On Amplification

Eliot Michaelson, Rachel Sterken, and Jessica Pepp[1]

1. Introduction

When thinking about online speech, it's tempting to start with questions like: What's new here? Do online speech environments enable new types of speech acts, new semantic phenomena, new expressive effects? In other words: how has the shift to online speech *fundamentally changed* how we use language to communicate, coordinate, obfuscate, rouse, empower, disempower, insult, etc.? What hidden truths might online speech reveal about the nature of meaning and communication more broadly?

We too have been tempted by this sort of question; indeed, we still are. But here we are going to make the case for something slightly more modest: not that online speech environments have given rise to a new sort of speech act, semantic phenomenon, or expressive effect, but rather that they are *structured* in a way that is very unlike offline speech environments and which has important ramifications for our social and moral lives. In fact, we will argue, there are certain important respects in which it is difficult to conceive of how these environments *could* be structured anything like our more ordinary offline speech environments. Online social networks, we claim, are pervasively *amplificatory* in ways that offline environments very rarely are. And it is difficult to see how they could be otherwise—though it is relatively easy to imagine how that amplification might function differently.

What do we mean by this? As a first pass, we take an amplificatory environment to be one that favors the speech act of amplification—one that makes amplificatory speech acts a natural and available option. But this can't be quite right. For we also think that amplification online is often a predictable side-effect of other sorts of speech acts—and that strikes us as an important structural change as well. So we will, tentatively at least, propose to think of amplificatory environments as those which *either* favor the speech act of amplification *or* which are structured in such a way that amplification is a predictable side-effect of engaging in a range of other speech acts. We'll return to the details of what we mean by all this below. To foreshadow a bit, however: one important upshot of amplification in online environments is that, whereas in offline environments it might once have been appropriate to try to fight hate speech by confronting it, this

[1] This work is entirely collaborative. Names appear in no particular order.

strategy looks far more questionable in online environments. That's because, we claim, the amplificatory nature of these environments means that most attempts to confront hate speech will inevitably have the effect of amplifying it. This becomes particularly clear in instances of so-called 'rage farming'.

Our primary goal below will be to expand on this understanding of amplificatory environments. We'll start in §2 by introducing the speech act of amplification in offline environments, taking care to distinguish it from its less savory analogue, *appropriation*. In §3, we'll look at what's distinctive about online amplification. In §4, we will finally be in a position to offer a fuller explanation of what we mean by an *amplificatory environment*. With this part of the project in hand, we'll turn in §5 to explain how social media environments generate what we'll call the 'amplification trap' for counter speech, how intentionally laying this trap gives rise to the phenomenon of *rage farming*, and what the extant options are for avoiding the trap. In §6, we will argue that there are ways of restructuring such environments so as to mitigate the risks of falling into the trap. But we are doubtful that social media networks are incentivized or governed in the right ways to undertake such restructuring. §7 concludes with some big-picture suggestions regarding how to better incentivize social media providers to mitigate some of the negative effects of amplification.

2. Offline Amplification

The phenomenon we call 'amplification' can take place in both online and offline environments—though, we claim, it is much more prevalent online. To start with, let's look at some cases of offline amplification.

Consider a philosophy seminar on lying and deception. Sam, a new PhD student, says:

(1) Lies are wrong in virtue of violating a fair play obligation.[2]

The seminar leader goes on as if nothing has been said. Another, more advanced PhD student, Kim, then goes on to say one of two things, either:

[2] We take this from Berstler (2019), not because we're aware of anything like this ever having happened to her but just because it's a cool idea.

(2) I think Sam made a really good point there: lies are wrong in virtue of violating a fair play obligation.[3]

Or just:

(3) It seems to me that the distinctive wrong of lying is that the speaker violates a fair play obligation.

The former is an instance of amplification. The latter, in contrast, is most naturally read as an instance of what we will call 'appropriation'.[4]

What's going on in each of these versions of the case? In both, we take it, Kim repeats what Sam said. In the first case, the repetition is verbatim whereas in the second it isn't—but that doesn't really matter. Kim could just as easily have paraphrased Sam's point in (2) and repeated it verbatim in (3) while engaging in the same sorts of speech acts. What seems to characterize cases of the former variety is that they involve the explicit acknowledgment that someone else has already made the relevant conversational contribution. In cases of amplification, the speaker takes care to credit the ideas involved to their original author. In cases of appropriation, in contrast, the speaker fails to take due care in this respect. Indeed, they might even actively try to avoid offering due credit for the ideas their utterance conveys.

Quotation is essential to neither the speech act of amplification nor that of appropriation. Paraphrase will do just as well. Also inessential is the setting; amplification and appropriation can take place in a wide variety of communicative situations. Quoting with due care in the course of a public speech, a journalistic article, or an academic presentation—all of these are acts of amplification. Likewise with instances of credited paraphrase. Plagiarism, on the other hand, is a sort of appropriation with which we are all sadly familiar, and which can take place in a range of contexts.

Now let's step back and try to characterize these two sorts of speech acts in a more general manner:

---

[3] Alternatively, the speaker might directly say 'I want to amplify Sam's point: lies are wrong in virtue of violating a fair play obligation.' We take the fact that one can say this sort of thing as weak evidence in favor of the claim that amplification is illocutionary.

[4] When the speaker in (3) is male and the original speaker female, this is often referred to as 'hepeating' in recent feminist work. See especially Horisk (2021) and McGowan (2021).

Appropriation:
A speech act is an instance of appropriation iff:
(i) it repeats some content (either via direct quotation, indirect quotation, or paraphrase) that has recently been entered into the conversational record, whether or not it is in the common ground,
(ii) it includes no sufficient acknowledgment (either tacit or explicit) that this content has already been entered into the conversational record, and
(iii) it serves to make this content more salient to the audience.

Amplification:
A speech act is an instance of amplification iff:
(i) it repeats something (either via direct quotation, indirect quotation, or paraphrase) that has recently been entered into the conversational record, whether or not it is in the common ground,
(ii) it includes a sufficient acknowledgment (either tacit or explicit) that this content has already been entered into the conversational record, and
(iii) it serves to make this content, and its authorship, more salient to the audience.

Appropriation and amplification overlap in the element of repetition—i.e., criteria (i) in the definitions above. Where they differ is with respect to conditions (ii) and (iii), having to do with acknowledging the authorship of the relevant content and making that authorship salient to other participants in the conversation.

A few clarifications are in order. First, we would flag that we are not entirely certain how to think about the broader category of speech acts to which amplification and appropriation belong. We are inclined to see amplifying and appropriating as *illocutionary* acts, at least in the broad sense of being a sort of act which amounts to more than just the (mere) production of a meaningful expression, but which is nonetheless accomplished by producing a meaningful expression.[5] One might wonder whether these are instead best understood as *perlocutionary* acts, since condition (iii) in the characterisation of each act concerns what looks like a downstream effect of the utterance—namely, the degree of salience to the audience of a content and its source. Adjudicating this taxonomical issue depends both on one's theory of illocutionary acts and on how one understands salience. If salience is a matter of the actual psychological states of the audience, this might augur for categorizing amplification and appropriation as perlocutionary acts. But if salience is instead a matter of the

[5] Here and throughout, we do not mean to commit to any particular theory of illocutionary acts, though we are, of course, drawing the label from the work of Austin (1962).

dispositional availability of a content to an audience, it is arguable that Kim's utterances themselves increase salience, regardless of actual downstream consequences. Without trying to settle these issues here, let us just note the following. Appropriation is not just asserting something, but rather asserting it in a special sort of way, in order to take credit for it where that cuts against the conversational record. Likewise, amplification isn't just asserting something, but doing so in order to make more salient something that was already entered in the conversational record, including the credit for that earlier entry in the record. Interestingly, if we're right that these are best understood as illocutionary categories, then it looks like some illocutionary acts can be unintentional.

Second, while we are not particularly wedded to the Austinian framework, we have difficulty seeing how we might productively theorize these phenomena without appealing to something like the resources that Austin proffers. Consider, for instance, the possibility of treating instances of amplification and appropriation as proposals to update the common ground.[6] In that case, an instance of appropriation will be understood as a proposal to update the common ground with some proposition *p*. But that just looks like the characteristic effect of an assertion! So we lose the important distinction between appropriation and ordinary assertion.[7] With amplification, it's likewise unclear how the intended effect could be characterized in terms of proposals to update the common ground: for, in many cases at least, *p* will already be a part of the common ground. One is trying to make that aspect of the common ground more salient or prominent—not add it again. Nor is one proposing to add something like: *that p ought to be more salient*. Rather, one is simply trying to make *p* more salient. So it is not at all clear to us that common ground-type frameworks are the right tool for the job here.

Third, we take it that the notion of 'sufficient acknowledgment' we rely on in each version of (ii) will often be contested. In certain contexts, it might be enough to, for instance, nod in the direction of the person who originally made the point while quoting them. In others, one is going to have to go out of one's way to explicitly acknowledge them if one is to have any hope of successfully getting the audience to credit the point to its original author. Indeed, we suspect that there will be plenty of cases where the speaker *intends* to amplify a point, but ends up unintentionally appropriating it—because they fail to grasp the standards for acknowledgment which are operative in this context *for the sort of person who was the original author of the point*. Those standards, we take it, are set by something like the reasonable likelihood of

[6] See Horisk (2021) and McGowan (2021) for discussion of some possibilities along these lines for hepeating.

[7] See Horisk (2021, 521-23) for more detailed—and, we take it, highly incisive—criticism of what the Stalnakerian can say about the hepeating. Such criticism which should also port over to the broader phenomenon of appropriation.

success in prompting the audience to acknowledge that the original author made whatever point they made; the more resistant the audience is to the idea that such-and-such a point might have been made by someone like *this*, the higher the standards for acknowledgment. In certain contexts, amplification, as opposed to appropriation, may even turn out to be impossible.[8]

Fourth, and finally, it is worth noting that, in offline environments—and indeed in online environments as well—individuals are going to differ when it comes to what we might call their 'amplificatory power'. By this, we mean the potential for amplification that an individual has, relative to a given context. So, for example, the amplificatory power of Professor Periwinkle will vary depending on whether they are giving a large public lecture or talking to their small upper-division seminar. Note that audience numbers aren't the only thing that matters for amplificatory power, however; there's also the degree to which one's intervention is likely to raise the point one is repeating to salience. If the public lecture is going well and Professor Periwinkle is a captivating lecturer, then they are likely to have more amplificatory power than someone whose audience is zoning out by the time they reach the relevant quote. Amplificatory power, as we understand it, is a loose notion, incorporating both the potential to reach more people and the degree to which one can affect what's salient in the conversation. While undoubtedly rough-and-ready, this notion of amplificatory power will prove helpful in what's to come.

3. Online Amplification

There has been a healthy debate in recent years regarding how best to understand online speech acts like liking, sharing, and retweeting (e.g. Rini 2017, Arielli 2018, Pepp, Michaelson, and Sterken 2019, Marsili 2020, McDonald 2021). In earlier work, we argued that sharing and retweeting are much like pointing at something—pointing at something online, specifically. While we are less inclined towards this view than we once were (see our forthcoming), we still think that it's right to say that *one* important function of online sharing is to offer a perspective on what's going on online that warrants attention. But whereas pointing tends to be inappropriate in instances where the thing one is pointing at has already been mutually recognized as salient, we take it

[8] As Patrick Connolly has pointed out to us, there are also likely to be types of speech where authorship simply isn't an issue. Memes are likely to be one instance of this. In such cases, we take it that the sufficiency condition is met by default. Condition (iii) of Amplification would need to be modified to make clear that authorship only need be made salient if relevant for a speech act to count as amplificatory. Our thanks to Patrick for calling our attention to this issue.

that amplification is always perfectly apt in such instances. Even when something is mutually salient, one can hope to make it *even more salient*.

So we are now inclined to think that sharing and retweeting (and perhaps liking as well) involve an essentially ampliative aspect; that is, we are now inclined to think that the pointing-like thing that we previously thought sharing and retweeting were doing is perhaps better understood as something akin to offline amplification. It involves a digital repetition of the relevant content—together with an indication of its authorship—thereby serving to make both the content and authorship information more salient.[9]

We should note, however, one reason for hesitation here: illocutionary types are standardly understood as determining the sincerity or aptness conditions for utterances. But, following Saul (2021), we take it that one can sincerely share or retweet something that one finds problematic—that one would like to see treated as *less* rather than *more* salient in the conversation. Typically, this sort of sharing or retweeting will be *commented* rather than *uncommented*. That is, if a Labour Party supporter were to retweet something by the @Conservatives account, they wouldn't typically do this without comment; rather, they might add 'This is a load of codswallop!' or 'Piss off Tory scum!' or whatever.[10] This sort of retweeting still involves an amplificatory element; it still makes one's audience more likely to see the original content, and to know who authored that content. But we don't think it's very plausible that this amplificatory element needs to be intended in order for the retweet to count as sincere.[11]

We're not sure exactly what lesson to draw from this. In offline amplification, it looks plausible that amplification will always come together with some other illocutionary act, as a sort of package deal. That is, in (2), Kim must assert that Sam asserted some content in order to amplify that content. Even if Kim quotes her while making a gesture that indicates she is simply repeating what Sam said, her amplification still comes attached to an act of quotation. So perhaps we could hope to appeal to those other

[9] See Marsili (2020) for helpful discussion regarding what this sort of digital repetition might involve.
[10] Perhaps one might tend to do differently on Facebook. In part, we take it, this is likely to have to do with norms that have developed around the default status of one's posts on these networks: Twitter feeds are public by default, whereas Facebook was (for a long time at least) a more private, 'siloed' network by default. We'll return in §4 to consider these structural features more explicitly.
[11] Granted, one might think that something similar happens in offline conversations when one says things like 'What you just said was really offensive!' The difference—and we take this to be a significant difference—is, in offline contexts, the amplificatory effects of such a speech act are effectively constrained by both proximity (one must have heard the offensive remark to have it amplified) and memory (one must remember it). So the amplificatory potential of many offline cases are relatively muted when contrasted with cases of online amplification.

illocutionary acts to set the sincerity conditions for the overall speech act. In the online case, in contrast, it's not at all clear that this option will be available. That is, in the offline case, one can quote without amplifying (i.e. by appropriating); in the online case, it would seem that amplification is an essential part of what it is to retweet or share something.

There is another important, and somewhat difficult to characterize, shift in the move online: while in offline settings, amplification is a sort of complex speech act—subsuming acts of repetition or paraphrase—in online settings, amplification is often presented as one of the *basic* things one can do, one of the most natural ways of interacting with some bit of speech in the Twitter or Facebook (or whatever) communicative space. Amplificatory speech acts like retweeting, sharing, and liking can, in the dominant social media environments of the present day at least, all be accomplished with the touch of a button.

To be clear, something akin to this feature of contemporary online communication predates social media networks like Facebook and Twitter. As soon as the 'forward' feature was introduced in email, amplifying the speech of others became one of the 'basic' things one could accomplish in this sort of speech environment. Still, amplification was more effortful than it is today: one still had to enter a list of email addresses, or the address of a listserv, or something along those lines. And emails, in contrast to actions undertaken on social media networks, are typically intended for private consumption. So even though this option was available, its effects were limited to, at most, one's contacts, their contacts, etc. What's more, most email programs display one's emails in chronological order. So forwarding didn't have further effects on how salient the content was going to be to the reader, beyond simply making it available to them in that sequence.

When you retweet something, by contrast, this raises the likelihood that your 'friends' or 'followers' will see it, not just by putting it in their 'feed' (if it wasn't there already) but also by putting it *higher* in their feed, or making it more salient. In fact, your having interacted with that post also raises the likelihood that any random Twitter or Facebook user—someone you have no direct or indirect connections to—will see the post. This effect is achieved in different ways on different social networks. But the common underlying mechanism is that the feed algorithm, which determines the order in which content is served to users, prioritizes content that has achieved more engagement, be that in the form of likes, retweets, shares, comments, etc. Different social media networks put different weights on engagement by those you follow as opposed to engagement by random strangers, but all take into account this sum total of

engagement in one way or another. For our purposes, we can safely set the details of the functions of these rival algorithms to the side.

Taking stock then, we can see at least four ways in which online amplification turns out to be interestingly different from its offline cousin.

First, almost any way of engaging with a bit of online content inevitably has the effect of amplifying that content. While there are exceptions (i.e. directly messaging the author, talking about the content offline with others who have already seen it, certain forms of subtweeting, etc.), these are not the 'normal' ways of engaging with a given bit of speech. In offline conversations, in contrast, most ways of engaging with a bit of speech will have only limited amplificatory effects. In one-on-one conversation, engaging with some bit of speech will keep that content salient, but the audience is only the original producer of that speech. Even in a group setting, the numbers tend to be low. Online, speech actions can have effects on all of one's friends and followers—and, as already noted, likely beyond that group as well.

Second, offline, amplification would seem to be parasitic on either quotation or assertion. That is, one cannot amplify some content without either quoting that content (and possibly asserting that so-and-so said such-and-such) or putting it forth again in the form of a paraphrase. Online, one can simply like or retweet something. While these acts arguably add something on top of their pure amplificatory effect, whatever it is that they add is far less clear-cut than in the case of quotation or assertion. In the case of an uncommented retweet or share, in particular, it is unclear what we would lose if we were to say that the characteristic illocutionary effect of such an act is one of pure amplification.

Third, online social networks allow for things we don't ordinarily tend to think of as 'actions' or 'speech acts' to have significant amplificatory effects. For instance, more visually-based social media networks like Instagram and TikTok rely on view time as users scroll through content in order to prioritize certain posts over others.[12] View time, for these networks, is user engagement. And even if they aren't yet doing so (something which is difficult to know for sure), it's hard to imagine that primarily text-based networks like Twitter and Reddit won't move in this direction soon enough. Offline, it's difficult to see what the parallel might be. Visually attending to the speaker might be

[12] While perhaps not exactly a social media network, the YouTube algorithm functions similarly. For a fascinating discussion of the development and effects of this algorithm, we very much recommend the New York Times' podcast Rabbit Hole.

taken as permission to continue speaking—but that's not much like amplification in the sense in which we are interested.

Fourth, and finally, social media networks make amplification *easy*. That is, whereas in pre-social networking times one typically had to make some effort to amplify an earlier bit of speech, online social media networks make this one of the least effortful ways of interacting with some content. All it takes to 'like' or 'share' or 'retweet' is a single click of a button or tap on the phone. This, we take it, amounts to a significant restructuring of the natural actions available in our speech landscape; whereas previously amplification was possible, it took some thought and effort. Now, in online social networks, amplification is a prominently available option.

4. Amplificatory Environments

We have already hinted at the basic idea of an amplificatory environment: a speech environment that favors the act of amplification and tends to make amplification pervasive. This, we take it, can happen in a few different ways. First, the relevant speech environment can make readily available something like a purely amplificatory speech act—something like what we take uncommented sharing or retweeting to be.[13] Second, the relevant environment might make it the case that amplification is an inevitable side-effect of at least one of the other, easy speech acts one can engage in. Third, speech environments can do both, making available purely amplificatory speech acts and, in addition, making amplification an inevitable side-effect of many other speech acts. This, we take it, is what our contemporary social media networks look like: while sharing or retweeting might be purely amplificatory speech acts, liking probably is not (cf. McDonald 2021). Still, even if liking isn't a purely amplificatory speech act, amplification is a predictable side-effect of any act of liking.

So amplificatory environments are alike in making amplification easy, but differ in how exactly they do this. They can also differ in some other important ways. For example, some amplificatory environments are *siloed*, in the sense that one's acts of amplification will only have a fairly limited reach. Many of the social media networks we presently use were once like this: their feed algorithms only attended to engagement by one's friends or those one followed. Everything else was discounted. What that meant was

[13] Suppose that we're right that uncommented sharing and retweeting are purely amplificatory speech acts. Would that entail that such acts are never used to do more than just amplify a certain content? Of course not. Even if the characteristic effect of sharing or retweeting is pure amplification, one might of course amplify some content in order to communicate approval, disapproval, puzzlement, etc.

that one's amplificatory power extended, directly at least, only to the bounds of one's friend- or follower-network. Because one's amplificatory power only extended this far, however, it was more significant with respect to those it reached.

Now these social networks are *connected*, in the sense that, while one's engagement is likely to count for more amongst one's friends and followers, engagement with a post makes it more likely that any given user will see that post. Effectively then, these networks have broadened their users' amplificatory reach—increasing this aspect of their amplificatory power while simultaneously diluting the amplificatory power that they once had with respect to their own friends or followers.

Plausibly, our present social media networks are even more extreme than the minimum bar for counting as an amplificatory environment. It seems sufficient to count as an amplificatory environment that there is *at least one* basic speech act of amplification readily available, or that one of the easy speech actions one can undertake has amplification as an inevitable side-effect.[14] Our present social media networks make many such acts readily available. In fact, almost all of the available actions have at least some amplificatory effect. So our present social media networks aren't just amplificatory environments, they are what we might call *strongly* amplificatory environments.

We will offer two caveats about this notion of an amplificatory environment. First, it relies on the further notions of being an *easy* or *readily available* action, and we have not specified these at all precisely. Second, like many philosophers, we're generally uneasy with disjunctive definitions. Still, we think this one captures something interesting and important about our present communicative environment. So, while remaining open to—and, indeed, hopeful for—future refinements, we'll adopt this as our provisional understanding of amplificatory environments.

5. The Amplification Trap and Rage Farming

[14] One might think that every offline speech context involving more than one person must be an amplificatory environment by our lights, since one can always repeat or paraphrase what someone else has said. We don't think so, since we take it that offline settings don't involve a basic speech act which either amplifies or involves amplification as an inevitable side-effect. Instead, amplification is going to involve asserting-by-quoting, asserting-by-paraphrasing, or something of the sort. While we are not entirely sure how to characterize the basicness of a speech act, these seem to us far less basic in their native environment than does hitting the like button on Facebook. If one disagrees, there is an alternative way of differentiating online from offline environments: online environments make amplificatory speech acts easier and more readily available and tend to make them more impactful for more people. This would make amplificatory environments matters of degree—with online environments tending to be more amplificatory than offline ones—but that hardly strikes us as a terrible result.

As we noted at the outset, amplificatory environments raise a special kind of problem for confronting bad speech, or engaging in what is often called 'counter speech'. In a recent paper, Saul offers a succinct overview of the problem:

> And this is a key problem with social media counter speech: 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. (2021, p. 148)

We take it that the harms Saul is concerned with here can be of various kinds: there are the direct, psychological harms that hate speech can inflict on those to whom it is directed, and the indirect harm it is likely to do to others by inculcating in them or reinforcing a set of bad attitudes. Then there are the sorts of physical harms that bad speech of a variety of sorts can trigger or otherwise contribute to.[15] Bad speech, in the sense we're interested in here, is speech that generates a real, foreseeable risk of any of these types of harm. This might be hate speech, but it can also be mere misinformation.[16]

Drawing on our earlier understanding of amplificatory environments, we can now expand on Saul's claim that 'objecting to something on social media is very likely to amplify it'. First, we can clarify what this notion of amplification amounts to: making the relevant content more salient for the participants in the given conversation, which, in the space of social media, amounts to either one's followers or to the users of the social media space as a whole, depending on how that space is structured. Second, we can clarify what the relevant sort of likelihood amounts to: social media spaces, or at least those of the sort we presently use, structure our communicative options so as to make the most natural ways of trying to engage in counter speech also inevitably amplificatory of the speech one is trying to counter. Allow us to expand on this latter point.

---

[15] Saul offers the example of the pastor Jerry Jones, who in 2010 gained widespread attention on social media for his plan to burn a copy of the Koran—something which generated huge protests and, in turn, a number of deaths as the side-effect of those protests. (op cit.)

[16] Of course, sometimes speech that we wouldn't ordinarily consider 'bad' will pose a real, foreseeable risk of harm. News reports of Donald Trump's having lost the election likely posed a real, foreseeable risk of causing riots to break out in certain parts of the United States. So this understanding of bad speech will need to be tweaked further in order to exclude cases like this one. This is easier said than done, however, and is tangential to the main aims of our essay. So we leave addressing this issue to another occasion.

Contemporary social media networks make it easy to interact with bits of content via a fairly limited array of basic speech actions: liking, sharing or retweeting, commenting, upvoting, downvoting, etc. Each of these (even downvoting, at least in certain circumstances; we'll return to this shortly) is likely to have the effect of amplifying the relevant bit of content—either to one's followers, or to any user of the network. If one wants to engage in counter speech, these are your basic options online. More specifically, the natural options are going to be commenting negatively on the relevant content, retweeting that content with a negative comment attached, or downvoting or otherwise disliking the content. But if we're right that all of these options serve to amplify the relevant instance of bad speech, then one appears to face what we will call the 'amplification trap': one's options for engaging in counter speech are such that engaging in any of those options will serve to make that bad speech more salient to others. So engaging in counter speech on social media networks has a cost with no analogue in most offline situations.

We say 'most' because, it should be noted, the amplification trap *does* arise in a few, very particular sorts of offline speech situations. Imagine, for instance, that it is the late 1970s. The now-notorious white supremacist David Duke—then grand wizard of the KKK—gives a hate-filled speech in Baton Rouge, Louisiana, which is covered by one of the local papers. Somehow this comes to your attention in your capacity as a well-known opinion writer for *USA Today*. You now risk falling into the amplification trap: by writing about Duke, even in order to denounce his hateful opinions and demonstrate your solidarity with Black Americans, you amplify his message. Now, more people will know about Duke and his hateful message; some will even sympathize and take comfort in the knowledge that there are others out there who think like they do, and who are willing to say what they think.

So the amplification trap did not arise with the advent of social media. But social media forces more of us into a position where we must face the trap, for social media networks serve to redistribute the sort of amplificatory power that was once the domain of the media elite. Now, even those of us with relatively little amplificatory power online still have *some* power—and that means that we face the question, when we encounter bad speech online, of whether to counter it and run the risk of contributing to more people seeing that bad content, or to simply let it go in the hope that it will fail to gain traction on the relevant network.

Having introduced the amplification trap, we can go on to clarify a particularly pernicious use of the trap that has recently arisen on social media networks like Twitter:

various accounts associated with the American political right have taken to posting content which invites easy responses by their opponents. Sometimes this involves things as simple as a misspelling—which, it seems, many on the American left cannot help but point out. Or it can involve bad, easily objected-to arguments. Consider, for instance, a recent post from the Texas GOP account:

This post invites replies by those on the American left, who were generally more sympathetic to a robust public health response to the COVID-19 pandemic and who it

basically accuses of being hypocrites for opposing Texas' recent attempts to restrict voting opportunities and, thereby, disenfranchise minority communities. And, indeed, the post has worked wonders: as of the time of this screenshot, the post had been commented on 15,600 times and retweeted over 9,000 times, but only liked 8,000 times. Most of those comments, and commented retweets, oppose the argument the tweet itself promotes. The argument is designed to be 'dunked on', because dunking amplifies the post and helps Texas Republicans publicize their message: their voting reforms are no more onerous or unfair than the public health mandates that their opponents sought to impose on the population of Texas.[17]

In this phenomenon, now commonly called 'rage farming', rage farmers set a trap—an amplification trap, specifically—by engaging in what they expect to be perceived as bad speech, in the hope that a substantial part of their audience will spring that trap and, thereby, help amplify their message. Good rage farmers don't just engage in bad speech, they engage in bad speech that is specifically designed to be easy to respond to in the comments section. So we face a situation on online social media networks where it is not only the case that the amplification trap looms as we try to discern how to respond to bad speech, but bad actors also routinely weaponize the amplification trap to make their bad speech all the more salient on these networks.

Given our present circumstances, is there any way out of the amplification trap? We see two potential ways out. Each, however, has its drawbacks. And even the combination of the two strategies fails to fully mitigate these drawbacks.

First, most social media networks allow for (at least some) direct messaging, and such messages have no amplificatory effects on the wider social network. But, by that same token, direct messaging avoids taking any sort of public stance against the relevant bad speech. So a third-party who happens upon the bad speech will not be able to see that one has engaged in counter speech. In other words, direct messaging manages to avoid amplifying bad speech only by making one's opposition to that speech private, a matter between the original speaker and the respondent. This may be appropriate where one's aim, in engaging in counter speech, is to help the original speaker reform, to help them realize and correct the error of their ways. But this is only one of the expected goods of counter speech: in conversations with larger audiences, in particular, counter speech can demonstrate an important kind of support for, and even solidarity with, the targeted group. Direct messaging breaks off the reforming ambition of counter speech from its function as a signal of solidarity.

[17] Thanks to John Scott-Railton for bringing this case, and the more general phenomenon it instantiates, to our attention.

Second, there is a particular sort of sub-tweeting that has arisen in response to the amplification trap, specifically on Twitter.[18] This involves taking a screenshot of some bit of bad speech, tweeting the screenshot, and commenting on that tweet.[19] Such tweets provide a copy of the original tweet—one that transparently displays the content of that tweet—but are disconnected from the engagement metrics of the original. Viewing the screenshot, in other words, doesn't register on Twitter as engagement with the original tweet—at least not yet. Effectively, this sort of subtweeting opens up a space for counter speech and conversation that is cut off from the original tweet in the sense that it does not automatically filter attention towards it. One can feel free to take part in this conversation without taking on any risk of amplifying the original tweet, at least not directly. (Of course, some users may, as a result of seeing the sub-tweet, look up the original and engage with it. In this manner, the original tweet's message may be amplified in the same way it could be in an offline environment).

There is a sense in which this sort of response to the amplification trap is the converse of the direct messaging strategy: here one is not trying to engage with the original speaker in the hope that they will reform their behavior. Rather, unless the original speaker ranks among one's own followers, one is opening up a space for their speech to be countered with little chance of them being able to respond. Via such sub-tweeting, or participation in the subsequent conversation, one can signal solidarity with the target group to one's followers, or try to convince the random Twitter user who comes across the thread that the original speaker's speech was bad, without amplifying the bad speech to which one is objecting (except in the traditional offline manner).

Even the combination of these two strategies, however, won't quite mimic counter speech in an offline, non-amplificatory context. Direct messaging addresses the speaker, inviting them to reform themselves. And sub-tweeting of this sort can signal solidarity without amplifying the original tweet. In offline counter speech, however, one's message of solidarity typically reaches all of those who were potentially negatively affected, directly at least, by the original bad speech—for they are party to the very conversation in which one is participating. This sort of subtweet is more likely to be viewed by one's followers, in contrast, who may or may not include a number of those

[18] To be clear, sub-tweeting comes in a variety of forms—many of them undoubtedly negative (see Ch. XX)—and we don't mean to attribute any sort of positive valence to these other sorts of sub-tweeting. See also Saul (2021, p. 151) for discussion of something along these lines.
[19] Our screenshot above is taken from an instance of this sort of sub-tweeting, namely: https://twitter.com/jsrailton/status/1479625958332243968?s=12&t=J4bipeIxjHoQicgqEOIOXg.

targeted by the tweet. In other words, the audience of this kind of sub-tweet fails to overlap with that of the original tweet in the way that we expect it to when it comes to offline bad speech followed by counter speech.

Even with our best tools then, the amplification trap remains. There may be other strategies for dealing with bad speech online—flooding the offending account with photos of kittens, for instance[20] or attempting to 'cancel' the original poster —but the ones we have seen require concerted group effort, leaving one facing the same question in the moment whenever one encounters bad speech online. Do I respond to this, understanding that my response will amplify this speech, or do I refrain?

We are skeptical that there is a satisfying answer at the individual level. That's part of what makes the amplification trap so frustrating. What is clear is that we must learn to resist rage farming, to spot attempts at it and treat them as something akin to a phishing attempt. The bar for responding to bad speech offline is, often at least, relatively low. On social media networks, at least as they are presently structured, we must learn to set the bar higher. From the point of view of a US progressive, "dunking" on the Texas GOP just isn't worth it; the costs of their being able to spread their message are much too high. Things may be different when it comes to certain instances of outright hate speech—indeed, we are tempted to think that they are. But even there, we think, one ought to attend to whether a post has already obtained widespread prominence on the network before deciding how to respond. A great many hateful posts online are quickly forgotten, and that is undoubtedly for the best. Now on to the question of whether things could be structured better, so as to mitigate the risk of springing the amplification trap.

6. Network Mitigation

As noted in §3, our various contemporary social media networks function differently from each other—and how they function has changed over time. Might any of these variants offer us some insight into how to mitigate the risks of the amplification trap?

One way of disarming the trap would, of course, be to strip the possibility of amplification out of the network. While this might sound impossible, early versions of the Facebook and Instagram feed functioned in a way that avoided the trap entirely. Those versions of these networks only displayed posts by those a user followed—so

[20] See Zaffarano (2015) and Saul (2021).

they were siloed networks in our terminology—and they displayed these posts in chronological order. In other words, these networks used following someone as a proxy for interest in their posts in general, and time stamps as a proxy for greater or lesser salience within this restricted set of items in which one might be interested. In a siloed, chronological network like this, interacting with a post has no effect on its salience. So there is no amplification in these networks, and hence no amplification trap.

To be clear, one need move only minimally away from networks like these for this happy state to collapse. Take a siloed, chronological network that shows me not just posts by those I follow, but also posts that have been commented on by those who I follow. Such a network allows for amplification, and is hence susceptible to the amplification trap; by commenting on a post, I make it salient to those of my followers who didn't follow the original poster. So networks that avoid amplification, though they really did exist in the past, are radically different from those we use today.

Would moving back to such networks be too high a cost to bear to avoid the amplification trap? We can't really hope to answer this question without knowing more about how much harm that trap causes—which is not something that promises to be easy to figure out. What we can say is this. The way social media networks are presently structured, we think, does add something of significant value vis-à-vis earlier siloed, chronological networks: our present connected, non-chronological networks serve to give us a sort of perspective on the internet, a sense of what's going on online that users of this or that network find to be worthy of attention. Social media networks aren't the only way to obtain such a perspective; this is what the Yahoo front page once did, and what content aggregators, like Apple or Google News, and certain sorts of online-focused blog sites, like Mashable or Gawker, still do. But social media networks promise to reflect a greater number of perspectives, and that is of non-trivial value.

Another natural option is downvoting—something which already exists on Reddit and has been trialed on Twitter. However, it's worth mentioning that Reddit downvoting isn't quite as straightforward as one might think. That's because, as best we understand it, early votes are weighted higher than subsequent ones. So a post that initially receives only upvotes and then subsequently receives downvotes will be ranked higher than one that receives the same number of up and downvotes but in a more even pattern of distribution.[21] We have also seen some speculation online (though not on any particularly reputable sources) that downvoting works differently in Reddit comments, as opposed to posts, with both upvotes and downvotes contributing to a comment's

[21] See https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9. Granted, this may be dated at this point.

place in Reddit's prominence ranking. In that case, there would be a strong bifurcation between what the act of downvoting amounts to on a post and a comment, with only the former serving an anti-ampliative role (though one that is often weaker than one might expect). Even if downvotes on Reddit comments don't presently work like this, it's easy to imagine a system that does: one in which there is something called a 'downvote' which is typically used to express a dislike or disapproval of a post or comment, but which serves to make that post or comment more salient to other users on the network.

Of course, more straightforward anti-ampliative systems are also conceivable: imagine a simple social media network that allows only upvotes and downvotes, no commenting. Now imagine that the algorithm takes account of the total number of upvotes, minus downvotes—and nothing about the timing votes. This looks like a system that will avoid the amplification trap, for here there is a very clear way to both express disapproval of a post while also making it *less* likely that that post will be seen by others. Granted, we probably don't want to give up on commenting, so a system like this would have to be modified to fit with our present social media networks. But there is no great difficulty in conceiving of those integrating a straightforward anti-ampliative response option (i.e. one that counts *against* the overall salience of the original post).

Should we want our social media networks to adopt such a downvoting scheme? The answer, we take it, is not entirely clear. The problem, as one might expect, is that such a scheme would be rather easily gamed: purveyors of hate speech, for instance, might well respond to their content being downvoted by inciting their followers to downvote those they view as responsible in retaliation for that downvoting.

Could such gaming be avoided? We think that it could be, though probably not without some substantial value-judgments being made by social media companies and regulators. Here is one way that such a judgment-laden scheme might work: start by introducing downvoting and a system that allows users to flag certain bits of speech as hate speech or misinformation. Now, of course, one should expect a substantial amount of gaming at this point by purveyors of hate speech and misinformation. This is where the judgments will have to come in: social media companies, on this sort of scheme, would need to review accounts that have been flagged for e.g. hate speech to see which of these appear to be actual purveyors of hate speech and which are accounts acting in good faith, but being retaliated against.[22] The final part of the scheme involves applying

[22] One particularly tricky aspect of these judgments is that recent empirical evidence suggests that in the United States political conservatives are more likely to spread misinformation online than progressives are (see DeVerna et al., 2022). If that is right, then any efforts to tamp down on the spread of

a discount to accounts which have been identified as retaliatory bad-actors, along with followers of those accounts whose pattern of engagement tends to mimic that of the accounts themselves (or at least their recommended behavior).

Would this sort of system work in practice? That depends on a number of factors relevant to implementation that we aren't in a good position to comment on. So our point is not that it would work, but rather that we can continue to change the structure of social media networks, just as they have already changed substantially since their beginnings. And we can and should experiment with doing so in ways that have a fighting chance of allowing us to circumvent the amplification trap, by, for instance, offering us a basic action on these networks that serves to de-amplify a post as opposed to amplifying it.

7. Conclusion

We started off by introducing the twin notions of amplification and appropriation. Then we tried to define the notion of an amplificatory environment—something that we take most contemporary social media networks to be, and which sets them apart from most offline communicative environments. We took care to show that not all social media networks are amplificatory, and likewise that certain offline environments are. Finally, we were in a position to clarify the sort of puzzle we face in trying to counter bad speech in amplificatory environments—the question of how to avoid what we called the 'amplification trap'—and offered some tentative suggestions for how we might try to defuse this trap both as individuals and collectively.

We want to flag, however, that we are not at all sanguine that our present political and economic environment is one in which the companies that run our social media networks will be tempted to experiment with ways of mitigating the costs of their networks' significant potential to amplify bad speech. That's not, we think, because they have looked seriously at the benefit of having a distributed or somewhat democratized perspective on what's salient online and argued that this benefit outweighs the cost of how they presently derive that perspective, but rather because the way they are presently doing things aims to maximize user engagement, and hence facilitates maximum profits from advertisers. To change things then, we would suggest that a

misinformation are likely to be interpreted as partisan by those on the American right. This is, we would suggest, an instance of perceived injustice that we may simply have to accustom ourselves to if we really want to limit the spread of bad speech on social media.

natural first step would be to change the incentive structure under which these companies operate.

We can imagine this happening in a few different ways. One would be to import the old libel model that was long used to regulate news publications—those adversely affected by news articles containing clear factual errors or unsubstantiated claims could sue publishers for damages. If we were to treat social media companies as publishers of their content, then a similar set of incentives would presumably apply here.

This, however, strikes us as unlikely to be a very productive way of going. For one thing, publishers were responsible for the content generated by their employees, which is a very different relationship from the one between social media networks and the content producers on those networks. And it is not at all clear that we should want to move content creators on those networks into employee roles, as doing so would compromise the somewhat more democratized perspective on the internet that these networks currently provide.

A better approach, it seems to us, would be a regulatory one involving periodic audits of the networks for compliance with standards regarding hate speech and misinformation.[23] We don't view it as essential that hate speech and misinformation be quickly identified and eradicated from these networks—in contrast to, say, child pornography—but rather that its spread be seriously constrained. If significant penalties were to be assessed for failing these audits (proportional, say, to revenues and scaling based on past offenses), we should expect social media companies to rapidly work to find effective tools to constrain the spread of hate speech and misinformation. We suspect that one of those tools might prove to be a trust-weighted downvoting system, like the one we sketched above, but other systems might work equally well. We proposed another system in Pepp, Michaelson and Sterken (forthcoming) whereby the retweets (especially of news) of users with high amplificatory power would be subject to greater scrutiny (such as content moderation) or labeling indicating that the user attests to the accuracy of the news they've shared. Even in the absence of downvoting or greater scrutiny of those with amplificatory power, one might call out an instance of hate speech with reasonable confidence that the amplificatory effects of one's comment

[23] Here we are assuming that the regulators in question correctly identify hate speech and misinformation. This is a strong idealization. It is all too easy to imagine those with regulatory authority treating actual hate speech or misinformation as patriotic and accurate, while treating accurate, reasoned discourse as hate speech or misinformation.

would be outweighed by the function of whatever dampening system would then have been built into that social network.[24]

References

Arielli, E. (2018). Sharing as speech act. *Versus*, *47*(2), 243-258.

Austin, J.L. (1962) *How to Do Things with Words*. M. Sbisà and J.O. Urmson (eds.), Oxford University Press, 1975.

Berstler, S. (2019). What's the good of language? On the moral distinction between lying and misleading. *Ethics*, *130*(1), 5-31.

DeVerna, M., Guess, A., Berinsky, A., Tucker, J. and Jost, J. Rumors in Retweet: Ideological Asymmetry in the Failure to Correct Misinformation. Pers Soc Psychol Bull. 2022 Sep 1:1461672221114222. doi: 10.1177/01461672221114222. Epub ahead of print.

Horisk, C. (2021) Can McGowan explain hepeating? *Res Philosophica*, 98(3), 519-27.

Marsili, N. (2021). Retweeting: its linguistic and epistemic value. *Synthese*, *198*(11), 10457-10483.

McDonald, L. (2021). Please Like This Paper. *Philosophy*, *96*(3), 335-358.

McGowan, M.K. (2021) New applications, hepeating, and discrimination: replies to Anderson, Horisk, and Watson. *Res Philosophica*, 98(3), 537-544.

Pepp, J., Michaelson, E., and Sterken, R. K. (2019). What's new about fake news. *Journal of Ethics and Social Philosophy*, 16(67).

_____ . (forthcoming) On retweeting. In L. Anderson and E. Lepore (eds.): *The Oxford Handbook of Applied Philosophy of Language*, Oxford University Press.

[24] For discussion and feedback, thanks to Fintan Mallory, John Scott-Railton, all of the participants in the Conversations Online Workshop, and, in particular, to the editors of this volume.

Saul, J. (2021) Someone is wrong on the internet: is there an obligation to correct false or oppressive speech on social media? In A. MacKenzie, J. Rose, and I. Bhatt (eds.): *The Epistemology of Deceit in a Postdigital Era*, Springer, 139-157.

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 12 April 2023.