SILVIA DONZELLI
Bielefeld University
silvia_donzelli@yahoo.com

COUNTERING HARMFUL SPEECH ONLINE. (IN)EFFECTIVE STRATEGIES AND THE DUTY TO COUNTERSPEAK

abstract

The concept of counterspeech denotes a non-coercive and non-censoring method for reacting to harmful speech, with the aim of impeding or at least diminishing its damaging effects. Remarkable work is being done by researchers and activist groups on elaborating practical strategies of countering hate speech online. Though, research in moral and political philosophy exploring the effectivity of counterspeech and grounding the reasons for engaging in it still remains in its early stages. In the following paragraphs I will address recent contributions which elaborate on the viability and normative aspects of counterspeech. Outlining their valuable insights, but also their failure to give due importance to the peculiarities of online speech dynamics, I will highlight relevant features on which future research about online harmful speech and counterspeech can build.

keywords

counterspeech, digital harmful speech, social media, bystander, complicity

Phenomenology and Mind, n. 20 - 2021, pp. 76-87
DOI: https://doi.org/10.17454/pam-2007
https://www.rosenbergesellier.it/eng/journals/phenomenology-and-mind

ISSN 2280-7853 (print) - ISSN 2239-4028 (on line)

1. Counterspeech goes digital

Recent decades have seen a growing body of theoretical and empirical research exploring the relationship between speech and harm. Along with the undoubted conceptual interest, the topic of harmful speech[1] raises egregious moral and political issues, especially when it comes to elaborating viable strategies of harm prevention.

Strategies to address harmful speech can be broadly classified into three main approaches: legal sanctions, content takedown regulations, and counterspeech. The first two strategies have a censoring nature. For this reason, they can come into conflict with the core democratic value of freedom of speech, raising challenging issues of political and economic coercion, definition and evidence of speech-related harm, proportionality requirements in balancing competing values, among others.

The third strategy, counterspeech, lacking a censoring approach, does not seem to pose similar normative issues and can therefore appear as an appealing and drawback-free solution. In spite of this, counterspeech is by far not unproblematic, as researchers have already pointed out, though mainly in the context of offline harmful speech.

The rise of digital communication to an overall pervasive phenomenon is bringing new challenges to harmful speech theory and policy. This does not mean that approaches from law, philosophy and social psychology, which have been designed with offline speech dynamics in mind, should now be dismissed. Rather, they should be up-graded, taking into account the particular ways in which digital speech is impacting on individuals, groups and socio-cultural structures.

In the following, I will deal with recent contributions on harmful speech and counterspeech, highlighting their valuable insights, but also their failure to give due importance to the distinguishing features of digital speech. I will start by outlining feminist philosophy of language insights on harmful speech and counterspeech strategies (2), as well as Lepoutre alternative proposal (3). Then, I will argue that both approaches fail to consider some specific features of digital speech dynamics profoundly affecting harmful speech force, and I offer one example of counterspeech strategy tailored for digital speech challenges (4). Finally, I will deal with the normative question of how to ground a possible individual moral duty to engage in

1 For present purposes, I define harmful speech broadly, comprising legally defined hate speech and incitement to hatred as well as forms of discriminatory speech which do not fit legal standards. Moreover, part of the paper deals with theories about speech constituting, as opposed to causing, harm.

counterspeech. Both Howard's proposal drawing on positive duties and the idea of complicity by failure to counterspeak will be considered, as well as their implications for offline and online speech (5-6).

2. Insights from philosophy of language

The concept of counterspeech denotes a non-coercive and non-censoring method for reacting to harmful speech, with the aim of impeding or at least diminishing its damaging effects. Usually, advocates of counterspeech do not exclude legal bans or content takedown as complementary strategies in order to achieve such goals. Yet, they trust in the power of speech as an effective and democratic valuable remedy to speech-related harms. Speech can be harmful in various ways – it can directly harm the feelings of the audience (both targeted and not), induce to harmful actions, contribute to the establishment and functioning of structural injustice. According to philosophers of language, speech can also constitute, as opposed to cause, harm, if certain conditions are given. I will start by considering this approach, since it offers valuable insights both on the ways in which speech can be harmful and on related counterspeech strategies – though, focusing on offline speech, it is not entirely able to match digital speech challenges, as should become clear in section 4.

Integrating philosophy of language conceptual resources with feminist critical thinking about language and systemic oppression, Rae Langton's seminal work elaborates on the topic of speech acts, specifically on the ways in which speech can *constitute* harm. Langton adopts John Langshaw Austin's distinction between locution (the content of an utterance), illocution (what the speaker does with the utterance, like ordering or promising) and perlocution (the multifarious causal effects of the utterance), in order to show that pornography can not only have harmful perlocutionary effects, but can indeed have normative illocutionary force, constituting a speech act subordinating and silencing women (Langton, 1993).

An utterance can have illocutionary force and thus constitute harm if some conditions are met (felicity conditions). Langton focuses mainly on authority felicity condition, particularly on speaker authority. To grant illocutionary force to an utterance, authority does not need to be institutional, nor formal. It suffices that the speaker has authority in the specific contextual domain in which speech is uttered.

There are various sources of authority, but I will focus here on one particular source, namely authority gained by default. As Ishani Maitra (2009) shows, a speaker can obtain authority in a given activity by simply assuming to have it: if the other participants and bystanders do not raise objections, the speaker's presupposed authority can become factual by default. Bystanders' silence in the face of harmful speech can have a licensing effect, i. e. the effect of conveying acceptance of speaker's authority. Failing to object can legitimize authority even if such an effect is unintended and even in case of dissent, if dissent remains unexpressed.

The crucial role of silence, understood as the omission of objection, in granting force to harmful speech has been explored also by Mary Kate McGowan, yet from a different perspective. McGowan focuses on one particular way in which speech can constitute harm, namely the dynamic process of adjustment and accommodation occurring within conversations (McGowan 2018; 2019). Accordingly, every contribution to a conversation can modify the conversational score, which is defined as the set of all elements counting as appropriate for that conversation. This includes, but is not limited to, the psychological and epistemic common ground of participants. In order for a contribution to change the conversational score, no authority is required. Indeed, changes in the conversational score occur routinely and mostly in a covert, "sneaky" way, remaining unnoticed by speakers and hearers. If none of the participants raises an objection, score changes are automatically accommodated in the conversation. This applies also to harmful presuppositions and associations, like sexist remarks, which, if not challenged, are automatically accommodated.

Utterances triggering implicit bias like "C'mon, even a woman could do it!" (Ayala & Vasileva, 2016, p. 206) are a case in point.

It should be noted that both authority felicity condition and conversational accommodation appear to hinge, at least in part, on the failure of hearers to object. This can have notable normative consequences: as Maitra notes, if remaining silent in the face of harmful speech can contribute to it, a moral obligation to speak up could follow (Maitra 2009, p. 20). This idea, as well as its implications for online speech, will be sketched below in section 6.

Philosophy of language insights downright suggest a particular counterspeech strategy. Since speech can constitute harm only in case certain felicity conditions are met, making such conditions unavailable would hinder speech-related harm. Langton talks of "blocking" harmful speech felicity conditions (Langton, 2018a). Hence challenging the speaker's epistemic and practical authority, or drawing attention on presuppositions covertly creeping into a conversation in order to undermine their accommodation appear viable ways of counterspeech. Though, counterspeech has shortcomings[2] and in particular the strategy of drawing attention to presuppositions.

3. Lepoutre's alternative approach: *ex ante* positive counterspeech

In a recent paper, Maxime Lepoutre addresses precisely this issue (Lepoutre, 2019). Explicitly focusing on speech spreading misinformation, Lepoutre outlines two kinds of it: speech diffusing falsehoods about policies (a case in point: "vaccine causes autism" rumors) and about persons, based on their group membership. The latter is a form of hate speech, delivering a distorted representation of out-group members.[3]

Lepoutre is concerned with the question of how to draw speech strategies which can effectively counter misinformative speech. Even if they cause different kinds of damages, the two prongs of "ignorant speech" (Lepoutre, 2019, p. 155) share a crucial feature: they can profoundly affect the beliefs of the audience, making it particularly difficult to eradicate the false beliefs – and thus to counter the related harms.

In order to account for this phenomenon, Lepoutre draws on the works of McGowan and Robert Simpson, among others. As Simpson (2013) explains by expanding McGowan's theory of presupposition accommodation, introducing an association (say, associating out-group members with dangerous attitudes or demeaning attributes) in a conversation or in a public debate makes such association *salient* in that context: this means, that the attention of the audience is drawn to such an association, increasing the possibility that someone would believe it.[4] This is likely to be the case especially if the association corroborates already preexistent bias or if it prospects a potential danger. Research on stereotyping and cognitive bias confirm this mainly unconscious process (Blum 2004). Associations are sticky: they are difficult to reverse. Hateful utterances and negative information are especially sticky (Simpson 2013). This explains why counterspeech attempting to challenge such associative claims can possibly backfire: since negating an association requires naming it, every counter-utterance

2 For a critical assessment of counterspeech flaws, see Mc Gowan 2018; Langton 2018a, pp. 16-18.

3 Note that Lepoutre does not focuses on hate speech, but on speech spreading falsehood, though maintaining that falsehood about groups can be crucial in hate speech. I endorse Lepoutre's view that discriminatory speech and hate speech are often corroborated by bias and falsehood. I should add that the epistemic component indeed plays a huge role in the debate about harmful speech and counterspeech.

4 Simpson, as well as Lepoutre, appear to use the adjectives *salient* and *relevant* interchangeably, endorsing an understanding of salience which is commonly used in psychology: salient is something that stands out to our perceptive and cognitive attention. For present purposes, I adopt this understanding of salience, for reasons that should become clear in section 4 in relation to Facebook's definition of relevance. McGowan uses salience in a more narrow and technical way, meaning picking up one particular noun or pronoun from a broader category within a conversation (see also Simpson 2013, footnote 17).

can make the association even more salient. "X are not lazy parasites" (Lepoutre 2019, p. 161) can be an instance.

Hence, as McGowan notes (2018, p. 191), in the attempt of designing an effective counterspeech strategy, we face an uncomfortable challenge: on the one hand, not speaking up could possibly back speech-related harms, favoring felicity conditions such as authority by default and presupposition accommodation. On the other hand, by replying to harmful speech one risks to strengthen its impact.

At this point, Lepoutre introduces his own suggestion for a counterspeech approach addressing falsehood which can avoid the "stickiness" impasse. Firstly, instead of negating false information and associations, counterspeech should be of a positive type, conveying facts and aspirational values. Secondly, Lepoutre recommends diachronic counterspeech: instead of reacting *ex post* to speech already uttered - which must deal with the impossibility of making at least some perlocutionary harmful effects retroactively undone - the diachronic approach aims at building an epistemic and moral common ground which is hostile to political and hateful misinformation. This should be achieved "in at least two ways: by entrenching important facts in the conversation's common ground, and by eroding ignorance-promoting speakers' status, so that they no longer count as authorities" (Lepoutre 2019, p. 182). Once such a background is created, misinformative utterances and hateful associations should hardly be possible, or, more accurately, they would not find the felicity conditions that can make them harmful, such as authority and the covert accommodation of harmful presuppositions. Lepoutre illustrates this with the example of a democratic elected head of state explicitly affirming that the government categorically rejects racism: "The hate speaker who *then* pronounces deeply racist views thereby marks himself as a minority voice, who cannot speak for the majority" (Lepoutre 2019, p. 182).

Some critical remarks are in order. Lepoutre appears to propose a form of citizens inoculation, which should be carried out by the state, against hateful ideologies. I do not dispute the legitimacy and possibly also the necessity of such enterprises. Yet when one considers that similar efforts were made in European Countries after the Holocaust, especially and very systematically in Germany, and one looks at the current rise of far right-wing discriminatory ideologies, including in that Country, then some doubt arises as to the effectiveness of the inoculation approach.

It should be noted that on the contemporary political scene worldwide racism, as well as right-wing extremism, cannot be properly defined as marginal phenomena. Currently, it appears not easy to reverse perceived authority, or factual acceptance, of racist and discriminatory speech in the way Lepoutre suggests. Possibly, it is already time to react, rather than to prevent. Moreover, even if inoculation strategies can do their part – and in some specific historical and political contexts, a huge part – in making the terrain unsuitable for speech to be harmful, the concept of counterspeech explicitly refers to a response, and thus to an *ex post* reaction to speech. Thus, Lepoutre partly bypasses the question of how to design viable counterspeech strategies.

Undoubtedly, since digital communication has been expanding as one of the most powerful means for conveying speech, research on counterspeech is presently urged to take online speech specificities into account. Since the philosophical positions outlined above focus on traditional forms of speech and counterspeech offline, the question arises as whether the tools they provide can be capable of matching the specific challenges of digital communication. This is what I would now like to explore.

4. Digital conversational dynamics

The recent proliferation of ideologies of discrimination and of far-right groups and political parties is accompanied by the rise of social media, which in many relevant respects differ from one-way communication. Importantly, social media are designed for interaction, in part following traditional non-digital conversational patterns, in part radically diverging from them. Physical unavailability, anonymity, spatial reach and temporal pervasiveness are just the most evident features of digital communication. Moreover, the peculiar interaction patterns of social media, made of an interplay of posts, re-posts, likes and comments require a rethinking of the relationship between speech and salience, confirmation bias and authority.

Let start by considering the problem of speaker's authority. As discussed above, authority can be granted by default, i. e. by the audience's failure to question the presupposed speaker's authority. The examples provided by Maitra, Langton and McGowan depict almost exclusively situations of small-scale personal interaction, mainly in public spaces: that of a man targeting an Arab woman with racist speech in a crowded subway car (Maitra 2009, p. 19) being a case in point.[5]

Now consider a racist post by an unknown private user on a public online platform.[6] Is the "authority by default" model, transposed in the context of digital communication, still convincing? I think the answer is no. Other users' failure to react to the racist post would be unlikely to have the effect of legitimizing it. Maintaining that the post can still have harmful perlocutionary effects on readers, both targeted and not, if nobody takes it up it can hardly exercise normative force, nor can its author be granted epistemic and practical authority in that digital conversation.

This does not mean that in digital communication other users cannot grant authority to a speaker, to a speech act or to a particular conversational tone. On the contrary, in the digital context speech impact - and the phenomenon of granting authority in particular - appear to be dependent on what other users do in a fundamental way. Though, it is not other users' omissions, but rather their active participation, which can substantially determine processes of legitimation and normalization; active participation meaning answering, sharing, rating or linking other users' contributions.

Importantly, the structure of harmful speech on social media is not, or by far not only, that of an authoritative speaker and an audience, according to a "top down" model of communication. More aptly, it can be described as a choral process of mutual confirmation and validation. Political leaders who do not endorse social equality values have quickly learned how to use digital communication participative patterns to strengthen their legitimacy.[7] A look at the online communication of right-wing political leaders reveals accurately calculated propaganda strategies: carefully avoiding explicit racist speech, which would probably backfire, though sometimes intentionally slipping into it, winking to supremacist or fascist followers, far-right leaders leave, more often than not, the "dirty job" of the most explicit hateful speech to their followers.

Beyond the use made by political leaders and on a more general level, other users' interaction by replying, linking and liking is necessary for posts and comments to get attention, fundamentally determining digital processes amplifying the impact of that speech. Such processes follow partly social patterns of communication and interactional dynamics

5 Langton briefly touches upon digital communication, for instance in 2018a, p. 2, though not elaborating on it.

6 I focus here on public communication platforms, since they play a crucial role in shaping public discourse about online harmful speech. In small private digital conversations among participants knowing each other, things may look different.

7 Thanks to an anonymous reviewer for bringing the work of Jennifer Saul (2017) and Tail Mandelberg (2001) to my attention.

characterizing offline speech: the attention, and the interest, of an audience is most likely to be attracted by speech already getting a considerable uptake.
But importantly, digital communication is also driven by automatic digital processes, which do not depend from patterns of human interaction. Besides human users, digital conversations are crucially shaped by non-human "participants", namely algorithms, which are responsible (in a technical sense) of some decisive, routinely activated online conversational dynamics. Facebook, for instance, automatically ranks posts on public pages according to the number of interactions. Closely connected with Facebook's dynamic ranking process is the visibility of posts and comments: they are shifted between top and bottom of the comment thread, which has a substantial impact on how many users will see and read them. In this way, contributions to a digital conversation obtain what Facebook calls *relevance*.
In light of this, we can reconsider McGowan's thoughtful insights on conversational kinematic, specifically on presupposition accommodation dynamics. It should be noted that McGowan is explicitly concerned with small-scale conversations occurring face to face. As explained above, her theory has already been expanded by others, notably Simpson and Lepoutre, in order to cover broader speech contexts and with a particular focus on the phenomenon of the "stickiness" of presuppositions and expectations introduced in a conversation. Now, it could be useful to consider the implications of the abovementioned digital speech dynamics for McGowan's theory and especially the role played by algorithms in determining speech relevance. We can hardly maintain the view that every contribution to a public platform digital conversation automatically changes the score, i. e. what counts as appropriate in that conversation, until someone tries to block this accommodation process. On the contrary, in absence of other users' active interaction, digital conversational contributions do not have much chance of affecting the score of that conversation.
Or more carefully, we could indeed maintain McGowan's view, if we integrate it with the consideration of the cumulative effect of interactions on further development of the conversation. Not the single contribution, but the aggregative effect of more interactions with that contribution can change the permissibility rules of the public digital conversation. In this process, human active interactions (rather than omissions) and algorithmic dynamics are closely interconnected. They determine which content stands out to attract users' attention and thus its salience (using Facebook's terminology: its relevance) which is crucially dictated by social media platform standards.
Beyond the level of the changes produced in a particular conversational score, digital speech relevance can have a huge impact on users' beliefs and behavior. The concept of norm perception, which has been developed in psychology, can be useful for understanding the relationship between online ranking, beliefs, values and behavioral choices. While individuals tend to perceive as norms the standards of persons and groups with which they identify (Tanckard & Paluck, 2015), they can also be conditioned by the value judgement of the majority, as Genocide Studies show: "sufficiently suffuse an ideological environment, so that a belief becomes something like 'everybody says' and that is liable to receive wide endorsement even if never properly substantiated" (Maynard 2014, pp. 10-11).
Though empirical work on this topic is surely needed, the impact of social media ranking on the perceptions and decisions of users is already well known - specifically, in the form of the relationship between accounts ranking and influencer economy. Further research on the phenomenon of online influencers' epistemic and practical power could surely shed light on the topic of speaker's authority in the digital domain. Needless to say, digital conversational dynamics are fundamentally driven by and exploited for economic interests.
However, the phenomenon of ranking can be turned to the advantage of counterspeech efforts as well. Activist counterspeech groups around the world are already practicing such approach,

showing a fresh way of engaging in effective action beyond the dichotomy of *ex ante* positive speech and *ex post* critical counterspeech.
The Swedish group #jagährär[8] have implemented an organized system of specifically online-tailored, collective counterspeech. #jagährär is not bound to a political party, and is framed by a set of rules, based both on democratic values like freedom of speech and diversity of opinion and on considerations of effectiveness. In order to avoid the drawback of making the posts they want to counter more relevant through their interactions (the problem MCGowan, Simpsons and Lepoutre are concerned with, though relating to offline speech), #jagährär activists rather focus on interaction with positive comments and counterspeech posts, in order to shift them higher in the comment thread. Researcher Susan Benesch explains: "Most of the news outlets have their comments ranked by what Facebook calls 'relevance'. Relevance is, in part, determined by how much interaction (likes and replies) a comment receives. Liking the counterspeech posts, therefore, drives them up in relevance ranking, moving them to the top and ideally drowning out the hateful comments." (Benesch 2019).[9]

5. Counterspeech as a positive duty to rescue: Howard's proposal

While the project of organizing counterspeech groups of private citizens appears praiseworthy, a related question emerges, which, as Jeffrey Howard rightly recognizes, has been unjustly neglected in the literature (Howard, 2018). Do individuals have a moral duty to engage in counterspeech, and if they do, which is the moral source of this duty?[10]
Touching upon Brettschneider's important work on state counterspeech (2012), Howard argues that individual action and state engagement in speaking back are not mutually exclusive, rather, they are complementary. Focusing on the question of how to justify the moral duty to counterspeak of private citizens, Howards concentrates on a specific kind of harmful speech, namely "speech that implicitly or explicitly encourages wrongful criminal violence, such as speech that advocates terrorism and incites racial hatred" (Howard, 2018, p. 2).
According to Howard, the normative source of the duty to counterspeak against incitement to violence is "the Samaritan obligation to rescue others from risks of harm" (Howard 2018, p. 1). Howard considers the Samaritan obligation a natural duty, pertaining to every moral agent, under the condition that the obligation is not too demanding (Howard 2018, p. 6). Since the counterspeech duty requires just to speak, and not to engage in dangerous rescue operations, the obligation is, *prima facie*, not too demanding.
Despite valuable insights on the counterspeech topic, Howard's route appears unconvincing in many respects, both on a general level and specifically when related to digital speech.
Firstly, Howard postulates the Samaritan duty to rescue others from risks of harm as universally accepted: "The general idea that we have natural duties to defuse unjustified threats posed by others is not controversial" (Howard 2018, p. 6). However, this issue is by far more controversial, both in philosophy and law, than Howard admits.[11]
The duty to rescue is *prima facie* imposed to private citizens under further specific conditions, in addition to the requirement that the intervention is not too demanding. There are various approaches to this. Generally, the duty relates to life-or-death-cases, in which the rescuer,

8 https://www.jagarhar.se/. Presently, #jagährär groups are present in 12 different countries. See also Cathy Buerger 2020.
9 On using "like" interactions as a strategy for countering ISIS propaganda, see Yadron 2016.
10 Though in philosophy the concepts of moral duty and moral obligation are sometimes differentiated, for present scope they are used interchangeably.
11 See for instance Schiff 2005, Dressler 2000.

standing close to the perilous event, is capable of effectively intervening. The example of a good swimmer's duty to rescue a drowning person is a case in point.
Incitement cases are structured differently, involving a speaker (the inciter), the possible perpetrator of violence and a bystander (the possible bearer of the duty to counterspeak). In Howard's example (2018, p. 6), an inciter is intentionally striving, with high chances of succeed, to induce someone to kill. In such clear structured cases, there can well be evidence of serious peril, and trying to dissuade the speaker or the perpetrator can be plausibly framed as a way to fulfill the moral duty to rescue; However, I would argue, in this case the duty also entails the moral (and possibly, legal) obligation to report to authorities and under certain conditions even to physically impede the violence, if the counterspeech strategy does not prove effective.
So, if Howard has just this type of cases in mind, then his argument of extending the duty to rescue to cases of incitement to violence can be plausible. But this does not seem to be the case. Indeed, Howards is also concerned with "speech that risks inspiring agents to engage in criminal violence" (Howard 2028, p. 5), which makes the analogy with life-or-death rescue cases less convincing. The category of speech that *risks* inspiring to criminal violence is very broad, surely including other kinds of speech rather than just incitement (whose legal definitions, incidentally, usually include the *purpose* to induce someone to violence).
It is extremely difficult to establish if and in which way a specific utterance would lead the audience to violence, since the effects of speech are highly context-dependent, hinging on a broad range of variable factors. Moreover, it should be noted that the variety of speech which is apt to inspire to violence is more often than not harmful even in the absence of such an effect: speech unjustly discriminating, subordinating, propagating distorted information such as the depiction of the target as dangerous can lead to grave harms, including the establishment and legitimation of social injustice and rights deprivation. Grounding a duty to counterspeak on the moral obligation to rescue from criminal violence leaves these harms unaddressed.
Thus, the duty-to-rescue route raises the following quandary: either the counterspeech obligation addresses just inciting speech which would bring about criminal violence with a high degree of probability, in which case other kinds of regulations could be better suited. Or the duty has a wider scope, including speech that risks inspiring to physical violence: then we face both the huge empirical difficulty of singling out which speech should be countered, and the problem of leaving other harmful effects unaddressed. Alternatively, we further relax the requirements, abandon the focus on physical harm in order to address other harmful effects and establish a general individual duty to counter potentially harmful speech – which would have troublesome implications, even if only for the vagueness of its scope.
Moreover, Howard does not seem to consider the specificities of digital speech. In the contemporary world, the main (though not exclusive) vehicle of incitement to terrorism and racial hatred is online communication. Words of hatred, refusal and dehumanization, scapegoating and misrepresentation of the other as a threat are spreading from myriads of digital sources, amplified by interactive processes and echo chambers among global users. Incidentally, it should be noted that the practice of linking and liking is not always driven by hateful purposes, nor accompanied by awareness of related harms: a subtle dynamic of reward and gamification-based incentives binds the users to social media participation and interaction. Facebook first president Sean Parker's comments are illuminating, explaining how the "like" button was specifically designed to induce in the users "a little dopamine hit" (Deibert, 2019, p. 30).
However playful the participation to hateful speech online may be: the harms it contributes to are multifarious and steered by specific digital conversational dynamics. Howards suggested

counterspeech strategies - among others, to change speaker's and audience's emotional and moral attitudes "introducing her to a new person, recommending a particular film or simply telling her a provocative story that powerfully conveys an alternative narrative" (Howard 2018, p. 14) - do not appear to be best suited for meeting the specific challenges of digital harmful speech.

6. The negative duty approach: complicity by failure to counterspeak?

Beside the proposal of grounding counterspeech duty in the moral obligation to rescue others from risks of harm, and thus in the positive duty to make others better off, it is worth considering an alternative strategy, drawing on the negative duty not to harm others. In the literature, this route has been repeatedly touched upon, though not systematically elaborated. As Maitra (2009) notes, failure to raise objection can contribute to speech-related harm, grounding the responsibility of a silent bystander as an accomplice in that harm; complicity being a form of responsibility for (reasonably foreseeable) contribution to harm committed by someone else.[12] Of course, in order to make someone accountable as an accomplice, some further conditions must be met, in addition to a possible causal contribution. Defining the appropriate epistemic, intentional and causal requirements for complicity is a challenging task, and the more so in cases of complicity by omissions.

Though, the phenomenon of contribution-by-silence and the concept of bystander complicity reveal important insights relating to moral choice. Notably, in the face of harm a bystander's options for action do not always include a threefold range of possibilities: actively supporting harm, objecting, or remaining neutral. The idea that a bystander's inactivity does not affect the setting, leaving things happening in the very same way, as if the bystander was not there, is, more often than not, an illusion. In case of small-scale offline context, the presence of a bystander can have remarkable causal and normative implications. Bystanders are in a privileged position, witnessing what is going on and having the chance to directly intervene to counter harm, which could make them the natural bearer of the duty to engage in counterspeech in that specific situation. Moreover, other actors being aware of bystander's presence can enact expectations, leading to record the omission of intervention and generating a variety of psychological and social harmful effects.

Though, these insights are not transferable to the context of digital communication without more work. Digital conversations on public platforms indeed seem to concede the possibility of remaining neutral by failing to interact: since what other users - and algorithms - track is active participation, not omission, it seems possible to look away from harmful speech without impacting on it. Thus, a user running into a potentially harmful digital conversation - becoming a virtual bystander in the privileged position of epistemic awareness and intervention capability - can simply decide to leave the platform and navigate into more peaceful digital waters. Nobody would notice, no expectation would arise, nothing seems to qualify this particular user as bearer of counterspeech duty.[13] The analogy between digital users and offline bystanders seems of little use to an account of digital counterspeech obligation based on the negative duty not to contribute to harm.[14]

However, digital harmful speech could be framed from a broader perspective, as a large-scale process systematically improving the erosion of aspirational values such as social justice and human rights. Refraining from dealing with this, shirking one's own political responsibility,

12 This definition is widely shared in law and philosophy, diverging both from Brown's (2019) and from definitions requiring participatory intention.

13 Note that the name of the Swedish digital counterspeech group #jagährär means "I am here".

14 See also Brown 2019, p. 217.

means leaving the development of unjust social patterns and institutional orders in the hands of others – which can still be a way of reinforcing it.
Perhaps, framing the question of why we should engage in counterspeech in terms of a moral duty, attempting to ground this duty in a single general moral obligation – either to benefit, or not to harm others – is neither adequate, nor necessary. We indeed have a variety of good moral and political reasons for taking an active part in countering both small-scale and systemic harm, and speech can be a powerful means at our disposal. This implies becoming aware of how counterspeech can be effective, and of possible shortcomings. Learning new ways in which digital technology can promote harm – but also how it can be effectively used against it – are mandatory first steps.

REFERENCES

Ayala, S. & Vasilyeva, N. (2016) Responsibility for Silence. *Journal of Social Philosophy* 47(3), 256-272;
Benesch S. & Jones D. (2019, August 13) Combating Hate Speech through Counterspeech. Retrieved from: https://cyber.harvard.edu/story/2019-08/combating-hate-speech-through-counterspeech;
Blum, L. (2004) Stereotypes and Stereotyping: A Moral Analysis. *Philosophical Papers* 33(3), 251-289;
Brettschneider, C. (2012) *When the State Speaks, What Should I Say?* Princeton: Princeton University Press;
Brown, A. (2019) *The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech*, in: S. Brison, K. Gelber, Free Speech in the Digital Age. Oxford: Oxford University Press;
Buerger, C. (2020, December 14) The Anti-Hate Brigade. Retrived from: https://dangerousspeech.org/anti-hate-brigade/;
Citron, D. & Richards N. (2018). Four Principles of Digital Expression (you won't believe #3!). *Washington University Law Review,* 95, 1353-1387;
Deibert, R. (2019) The Road to Digital Unfreedom: Three Painful Truths About Social Media. *Journal of Democracy*, 30(1), 25-39;
Dressler, J. (2000) Some Brief Thoughts (Mostly Negative) About "Bad Samaritan" Laws. *Santa Clara Law Review* 40, 971-989;
Howard, J. (2019) Terror, Hate and the Demands of Counter-Speech. *British Journal of Political Science* 1-6. doi:10.1017/S000712341900053X;
Klonick, K. (2018) The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, 1598-1670;
Langton, R. (1993), Speech Acts and Unspeakable Acts. *Philosophy and Public Affairs*, 22(4), 292-330;
Langton, R. (2018a), Blocking as Counter-speech, in: *New York on Speech Acts*, Oxford University Press;
Langton, R. (2018b), *The Authority of Hate Speech*, in: J. Gardner, L. Green, B. Leiter (eds.), Oxford Studies in Philosophy of Law, vol. 3, Oxford: Oxford University press 123-152;
Lepoutre, M. (2019). Can "More Speech" Counter Ignorant Speech? *Journal of Ethics and Social Philosophy*, 16(3), 155-191;
Maitra, I. (2012) *Subordinating Speech*, in: I. Maitra, M.K. McGowan (eds.), Speech and Harm: Controversies Over Free Speech. Oxford: Oxford University Press;
McGowan, M.K. (2009) Oppressive Speech. *Australasian Journal of Philosophy* 87(3), 398-407;
McGowan, M.K. (2018) *Responding to Harmful Speech. The More Speech Response, Counter Speech, and the Complexity of Language Use*, in: Johnson, C.R. (ed.), Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public. London, New York: Routledge;

McGowan, M.K. (2019) *Just Words: On Speech and Hidden Harm*. Oxford: Oxford University Press, 2019;
Schiff, D. (2005) Samaritans: Good, Bad and Ugly: A Comparative Law Analysis. *Roger Williams University Law Review,* 11(1) 77-141;
Simpson, R. (2013) Un-Ringing the Bell: McGowan on Oppressive Speech and the Asymmetric Pliability of Conversations. *Australasian Journal of Philosophy* 91(3), 555-575;
Tankard, M., Levy Paluck E. (2015) Norm Perception as a Vehicle for Social Change. *Social Issues and Policy Review,* 10(1) 181-211;
Yadron, D. (2016, Januar 21) Facebook's Sheryl Sandberg: "likes" can help stop Isis recruiters. The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/jan/20/facebook-davos-isis-sheryl-sandberg.