Chapter 8
Someone Is Wrong on the Internet: Is There an Obligation to Correct False and Oppressive Speech on Social Media?

Jennifer Saul

Introduction

The experience is not an uncommon one: over your morning coffee, you open up Facebook and find a friend of yours discussing someone else's false and offensive post and urging all right-minded people to go and comment on it. Alternatively, perhaps you read a false and offensive tweet from an acquaintance from high school, or even from someone close to you. You feel appalled and you feel a pressing obligation to say something, not to just let it sit.

Perhaps at this point you pause for a moment wondering what to do. And here the story gets a little more fanciful. If you are a philosopher (or interested in philosophy), you might turn to what philosophers have said about responding to false and hateful speech. You will find some arguments that may make you feel an urgent need to respond, reinforcing what your friends are saying, or what you are already feeling. For example, you will see arguments from Rae Langton on the importance of what she calls 'blocking' oppressive speech, which include powerful statements like this: 'Hearers and bystanders who do not block will sometimes, through that omission, make a speech act more evil, whether they mean to or no.' (Langton 2018: 161) You are likely to be deeply concerned by Ishani Maitra's (2012) argument that not objecting can confer authority on an utterer of hateful speech. Sandy Goldberg

I am very grateful for discussions of earlier drafts of the paper with audiences at the University of Sheffield and the University of Waterloo; for conversations with Patrick Connolly, Shannon Dea, Ray Drainville, Anna Klieber, Mathieu Marion, and Martina Rosola; and for research assistance from Anna Klieber. I'm grateful also for helpful feedback from the editors.

J. Saul (✉)
Department of Philosophy, University of Waterloo, Waterloo, ON, Canada
e-mail: Jennifer.saul@uwaterloo.ca

A. MacKenzie et al. (eds.), *The Epistemology of Deceit in a Postdigital Era*, Postdigital Science and Education,
https://doi.org/10.1007/978-3-030-72154-1_8

(2020a) will tell you that there is a default entitlement to take silence as acceptance, and you will worry about what it means if you do not speak up. You will also find Casey Rebecca Johnson (2018) arguing that there is an epistemic obligation to voice disagreement. All this may make you feel, all the more pressingly, that you need to throw yourself into those threads and speak your mind.

And yet, there is also another strand of thought. These authors acknowledge some serious problems that one may encounter in trying to speak up. And many others also discuss these problems, some giving them even more weight (Lepoutre 2019; McGowan 2012, 2018). Counterspeech may fail or even backfire. It may not be safe to speak up in a particular context, or for members of a particular group. Indeed, it may be that the structural injustices present in society make any claim of a duty to speak up a piece of ideal theory (Lackey 2018).

At this point, you might feel very uncertain what to do. You might think, perhaps with some reason, that philosophers may not have been the best people to consult on this matter. But one interesting fact is that these discussions have been built almost exclusively around a model of speech acts taking place in face-to-face conversations. In this paper my plan is to begin from the work of philosophers on counterspeech in general, but then turn to the particular issues posed by social media. I will discuss the special problems presented by social media, but also some interesting approaches to false and offensive speech that social media makes possible. In the end, I will offer some advice to the fictional you who is appalled by what you see on social media, but (as usual for philosophy papers) it will not be as detailed and concrete as you might like.

Why You Might Feel You Should Speak Up

Speech Act Strand

There is a substantial speech act tradition in social and political philosophy, which forcefully presses the point that speech acts can dramatically alter what is permissible, making racism and sexism more acceptable. The focus here is particularly on *unchallenged* speech acts, which can have the effects that they do precisely because they are not challenged. It is not hard to see how this focus could make one feel the pressing necessity of challenging. (Importantly, as we will see, these authors do acknowledge and grapple with difficulties for such challenges.)

Mary Kate McGowan (2012, 2018) is focused on the ease with which facts about what is permissible can be changed, especially with regard to racism and sexism. She argues that the ease with which permissibility facts change more generally leads to an overly smooth facilitation of racism and sexism. Any utterance, for McGowan, changes some permissibility facts simply by virtue of changing the state of the conversation. If I say, ‘hang on, my kitten is chewing the power cord’, it becomes permissible to assume that I have a kitten and impermissible to ask if I

have any pets—as long as nobody follows up with, 'hey, you don't have kittens, what are you talking about?'. That much is unremarkable. But McGowan argues that the same sort of thing happens with racist utterances. If someone in a conversation makes a racist utterance and nobody objects, then racism (or at least racist utterances) becomes permissible in that conversation. For this reason, objecting becomes paramount. There are very high costs allowing a racist utterance to go unchallenged, because not challenging it allows the permissibility facts to change and make racism permissible for the conversational context and even beyond. McGowan extends this approach to oppressive speech and oppression more generally, building a compelling picture of how unchallenged speech acts can have serious oppressive effects. As we will see, however, McGowan does also raise concerns about the difficulty of such challenges.

Ishani Maitra (2012) directs our attention to the way that speakers may come to be vested with authority simply by other speakers not questioning linguistic moves that are made. So, for example, consider a case in which someone is racially abusive on a subway car, and nobody challenges them. The speaker, in her example, aims to rank his target as inferior, and this is licensed by the silence of the other passengers, who thereby give him the authority for this ranking (Maitra 2012: 115). It is crucial to Maitra's picture that this happens even if the other passengers are quite uncomfortable with what is taking place: 'Even if the hate speaker's fellow passengers have strong reservations about what he says to the woman he targets, as long as they fail to speak up, the speaker can end up with authority.' (Maitra 2012: 116)

Rae Langton's (2018) recent work has been on 'back door speech acts', which smuggle in assumptions in ways that may go unnoticed by participants. Presuppositions are a key example of an ordinary way that this happens—if I say, 'he is from Michigan but doesn't like Trump', my interlocutors may take on the assumption that a Michigander who dislikes Trump somehow defies expectations. And they may do this without even realizing that this is what they are doing. This can be easily stopped, though, by pulling out the assumption and criticizing it: 'Hey! Actually more than half of Michigan voters were opposed to Trump in 2016, and he only won the state due to third-party voting. And Biden won in 2020!'

More perniciously, racist and sexist background assumptions can be smuggled in this way. Here is an example from John McCain. Remarkably, this interchange has been repeatedly cited as an instance of John McCain's brave *anti*-racism.

> 'I gotta ask you a question,' Quinnell told McCain, who leaned in closer to hear her.
>
> 'I can't trust Obama,' she told McCain, and the world. 'I have read about him and he's not… he's not… he's an Arab. And…'
>
> This is when McCain politely took back the microphone and started shaking his head back and forth. He did so instinctively, without a hint of political motivation or strategic forethought.
>
> 'No?' Quinnell asked, her voice trailing off.
>
> 'No, ma'am,' McCain replied decisively.
>
> *'He's a decent family man, citizen, that I just happen to have disagreements with on fundamental issues,' McCain told the crowd.* (Davich 2018) (emphasis added)

McCain's utterance in italics presupposes that there is a contrast between being Arab and being a decent family man and citizen—it does not make sense without it. Yet this is hidden enough that most commentators (and quite likely McCain himself) failed to notice this. It is precisely this sort of thing that Langton is concerned with (although the example is mine). The right thing to do, she says, is to pull out these sorts of assumptions and criticize them. To do so is to engage in the speech act of *blocking*, an important counterspeech obligation. Indeed, she writes, '[h]earers and bystanders who do not block will sometimes, through that omission, make a speech act more evil, whether they mean to or no' (Langton 2018: 161). Although Langton very much acknowledges that blocking is not always possible, she builds an appealing picture of its power, describing it as offering 'a modest time machine' because it offers a way to undo what may have seemed successful pernicious speech acts.[1]

Silence as Assent Strand

There is another strand of argument that might make you feel even more uncomfortable about not speaking up. This is Sandy Goldberg's (2020a) argument that not speaking up will (generally) be interpreted as assent or at least acceptance.[2] The idea that we could be seen as accepting false and oppressive claims is one that motivates many people to feel a pressing urge to comment quickly or retweet angrily.

Goldberg (2020a) argues that if one is engaged in a cooperative conversation (in the sense of Grice 1991), then—if one does not speak up—one will generally be assumed to agree with what a speaker has said. This is because there is an assumption that if one rejects an assertion, one will say so. More precisely, he endorses the claim that he calls NO SILENT REJECTION (the capitals are his):

> In all speech exchanges which are Gricean conversations, all competent language users enjoy a default (albeit defeasible) entitlement to expect that an audience regarding whom it was manifest that he has remained silent in the face of a publicly made assertion has not rejected that assertion. (Goldberg 2020a: 176)[3]

If A says something false (or otherwise rejection-worthy) to B, then—if B is being cooperative and is aware of the falsehood—B should speak up. If B doesn't speak up, then there is a default entitlement to assume that B agrees. This is only a

[1] For a very close examination of undoing speech acts, see Caponetto (2020).

[2] Philip Pettit (2002) also argues for this conclusion, more briefly. His argument also turns on Gricean considerations (Grice 1991). According to Pettit, in conditions of genuine freedom of speech, silence can (generally) legitimately be presumed to communicate assent. This could be seen as motivating a very strong obligation to speak up on social media. Here I focus on Goldberg, due to his much more detailed discussion of possible defeating conditions for the presumption of acceptance. (For a response to Pettit 2002, see Langton 2007.)

[3] For an argument against Goldberg (2020a), see Klieber (2020).

default, not a guarantee: there may be reasons for not speaking up. But this default means that silence will, in general, be taken as assent. And this is what generates the obligation to set the record straight by speaking up. This duty holds, Goldberg argues, for both morally and factually problematic utterances. Indeed, one form of support that Goldberg offers for this is a collection of powerful statements about the obligation to speak up in the face of wrongdoing (not just wrong speaking). For example:

> When I was the rabbi of the Jewish community in Berlin under the Hitler regime, I learned many things. The most important thing that I learned under those tragic circumstances was that bigotry and hatred are not the most urgent problem. The most urgent, the most disgraceful, the most shameful and the most tragic problem is silence. (Joachim Prinz, quoted in Goldberg 2020a: 170)

Goldberg (2020a) acknowledges, as we will see, that this is merely a default—and that there will be many circumstances in which this obligation does not hold. But the mere fact (if it is one) that there is a default of interpreting silence as acceptance will contribute substantially to the felt need to *respond now* to problematic claims on social media.

Epistemic Obligation Strand

Rebecca Casey Johnson (2018) argues that there is a distinctly epistemic obligation to voice disagreements. She identifies several sources for this obligation, focusing in particular on self-regarding duties, obligations arising from cooperative endeavors or from efforts at enquiry, and obligations stemming from Millian concerns about the need to submit beliefs to proper scrutiny. Again, this obligation is merely a default. But it is the sort of default that could make you think you do need to put down your coffee and respond to that false and offensive tweet in the story that I mentioned at the start of this chapter.

Problems with Counterspeech

There is also a very substantial literature, on the other hand, itemizing serious problems with counterspeech.

Jennifer Lackey (2018) gives one of the most powerful statements of this in her objection to Goldberg (2020a). She argues that his NO SILENT REJECTION claim is motivated by ideal theory—by a picture of the world in which people are situated as equals—and that this claim only makes sense in such a world. However, in our actual world, power dynamics are omnipresent, and huge numbers of people are unable to object safely or to be taken seriously:

> No conversation is entirely free of differences in the distribution of epistemic goods, status, power, psychology, cultural expectations, practical constraints, or some combination thereof. Speaking up against others almost always involves a calculation—whether conscious or not—that is based on one's position and the costs and benefits of dissent on this topic at this time with this conversational participant. (Lackey 2018: 90)

To some extent, Goldberg's (2020a) view *does* acknowledge these varying power dynamics. He discusses circumstances in which NO SILENT REJECTION fails, and those circumstances include those in which objecting is very costly and those in which objecting is ineffective. But Goldberg treats these as particular circumstances which may or may not arise for individuals, noting that when they do, NO SILENT REJECTION will be undermined. Lackey argues that the corrosive effects of power dynamics are so widespread and systematic that there cannot be a *general* NO SILENT REJECTION, which she rightly notes would underpin a very general duty to reject. Instead, she suggests that there are different sorts of duties for members of different sorts of social groups (as well as the more contextual variations that Goldberg 2020a discusses). Importantly, she notes also that members of some social groups will have an *easier* time speaking up and being taken seriously—and that this gives them a heightened duty to object.

Mary Kate McGowan (2012, 2018), Robert Simpson (2013), and Maxime Lepoutre (2019) raise serious concerns about the idea that counterspeech can be effective or safe. These are rooted in the same concerns that Lackey has about the power dynamics of society. McGowan notes that speaking up—which can be dangerous—constitutes a serious extra burden for those targeted by hate speech, who are already suffering the harms of that speech (as well as of oppression more generally). She also notes that power dynamics may render counterspeech ineffectual: a woman who speaks up in a sexist environment, for example, is not all that likely to have her objections properly understood, taken seriously, and acted upon in an appropriate manner. She further notes that there is an important asymmetry: while it is relatively easy to carry out a speech act of oppression, it can be much harder to reverse it. Finally, building on work from Simpson (2013), she notes that counterspeech can backfire—a point that will be important later in this paper. Her concern, and Simpson's, is that objecting to something may serve to reinforce associations: repeatedly explaining that Black men are not disposed to violence may, for example, unwittingly reinforce the association between Black men and violence.[4] Lepoutre (2019) extends this point beyond hateful and oppressive speech, noting that the association reinforcement can happen with any attempt to correct

[4] McGowan (2018) suggests that these difficulties are serious, but sometimes possible to overcome. A skillful interlocutor can, in some circumstances, succeed in shifting the focus of a conversation. McGowan's example involves someone very politely and skillfully making sure that a colleague of color is considered for a leadership role (McGowan 2018: 192). As McGowan acknowledges, this will often not be possible. And as anyone who has tried to have a tactful conversation on the Internet knows, special challenges for this technique are presented by the dynamics of social media.

falsehoods, including reinforcing the association between vaccines and autism through one's speech denying this link.[5]

Importantly, the authors we have discussed are all sensitive to these concerns. Some of their arguments, as we have seen, can be taken to provide strong motivation for counterspeech. But they also show real concerns about how often this will be effective or safe, especially for members of marginalized groups. Rae Langton (2018) discusses the way that social location can impact one's efforts to object. Her concern is that oppressive power and social norms can render counterspeech less effective. The oppressive power of racism can mean that a Black person's utterance is not listened to or taken seriously. Norms that prescribe passivity and agreeableness for women may mean that when they object, they are dismissed as difficult and not worth paying attention to. Considering both of these, it may be far less effective when members of subordinated groups speak up. Both Langton and McGowan suggest that, due to these issues, members of dominant groups have a greater obligation to speak up. (Though, as they are aware, this does not solve all of the problems that they raise.)

Goldberg (2020a) and Johnson (2018) also acknowledge key factors which can affect an obligation to speak up. Goldberg notes that the default entitlement to think that a silent audience agrees does not hold unless one is in a cooperative conversation. So if the situation is either not cooperative or not a conversation, this entitlement is not present and there is not the same obligation to speak up. This will include, for example, situations in which a back and forth is not expected (a formal meeting with a hierarchy, perhaps), or situations in which one has opted out of cooperating (perhaps in protest). The entitlement is also removed if there is an outweighing explanation for the silence—for example, if the room is too noisy, or one has laryngitis. Or, importantly if there are difficult power dynamics—one might not be able to tell one's boss that they are saying something false. The cost of objecting might be too great—especially in cases like hate speech, where objecting could lead to violence. Johnson similarly notes that considerations of practicality, safety, and power dynamics can override the default obligation.

None of these authors specifically addresses the issues that arise with objections and counterspeech on social media. I'll turn to those now.

[5] Lepoutre (2019) invokes, in addition to other arguments, the much-discussed 'backfire effect' (Nyhan and Reifler 2010), a purported psychological phenomenon in which correcting a falsehood seems to lead to increased belief in the falsehood. It is not so clear, however, that this effect actually exists (see Swire-Thompson et al. 2020). If this effect is proven to exist, it only strengthens the arguments for difficulties of counterspeech.

Problems with Objecting on Social Media

So far, we have seen that there are arguments which seem to back up that pressing felt need to speak up in the face of false or oppressive speech. But we have also seen ways that it may be mitigated, especially if we are members of oppressed groups who may find objecting unsafe—or for whom it may also be less effective. But, this line of thought goes, unless there are such power dynamics present it really would be much better to speak up. However, these arguments come from authors focusing on our duties in face-to-face conversations. It is important, then, to think carefully about what happens when we turn our attention to speech on social media.

One reason for hesitating over the felt obligation to object on social media, which we will not dwell on much, is that false and offensive claims are *constantly* being made on social media. Trying to object to every claim would be a never-ending, exhausting, and fruitless task. But we'll set this aside—the demandingness of duties is a well-worn topic and not something specific to this one.[6] Instead, we will look at other features of online communication which set it importantly apart from face-to-face communication.

Even before the pandemic, social media speech had become enormously influential. But with a dearth of face-to-face human contact taking place, it takes on even more importance. The effects in the world are by now beyond doubt. Moreover, there are both a vast number of falsehoods circulating on social media and a torrent of opportunities to object to them, as well as considerable pressure to do so. Since conversational dynamics on social media are very different from those face to face, it is important to carefully consider their consequences for the duty to object. I will argue that these consequences are profound.

First, however, let's think a bit about some of the ways that social media conversations differ from face-to-face ones.[7] Here are some key ones that will be important to our discussion.

- Uncertain/changeable audience: In a face-to-face conversation, one usually knows to whom one is speaking. Even in the case of addressing a crowd, one knows that one is addressing a crowd. While the audience may sometimes change—one person leaves the table at the bar; another sits down—this is not a constant feature, and the speaker is reasonably likely to be aware of it. This is not at all the case on social media.
- Responses can drastically alter the makeup of a conversation. If someone with a large social media following—or just one very different from one's own—weighs in, the audience can be dramatically altered. This is unlikely to happen face to face.
- Audiences can be indefinitely vast—there is virtually no limit on the number of people who may become participants in, or audiences for, a social media

[6] Johnson (2018) has a nice discussion of this sort of limitation on the duty to object.

[7] Here I draw on work by Connolly (2020) and Goldberg (2020b).

conversation. Physical constraints alone mean that face-to-face conversation is not like this.

Amplification

As mentioned above, the philosophers who have pressed the importance of objecting have been very much focused on face-to-face interaction. In face-to-face conversations, a lot of awkward, unfortunate, or even terrible things can happen when one objects. One may be ignored; one may be made to feel that one was being rude; one may be attacked, verbally or physically. And all of these responses are more likely for those from marginalized groups. That is presumably a key reason why the authors we discussed all shy away from asserting a fully general duty to object. And yet, they argue, if one is safe from attack, then there may be good reason to endure the awkwardness or the frustration of being ignored—it is possible that one will be listened to. And, importantly, not objecting may either communicate agreement or confer authority on the speaker. If possible, it is important to object in order to avoid this.

One thing that will not happen in a face-to-face conversation: one will not, in general, bring it about that more people hear or pay attention to the problematic utterance. There are certain exceptions—e.g., if someone mutters something appalling, and many more people come to hear it when it is repeated by an objector. This exception, however, is an important one. The importance comes from the fact that it is not a bad analogy for what may happen on social media.

To understand this point requires a bit of information about how social media algorithms determine which posts are most likely to be presented to users. While this has changed over time (and there are also variations between platforms), a key feature has long been degree of engagement.[8] Engagement can take the form of simple reactions such as *likes*, but sharing, retweeting, or replying to a post is far more powerful. Such engagement is vital to—and much sought by—those hoping to build a following. Given these dynamics it is very easy to see that objecting to a post makes it *more* likely that it will be seen. And this is so whether one's reaction takes the form of a reply on Facebook or a retweet with commentary on Twitter (though the latter is the closest analogy to repeating something that has been muttered in order to criticize it). If one has politically engaged friends on social media, it is very common to see an angry post criticizing someone else on social media and directing all right-thinking people to go and register their objections. This instruction, when widely disseminated and followed, guarantees that the offensive original post will be seen far more widely than it otherwise would have been and is likely to bring notoriety to the original poster. In some instances, this

[8] This is just one key feature, but it is the most important one for the present argument. For more details, see, for example, Cooper (2020).

can be a devastating public shaming. But in others, this is how a career as a provocateur is built.

In 2010, Terry Jones, a small-town pastor without a huge following, declared his intention to burn the Koran. Initially, this was ignored by mainstream media. However, it was picked up on social media, largely by those who found his plan abhorrent. The outrage over his plan made it famous, so famous that eventually the mainstream media felt they had to cover it. Eventually, he backed away from the plan, but when that led to a reduction in media coverage, he decided to go ahead after all. The ensuing worldwide protests led to 12 deaths in Afghanistan. And none of this would have happened without social media sharing *by his opponents.* This sharing enormously amplified his message and led to horrendous real-world consequences.[9]

And this is a key problem with social media counterspeech: objecting to something on social media is very likely to *amplify* it. Since a central reason for thinking we should object is risk of harm from the utterance, we should be very worried about increasing that risk by increasing the number of people who are reached by the utterance. This concern applies equally strongly to the issue of correcting oppressive speech and to the issue of correcting falsehoods.

Generation of Sympathy

It is, and always has been, important to object in the right way, in order to avoid generating sympathy for those one is objecting to. Viciously insulting responses have always run the risk of alienating potential allies and of recruiting sympathizers to the cause one opposes. All this is true of ordinary conversations as well as social media ones. However, objecting on social media poses special risks that are worth taking note of.

When you speak up in a face-to-face conversation, you are aware of whether this is a part of a large pile-on or just an individual comment. If somebody else is raising the objection you want to raise, you generally hold back and let them do it. You do not add your voice repetitively. If you decide to add your voice to a crowd of people objecting, that is a decision you make—you do not accidentally find yourself doing it. On social media, however, things happen very quickly. You may object to something, thinking you are the only one speaking up, and then quickly find yourself part of a very large group. And this matters a great deal—small groups on social media quickly become large, and a large group may be perceived as a mob, and therefore generate sympathy for the person criticised.

Research bears this out. A study by Sawaoka and Monin (2018) compares reactions to criticisms of offensively racist or sexist posts, depending on the amount of

[9] For a full discussion of this case and its implications for social media amplification, see boyd (2018).

opposition they receive in comments. They found that a single commenter may be viewed favorably, but if there is a large group of commenters (even ten, quite a small group by social media standards), that same commenter will be viewed negatively. They take this to result from sympathy generated for a person who seems to be ganged up on by a large number of people, who come across as bullying. Since generation of sympathy is far from the desired goal of those who speak up against an offensive post, it does look like counterspeech may become counterproductive if it is too widely taken up. This concern is only enhanced by the fact that it is not always easy to know whether one is a part of a social media mob or not. As you sit there, over your coffee, you may see that nobody has responded to that problematic tweet. But by the time you retweet it with a pithy criticism, you may be one of hundreds.[10] This possibility of unknowingly inciting or joining a mob is a part of what motivates Norlock (2017: 188) to note that social media brings with it 'new responsibilities [which] include sorting out the extent to which we each have more power than we believe we do or than we think carefully about exerting, even as we exert it in online communication'.[11]

Abuse

When raising objections face to face, there is always a risk of verbal or physical abuse, especially if racist or sexist speech is at issue and if one is a member of a marginalized group. This is something that the philosophers we have discussed are well aware of and take into account. However, the risk of verbal abuse, including serious threats, is greatly magnified by social media. There are several relevant factors:

1. It is harder to assess the risk one faces, because one does not know who will end up seeing anything one writes.
2. It is very easy, and very fast, to mobilize armies of commenters to respond with vile threats and abuse.
3. This abuse can include such things as doxxing, which put one potentially in physical or financial jeopardy.

None of this is speculative. All of these behaviors are well-established. Attacks of this sort are especially common for members of marginalized groups, and especially when they speak up about racism or sexism. Soraya Chemaly writes:

> The phrase 'online harassment' is an anodyne catchphrase for a spectrum of behaviors, many of which break unenforced laws, degrade people's civil rights, reduce their ability to work, cause emotional and psychological harm, and actively inhibit their freedom of

[10] For further discussion of proportionality worries, specifically with respect to online shaming, see Billingham and Parr (2020). For further criticisms of online shaming, see Aitchison and Meckled-Garcia (2020).

[11] For more on these complexities, see also Aly and Simpson (2019).

> expression… The harassment often involves public shaming meant to humiliate and generates anxiety that comes with stranger threats and mob attacks. It also almost always alters, sometimes permanently, a person's ability to feel safe in 'real' space, to make a living, and to engage publicly and politically. (Chemaly 2019: 150)

Karen Adkins (2019: 83), similarly, discusses two prominent attempts at feminist online shaming, noting the disastrous consequences for the feminists who spoke up: 'Richards and St. Louis, as aspiring shamers, themselves became the objects of shame and rejection from their professional communities. They became the objects of sustained, public, and shaming scrutiny; they were condemned as professionals….' In the end, as Adkins notes, both were forced out of their professions.

Summing Up

Without looking at social media, the literature on counterspeech already included concerns about safety and power dynamics. We can see that these concerns are only intensified on social media. A post which might be meant as a response to one individual can end up travelling far and subjecting the poster to a quite stunning amount of vile abuse—and this is especially common for members of marginalized groups. The epistemic uncertainty introduced by this means that it can be far more difficult to know whether one is in a safe position to speak. The non-social-media literature also discussed the possibility that counterspeech might not be effective, either for practical reasons (e.g., the room is too noisy) or for political ones (bias makes one unlikely to be taken seriously). And it touched on the potential for backfire through reinforcing associations. But social media presents important new ways that this backfire can take place: through amplification via algorithms and through generation of sympathy if a lone voice unwittingly becomes part of a mob. Concerns about these sorts of backfires make social media counterspeech especially problematic. With that in mind, we now turn to alternatives. In other words, what other options do you have, as you survey the problematic posts you see online, and sip your coffee?

Not Objecting on Social Media

Alexander Brown (2019) has argued, quite compellingly, that remaining silent on social media does not have the same meaning or consequences that it has in face-to-face communication. There are three key points he makes here. First, it is rarely clear who has or has not seen a particular social media utterance. He makes a helpful comparison here to Maitra's subway car example. In that case, it is generally safe to

assume that anyone conscious and in earshot hears the racist abuse,[12] which is why their silence has an authorizing effect.[13] On social media, however, any particular speaker's silence could be due to their not being aware of an offensive or false utterance. Remaining silent, then, does not have the same licensing effect. Returning to Goldberg's concerns, it is not at all clear that any one person's silence will be interpreted as acceptance, and a general lack of response seems also unlikely to be interpreted in this way.

Brown (2019) also notes the importance of different conventions in different speech communities. There may be some online communities in which the lack of critical comments is taken to signify acceptance, but this does not seem generally to be the case. Social media users do not in fact usually feel obligated to take the time to pass negative judgment on each post that they find or false. Finally, Brown (2019) draws attention to ways of registering disagreement or offense that do not involve replying to the original utterer. Social media presents many different ways that one can communicate, not all of which have clear correlates in face-to-face communication. When someone says something which is offensive or false, an objector may respond directly. But they may also start an entirely new thread discussing the problem that they witnessed. In order to avoid amplifying an offensive or false comment, it is important not to link to that comment in any way and not to discuss it in a way that will raise the profile of the person who made it. This is why social media conversations on controversial matters can sometimes take the form of rather roundabout descriptions of problematic utterances and what is wrong with them. This strategy can be very useful for avoiding both amplification and abuse. But it may be difficult to know what is being referred to, and the discussion may not be seen by those who most need to see it—those who might have been adversely affected by the original post.[14]

A decent case can be made, then, that any obligation to respond to false or offensive speech on social media is significantly less than it would be in a face-to-face situation, simply because a lack of direct response will not have the same meaning or effects. This case is only strengthened by considering the problems we saw earlier for responding on social media.

But we are still left with a serious problem to be solved. What should be done in response to false or offensive speech on social media? Given the lessened responsibility to directly respond, one might wonder whether silence could usefully act as a response. Alessandra Tanesini (2018) has recently argued that silence can

[12] This assumption, of course, may be false if the people within earshot are hearing impaired or if they are listening to headphones. But an assumption like this is nonetheless more plausible offline than online.

[13] Brown (2019) in fact argues that it should often not be seen as having this effect even offline, but I will not go into that here.

[14] This problem is very much heightened by the presence of online epistemic bubbles (which have the consequence that one's post may be seen only by those one already agrees with) and echo chambers (which have the consequence that one's objections will not be viewed as credible by those one is disagreeing with). See Nguyen (2020) for this distinction.

offer an eloquent means of objecting. However, for this to succeed, the silence has to be witnessed and understood as an objection. Key examples are resistant refusals to meet demands that one speaks and large silent protests. It is difficult to see how a failure to comment on a social media post could have the sorts of effects that one can obtain with these techniques. Eloquent silence, then, seems more likely in face-to-face situations.

So we still need to think through effective responses. The reflections so far have, I think, made it clear that there are very serious difficulties with any individual responding directly to offensive or false speech and hoping to correct it. So we turn now to institutional/group responses.

Institutional and Group Responses

Institutional Responses

To some extent, as we have already hinted, these issues are not new. In the 1960s, George Lincoln Rockwell, head of the American Nazi Party, carefully exploited the way that counterspeech could amplify and generate sympathy for his political movement. He did this by booking talks on liberal college campuses, where he knew he would be met with protests. This generated media coverage. The media coverage of the protestors who were (quite reasonably) angry also generated sympathy and financial donations for his cause. In response, Jewish groups developed a strategy that they called 'quarantine': they called on people, crucially including the media, to ignore the speeches. They asked campaigning groups not to protest, and they asked media not to cover the speeches.[15]

A key thing which is different now is the massive difficulty of succeeding in a quarantine strategy. Online responses are often not centrally organized in the way that a campaigning group might be, and even when they are, there may be many online campaigning groups—it would be very difficult to succeed in reaching all the relevant people and convincing them not to object. One *might* be able to get some mainstream media to agree not to cover something, but given media polarization, it seems clear that some large media operations, seeking to profit from controversy, would not be amenable to this idea. Even if they were, however, the Terry Jones case is quite a cautionary tale: mainstream media did ignore his announcement. But eventually the outcry on social media became strong enough that they felt they needed to cover it.

A quarantine as previously practiced, then, will not succeed. But social media also presents new methods that may be used. Certain recent moves by social media companies can be seen as attempts to reimagine quarantine for our times. Both

[15] See Beckett (2017); Donovan and boyd (2019).

Twitter and Facebook have made efforts to remove Q Anon conspiracy groups.[16] Social media companies are attempting to remove falsehoods about Covid-19.[17] They are also labelling certain false claims, about Covid-19 or the 2020 US Presidential election result, as false (not quarantining in this case, but providing a real-time correction rather than relying on users to do so).[18] Donovan and boyd (2019: 14) argue for a more nuanced proposal, *strategic amplification*, which they describe as 'a complex recognition that amplifying information is never neutral and those who amplify information must recognize the costs and consequences of publication'. One suggestion they offer is that amplification could be thought of as something that has to be *earned* from social media companies, suggesting that 'platforms can define successful recommendations and healthy feeds as those maximizing respect, dignity, and other productive social values. They can actively downweigh divisive, cruel, hateful, or antagonistic content.' (Donovan and boyd 2019: 13)

It remains to be seen how effective these efforts will be, but they seem clearly preferable to relying on individual users to take huge personal risks to raise objections and corrections that may only succeed in amplifying falsehoods and generating abuse.[19]

Group Responses

There are also some highly innovative group efforts to respond to false or offensive speech by reducing its prominence and/or raising the profile of countervailing views. In 2015, social justice advocates in Italy began a novel campaign to block the hatred being spread by far-right politician, Matteo Salvini. Salvini's social media feeds had become very effective vehicles for spreading anti-immigrant and racist sentiments, particularly targeting refugees and the Roma. In response, Progetto Kitten was born. This project involved activists flooding Salvini's social media feeds with photos of kittens—making it difficult to even find the posts stoking hatred.[20] An even more recent such effort involved taking over the twitter feed of the Proud Boys, a violent far-right group, with photos of 'proud boys', understood as meaning gay men expressing their pride (Bryant 2020). Similarly, the #iamhere group works to flood comment sections and social media feeds with supportive and accurate posts in order to combat online falsehood and misinformation (Eyre and Goillandeau 2019).

[16] See BBC News (2020), Timberg (2020).

[17] See Scott (2020), Reuters (2020).

[18] Individuals may, however, still have a role to play: these responses depend in some significant part on individuals reporting problematic posts.

[19] One problem with institutional responses so far has been biases and errors in how they are applied. See Chemaly (2019) for a discussion of these.

[20] My thanks to Martina Rosola for calling this to my attention. See Zaffarano (2015).

Responses like the above present one form of positive group-based counterspeech. But they are not the only one. Maxime Lepoutre (2019) discusses the importance of positive counterspeech, noting substantial evidence that it is more effective to combat a false story with a different, true story than with the mere negation of the false one. But Lepoutre also takes very seriously the difficulty of undoing the harms of false or oppressive speech. This is why he argues for a focus instead on *preemptive* counterspeech—educational efforts that can 'condition the conversational setting to make it inhospitable to ignorant speech' (Lepoutre 2019: 181). Although it may be too late to do anything helpful about your acquaintance's problematic social media post, Lepoutre would suggest, perhaps, it should serve as motivation to make preemptive efforts. This could consist of posting articles about how to spot untrustworthy sources, but at its most effective, it will surely involve large-scale educational efforts. Again, group-based (possibly institutional) counterspeech is likely to be more effective.

Adkins (2019) discusses one especially effective group-based response to sexist speech, especially notable because the original individual attempt at online shaming led to devastating consequences for the shamer while the group response was almost universally acclaimed. This was the case of Nobel Laureate, Tim Hunt, who made a stunningly sexist joke at a conference: 'let me tell you about my trouble with girls. 3 things happen when they are in the lab; you fall in love with them, they fall in love with you and when you criticize them, they cry' (Adkins 2019: 80). Connie St. Louis criticized this joke and ended up being forced out of science journalism. But women scientists started a hashtag, #distractinglysexy, featuring 'pictures of women vamping while holding test tubes, captions sarcastically praising themselves for managing to stave off sobs as they examine slides of tissue under microscopes or excavate archaeological sites' (Adkins 2019: 89). Adkins notes that this constructively redirected attention to the wide range of women scientists rather than to Hunt and also that participants gained safety and anonymity through the collective nature of the effort.

Conclusion

Many problems for counterspeech were already recognized in the philosophical literature which focused on face-to-face communication, such as oppressive power dynamics and dangers of speaking up (especially for members of marginalized groups). There were also concerns about impracticality and situations which make successful counterspeech less likely to succeed. All of these problems are greatly magnified by the workings of social media. Moreover, social media adds to this new ways of bringing about unwitting amplification and uncertainty regarding one's conversational context. Social media, however, also presents some new and potentially promising avenues for institutional and group approaches to false and oppressive speech. It is too early to tell which methods will be most effective for combatting falsehoods and hate-filled utterances on social media. But it does seem

clear that direct individual responses are less likely to succeed than either group or institutional responses. The individual responsibility to issue corrections, then, is at the very least substantially lessened. So, to return to the scene from which we started: when you see those problematic utterances on your social media feed, you might be better off having another cup of coffee and thinking carefully and strategically about how to involve groups and institutions in fighting this problem, which is unlikely to be conquered through individual direct confrontations.

References

Adkins, K. (2019). When Shaming is Shameful: Double Standards in Online Shame Backlashes. *Hypatia, 34*(1), 76–97. https://doi.org/10.1111/hypa.12456.

Aitchison, G., & Meckled-Garcia, S. (2020). Against Online Public Shaming: Ethical Problems with Mass Social Media. *Social Theory and Practice.* https://doi.org/10.5840/soctheorpract20201117109.

Aly, W., & Simpson, R. (2019). Political Correctness Gone Viral. In C. Fox & J. Saunders (Eds.), *Media Ethics, Free Speech, and the Requirements of Democracy* (pp. 125–143). New York: Routledge. https://doi.org/10.4324/9780203702444.

BBC News (2020). Facebook Bans QAnon Conspiracy Theory Accounts Across All Platforms. BBC News, 6 October. https://www.bbc.com/news/world-us-canada-54443878. Accessed 11 December 2020.

Beckett, L. (2017). George Lincoln Rockwell, Father of American Nazis, Still in Vogue for Some. The Guardian, 27 August. https://www.theguardian.com/world/2017/aug/27/george-lincoln-rockwell-american-nazi-party-alt-right-charlottesville. Accessed 11 December 2020.

Billingham, P., & Parr, T. (2020). Online Public Shaming: Virtues and Vices. *Journal of Social Philosophy, 51*, 371–39. https://doi.org/10.1111/josp.12308.

boyd, d. (2018). Media Manipulation, Strategic Amplification, and Responsible Journalism. Data and Society: Points, 14 September. https://points.datasociety.net/media-manipulation-strategic-amplification-and-responsible-journalism-95f4d611f462. Accessed 11 December 2020.

Brown, A. (2019). The Meaning of Silence in Cyberspace. In S. Brison & K. Gelber (Eds.), *Free Speech in the Digital Age* (pp. 207–223). New York: Oxford University Press. https://doi.org/10.1093/oso/9780190883591.001.0001.

Bryant, M. (2020). LGBT Twitter Users Tease Far-Right Group By Taking Over Proud Boys Hash Tag. The Guardian, 5 October. https://www.theguardian.com/world/2020/oct/05/proud-boys-hashtag-lgbt-twitter-users. Accessed 11 December 2020.

Caponetto, L. (2020). Undoing Things With Words. *Synthese 19*, 2399–2414. https://doi.org/10.1007/s11229-018-1805-9.

Chemaly, S. (2019). Demographics, Design, and Free Speech. In S. Brison & K. Gelber. (Eds.), *Free Speech in the Digital Age* (pp. 150–169). New York: Oxford University Press. https://doi.org/10.1093/oso/9780190883591.003.0010.

Connolly, P. (2020). *Two Types of Conversation: Face-to-face and Digital.* PhD dissertation. Sheffield: University of Sheffield.

Cooper, P. (2020) How the Facebook Algorithm Works in 2020 and How to Make it Work for You. Hootsuite, 27 January. https://blog.hootsuite.com/facebook-algorithm/. Accessed 12 January 2021.

Davich, J. (2018). 'No, ma'am': McCain wasn't merely a politician, but a true patriot. Chicago Tribune, 30 August. https://www.chicagotribune.com/suburbs/post-tribune/opinion/ct-ptb-davich-john-mccain-no-maam-moment-st-0831-story.html. Accessed 15 January 2021.

Donovan, J., & boyd, d. (2019). Stop the Presses? Moving From Strategic Silence to Strategic Amplification in a Networked Media Ecosystem. *American Behavioral Scientist 65*(2), 333–350. https://doi.org/10.1177/0002764219878229.

Eyre, M., & Goillandeau, M. (2019). Here, Here: The Swedish Online Love Army Who Take on the Trolls. The Guardian, 15 January. https://www.theguardian.com/world/2019/jan/15/the-swedish-online-love-army-who-battle-below-the-line-comments. Accessed 15 January 2021.

Goldberg, S. (2020a). *Conversational Pressure.* New York: Oxford University Press.

Goldberg, S. (2020b.) The Promise and Pitfalls of Online Conversations [YouTube Video]. 7 February. London: Royal Institute of Philosophy London Lecture Series. https://youtu.be/DfxhOmRrDcE. Accessed 10 December 2020.

Grice, H. P. (1991). *Studies in the Way of Words.* Cambridge, MA: Harvard University Press.

Johnson, C. R. (2018). Just Say 'No': Obligations to Voice Disagreement, *Royal Institute of Philosophy Supplement 8,* 117–138. https://doi.org/10.1017/S1358246118000577.

Klieber, A. (2020). Conversational Silence Reconsidered (draft).

Lackey, J. (2018). Silence and Objecting. In C. R. Johnson, (Ed.), *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public* (pp. 82–96), New York: Routledge. https://doi.org/10.4324/9781315181189.

Langton, R. (2007). Disenfranchised Silence. In M. Smith, R. Goodin, & G. Geoffrey (Eds.), *Common Minds. Themes from the Philosophy of Philip Pettit* (pp. 199–214). Oxford: Oxford University Press.

Langton, R. (2018). Blocking as Counterspeech. In D. Harris, D. Fogal, & M. Moss (Eds.), *New Work on Speech Acts* (pp. 144–164), New York: Oxford University Press. https://doi.org/10.1093/oso/9780198738831.001.0001.

Lepoutre, M. (2019). Can 'More Speech' Counter Ignorant Speech?. *Journal of Ethics and Social Philosophy 16*(3), 155–191. https://doi.org/10.26556/jesp.v16i3.682.

Maitra, I. (2012). Subordinating Speech I. In I. Maitra & M.K. McGowan (Eds.), *Speech and Harm: Controversies Over Free Speech* (pp. 94–120), Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199236282.001.0001.

McGowan, M. K. (2012). On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racial Discrimination. In I. Maitra & M.K. McGowan (Eds.), *Speech and Harm: Controversies Over Free Speech*, (pp. 121–147). Oxford: Oxford University Press.

McGowan, M. K. (2018). Responding to Harmful Speech. In C. R. Johnson (Ed.), *Voicing Dissent* (pp. 182–200). New York: Routledge. https://doi.org/10.1093/acprof:oso/9780199236282.003.0006.

Nguyen, C. T. (2020). Echo Chambers and Epistemic Bubbles. *Episteme, 17*(2), 141–161. https://doi.org/10.5840/socphiltoday201762343.

Norlock, K. (2017). Online Shaming. *Social Philosophy Today,* 33, 187–197. https://doi.org/10.5840/socphiltoday201762343.

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior, 32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2.

Pettit, P. (2002). Enfranchising Silence. In P. Pettit, *Rules, Reasons, Norms* (pp. 367–378). Oxford: Clarendon Press. https://doi.org/10.1093/0199251878.001.0001.

Reuters (2020). Facebook Bans False Claims About Covid-19 Vaccines. NBC News, 3 December. https://www.nbcnews.com/tech/tech-news/facebook-bans-false-claims-covid-19-vaccines-rcna188. Accessed 11 December 2020.

Sawaoka, T., & Monin, B. (2018). The Paradox of Viral Outrage. *Psychological Science 29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658.

Scott, M. (2020). Social Media Giants Are Fighting Coronavirus Fake News. It's Still Spreading Like Wildfire. Politico, 3 December. https://www.politico.com/news/2020/03/12/social-media-giants-are-fighting-coronavirus-fake-news-its-still-spreading-like-wildfire-127038. Accessed 11 December 2020.

Simpson, R. M. (2013). Unringing the bell: McGowan on oppressive speech and the asymmetric pliability of conversation. *Australasian Journal of Philosophy*, *91*(3), 555–575. https://doi.org/10.1080/00048402.2012.704053.

Swire-Thompson, B., DeGutis, J., & D. Lazer. (2020). Searching for the Backfire Effect: Measurement and Design Considerations, Journal of Applied Research in Memory and Design Considerations. *Journal of Applied Research in Memory and Cognition, 9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006.

Tanesini, A. (2018). Eloquent Silences: Silence and Dissent. In C. R. Johnson (Ed.), *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public* (pp. 109–128). New York: Routledge. https://doi.org/10.4324/9781315181189.

Timberg, C. (2020). Twitter Banished the Worst QAnon Accounts. But More Than 93,000 Remain on the Site, Research Shows. Washington Post, 3 October. https://www.washingtonpost.com/technology/2020/10/03/twitter-banished-worst-qanon-accounts-more-than-93000-remain-site-research-shows/. Accessed 11 December 2020.

Zaffarano, F. (2015). La Pagina Facebook di Salvini e Invasa Dai Gattini. La Stampa, 7 May. https://www.lastampa.it/politica/2015/05/07/news/la-pagina-facebook-di-salvini-e-invasa-dai-gattini-1.35259558. Accessed 11 December 2020.