin, Wenjie and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Yu, Xinchen, 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*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Yu, Xinchen, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289, Singapore. Association for Computational Linguistics.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zheng, Yi, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Zhou, Chunting, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zhu, Wanzheng and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.