Journal of Ethics and Social Philosophy
VOL. 25, NO. 2 · AUGUST 2023
https://doi.org/10.26556/jesp.v25i2.1561

WHO DO YOU SPEAK FOR? AND HOW?

ONLINE ABUSE AS COLLECTIVE SUBORDINATING SPEECH ACTS

Michael Randall Barnes

> Internet trolls, predominantly anonymous posters, realized they could work together to try to destroy the lives of people who disagreed with them.
>
> —Ian Sherr and Erin Carson, "GamerGate to Trump"

THIS PAPER is about online abuse. I have two goals in directing our attention here. First, I want to show that this is a serious but neglected area of subordinating speech and that social philosophers of language have good reason to pay more attention to the specific harms of online discourse. Second, I will argue that accounting for the realities of online abuse shows that speaker authority—the thing that makes harmful speech harm in the way it does—is dynamic and emergent and often depends on the broader community of both audiences and other "speakers" in ways that current theories are ill equipped to explain.[1] I argue that much of online abuse is best understood as a type of *collective subordinating speech act*, where this collective is an *ad hoc* group that constitutes itself through speech, and it is (partly) this group that gives online abuse the subordinating force that it has. Overall, my hope is to show that attention to online abuse is useful both for illuminating the harmfulness of that important phenomenon itself and also for clarifying features it shares with "in real life" (IRL) hate speech that regularly go underemphasized in the existing literature.

It is not controversial to say that a lot of harmful speech now occurs online. Yet much of the philosophical work in this area has focused on offline life. This immediately raises two questions: (1) Can current accounts of oppressive speech adequately capture digital hate? (2) How does the (perceived) anonymity of (many) online harassers contribute to the force of their speech? To answer

1 A quick note on the term "speaker," which is a bit ill fitting for online contexts. A more accurate term may be "poster" or "user." However, throughout, I mainly use "speaker," as the alternatives are not perfect either, and because I aim to make a contribution to the speech act theory tradition, which tends to use "speaker."

these questions, I argue that the combination of anonymity and shared language offers online abusers a path to a type of group authority that lends more power to their speech than they might first appear to have. While most abusive messages online—tweets, emails, direct messages (DMs), and the like that harass, threaten, or otherwise potentially harm their targets—are uttered by individual users acting for myriad reasons, I claim that the cumulative effect of receiving dozens, hundreds, or even thousands of these messages impacts the force of these speech acts in a significant way, backing them up with a unique type of authority and making them unlike offline hateful speech. Thus, I argue that online abuse is best understood as a type of collective subordinating speech act. In other words, online abusive speech is a form of subordinating speech where the "speaker" of these messages is better conceived as a collective, though often an *ad hoc* one.[2]

To make this argument, I explore the popular model that claims that speakers can gain subordinating authority through processes like *licensing* and *accommodation*. The basic idea is that while a hate speaker can lack the necessary authority to subordinate before they make their utterance, because of the silence of bystanders, the audience fails to block the speaker's speech act, imbuing it with subordinating force. This approach has proven quite fruitful at explaining the outsized harm a seemingly powerless individual can achieve through their speech. Yet I argue that it fails to explain the dynamics of online abuse and that this failure reveals a more widespread tension in the concept.

I begin in section 1 by outlining some ways in which internet speech is different from noninternet speech. This will, in many ways, be fairly familiar to most readers, but it is worth making explicit as these features shape our speech acts in profound ways but too often fall from view. After this general overview of the distinctiveness of online speech, I then describe, in section 2, some of the key features of my narrower topic: online abuse. In section 3, I explain how these features pose a problem for existing accounts of subordinating speech, particularly around the notion of authority. This leads me to develop an alternate conception of the subordinating authority at work in online abuse. Section 4 is devoted to developing this idea, focusing on (1) the role of anonymity and

2 My use of the term "collective subordinating speech" is a bit different from Anthonie Meijers's use of "collective speech acts" (in "Collective Speech Acts"). This discrepancy is worth clearing up right away. In short, Meijers follows a broadly Searlian framework, and for that reason explains collective speech acts in terms of collective intentions. For my part, I am interested in uncovering the authority conditions that affect the force of particular speech acts, rendering them harmful—hence my inclusion of "subordinating" in the term. Because of this harm-centric approach, I am more concerned with identifying speech acts that an audience might, for various reasons, take to be representative and backed up by a group of agents, and give them uptake that reflects this perception. But this need not be tied to collective intentions, so I do not adhere to Meijer's analysis.

(2) the use of shared language in constructing and sustaining this distributed and collective speaker authority.[3] Here I include some considerations about how this conclusion could impact the types of mechanisms social media platforms use to mitigate the harms of abuse on their platforms.

Throughout, I argue that much online abuse challenges existing accounts of subordinating speech. It, therefore, represents in some ways a distinct phenomenon. At the same time, though, I believe this analysis can also shed some light on IRL subordinating speech. That is, I aim to show how online speech makes explicit many features that it shares with offline hate speech but that tend to be ignored or de-emphasized in existing accounts. Despite the internet offering bigots and abusers new tools and strategies, for the victims, the experience of being targeted by such abuse can be remarkably similar. The examination of online abuse, therefore, helps reveal key features of subordinating speech across mediums. I make these connections explicit in my conclusion.

1. INTERNET SPEECH: IT'S DIFFERENT

To state the obvious, online speech is different from offline speech. Terms like IRL, "meat space," and others make plain what we all know at a moment's reflection: what occurs online and through our screens is different and distinct from what happens outside of those parameters. This is not to claim, though, that "Twitter isn't real life." Far from it, my position is that what we do online is just as real and significant as our offline actions but that we must appreciate the differences the medium presents.

For starters, unlike standard in-person speech, online speech is mediated by an immense infrastructure of cables, wires, servers, satellites, modems, internet service providers, electricity grids, data networks, computers and smartphones, and so much more that is at the same time incredibly obvious as well as somewhat hidden from view. This infrastructure plays a role in determining *who* is able to perform online speech, as well as how early users often set the tone for acceptable behavior long after a much larger and more diverse group of users comes online. The fact that we can trace a line from the early history of "trolling" to current tactics in online harassment suggests a lineage from the sociological history of the internet to some of the problems we now face.[4]

3 For consideration of how similar features are at play in offline contexts for some types of propagandistic hate speech, see Barnes, "Presupposition and Propaganda." For consideration of the protest speech of social movements, see Barnes, "Positive Propaganda and the Pragmatics of Protest."

4 For accounts of early trolling, see, for example, Phillips, "The Oxygen of Amplification"; Bartlett, *The Dark Net*; and Quinn, *Crash Override*. And for brief philosophical analyses

And beyond the physical infrastructure of the internet, along with its economic history, we must also acknowledge that the platforms that currently host the bulk of online speech—Meta Platforms, Google, Microsoft, Amazon, and Apple—make decisions that shape the contours of online speech. Perhaps most significant is the invisible and opaque algorithmic amplification and moderation that each platform employs.[5] However, more mundane aspects like the default settings about public and private profiles, who can send DMs to whom, message-length restrictions, image capabilities, limits on sharing, forwarding, or replies, and much more are all features that have concrete impacts on what speech acts are possible in online environments.

At this point, it must be admitted, though, that the internet is a big place and that different platforms offer different affordances.[6] So, with an admission that none of the following is universally true for all online speech, let us consider some further distinguishing features of *much* of how we now communicate over the internet.

First, a lot of online speech, as written text, is in an important way less embodied than offline speech—or at least *differently* embodied. Our texts, tweets, emails, and the like usually occur within a small screen that we interact with mainly via our thumbs and fingers. This fact is both banal and significant. It has the consequence that when reading the words of another, we can experience it as a voice within our head, perhaps in our own voice, rather than as speech directed at us from the actual lips of another agent. Talking with another becomes, in some cases, talking with oneself.

Additionally, most online speech is asynchronous, at least to an extent. This sits along a spectrum, with some formats (such as email and message boards) on one end and other nearly but not quite real-time formats on the other. But even supposedly instantaneous platforms (e.g., WhatsApp, Zoom) admit to delays, outages, and buffering that manage to interrupt what we might think of as the "normal" flow of a conversation. The effect is that entirely different norms take hold when we cannot rely on the immediate feedback of our interlocutor, even when using supposedly "live" chat applications. Simple things like how long it is appropriate to wait before following up are norm-governed practices impacted by features like read receipts and time stamps.

---

of trolling, see Barney, "[Aristotle], On Trolling"; and Cherry, "Twitter Trolls and the Refusal to Be Silenced."

5 See, for example, Tufekci, *Twitter and Teargas*; Noble, *Algorithms of Oppression*; and Lynch, *The Internet of Us*.

6 For a recent account of the notion of affordances, see Davis, *How Artifacts Afford*.

Anonymity or pseudonymity is an often-cited feature of online communications.[7] This too is best conceived along a spectrum, and one with multiple axes. While at times we may be speaking anonymously to the other participants on a forum, this does not imply that we are anonymous to the site's host. It comes in degrees, from the relatively rare total anonymity one might have in certain parts of the web, to the anonymity of a screen name that does not easily lead back to one's offline life. And I include here the kind of anonymity one finds in a crowd even when they use their real name.

There is also the ambiguous state of the audience that is typical of so much online speech. Social media posts are often characterized by a genuine uncertainty about to whom one can be said to be speaking. One's tweets, for example, might be read by only a handful of one's followers, or perhaps by thousands of strangers with whom this could be one's only interaction ever. For most users, it is simply unknown exactly to whom they are talking when they hit "send."

And, as many cases of sudden online infamy show, we can be drastically wrong about who our actual audience ends up being, like when a larger public gives uptake to utterances meant only for semi-private consumption. That this occurs demonstrates how our online speech acts are not in our control. As we speak online, our communicative goals can be seemingly outstripped by the medium, where the broader community's norms may play a greater role in determining what exactly we meant, and what we did with our words, than our own intentions.[8]

One reason this can occur is that platforms take some effort to hide from us the algorithmic architecture that renders this speech situation entirely unnatural. Facebook may ask you "what's on your mind," and Twitter might goad you to tell it "what's happening," but this is merely in support of their underlying goal to incentivize you to produce more (free) content for them. The fact is that our seemingly ephemeral expressions are cataloged in their servers where the data is mined to sell advertisements. And our willingness to share is fed by the rush of endorphins caused by carefully crafted notification systems and user-interface designs.[9]

In a classic article on the topic, John Suler notes that similar features lead to what he calls the "online disinhibition effect."[10] People, he noted, acted *differently* online than offline. He was careful to note that there is both benign and

7 See Levmore, "The Internet's Anonymity Problem"; Levmore and Nussbaum, *The Offensive Internet*.

8 Online shaming offers an instructive case. See Ronson, *So You've Been Publicly Shamed*; Norlock, "Online Shaming"; and Adkins, "When Shaming Is Shameful."

9 For one articulation of this idea, see Lanchester, "You Are the Product."

10 Suler, "The Online Disinhibition Effect."

toxic disinhibition, and more significantly that this was not meant to suggest that one's online self was somehow more *real*. Similarly, I do not mean to imply that these features of online speech make it somehow more artificial, more constrained, or less genuine than offline speech. Adaptation to online, mediated, text-and-image-based environments has been swift and full of ingenuity far beyond what platform designers could predict. A whole language of emojis and GIFs sits at our fingertips. My point here has simply been to remind us of these differences as they point to a noteworthy architecture that scaffolds our communicative acts. Fundamental questions like who or what should count as a "speaker," or how retweets, "likes," tagging, and emojis should fit into an account of utterances all need to be reexamined, as does my current question: How has the internet changed *harmful* speech?[11]

The philosophical literature on subordinating speech has seen steady growth for a few decades. And while the internet has been around almost as long, much of the philosophical work on hate speech, propaganda, and subordinating speech in general has focused on offline life.[12] In-person hate speech, like what you might see in public spaces, propaganda as it is disseminated in print or on the radio, and, more recently, microaggressions as they occur in settings like a workplace or college classroom, are the main examples.[13] This has remained the case even as more and more of our lives have migrated online.[14]

But online speech raises many new issues for social philosophers of language. The overall context of online communication—the total speech situation, as Austin would call it—is radically different from that of offline communication. To begin to explore these differences, let us briefly consider the internet's impact on propaganda—including notable subcategories like "fake news," or mis- and disinformation. One initial thought might be that all the internet has done is make it easier to spread propaganda to more people more quickly. And

11 For retweets, see Marsili, "Retweeting."

12 For a quick and nondecisive example, consider that the index for the 2012 anthology *Speech and Harm* has no entries for the terms "internet," "website," "online," or other specifically online communication mediums (see Maitra and McGowan, *Speech and Harm*). There are, of course, a few noteworthy exceptions, some of which I note below.

13 This is a large and growing literature. For important contributions, see Maitra, "Subordinating Speech"; Langton, "The Authority of Hate Speech" and "Blocking as Counter Speech"; McGowan, "On 'Whites Only' Signs and Racist Hate Speech" and *Just Words*; Stanley, *How Propaganda Works*; Tirrell, "Genocidal Language Games"; Rini, "How to Take Offense"; Saul, "Beyond Just Silencing"; and Liebow, "Microaggressions."

14 The anthology *Free Speech in the Digital Age*, edited by Susan Brison and Katharine Gelber, is an important recent entry in this area.

that would be problem enough.[15] However, the reach and speed of the internet is but one concern. Beyond these issues, further complications arise.

Regina Rini argues that social media posts can be considered a "bent" form of testimony whose features exacerbate preexisting problems. That is, our unstable norms around sharing information online—e.g., a retweet ≠ endorsement—enable old tensions to flourish in new ways. For Rini, fake news is not limited to online communications, but there is, as she says, "a strong contingent relationship between fake news and social media," making the one ripe for the other.[16] As she says:

> Perhaps people are less inclined to subject ridiculous stories to scrutiny *because* we have unstable testimonial norms on social media. A friend posts a ridiculous story, without comment, and *maybe* they don't really mean it. But then other friends "like" the story, or comment with earnest revulsion, or share it themselves. Each of these individual communicative acts involves some ambiguity in the speaker's testimonial intentions. But, when all appear summed together, this ambiguity seems to wash away.[17]

Rini's analysis shows how fake news can spread organically, where little to no malicious intent is needed, *because* of the distinct features of online communication, specifically social media. Other theorists, such as Zeynep Tufekci and Michael Lynch, worry that the personalization algorithms used on Facebook, YouTube, and other platforms make a hard problem—what to believe in our saturated information environment—even harder.[18] And, as Tufekci adds, "social media's business model financed by ads paid out based on number of pageviews makes it not just possible but even financially lucrative to spread misinformation, propaganda, or distorted partisan content that can go viral in algorithmically entrenched echo chambers."[19] The worry, therefore, is not simply that social media permits the rapid spread of propaganda, but that it has also incentivized new forms of propaganda to emerge, reach their targets, and further entrench themselves in communities.[20]

15 For an analysis of the "instantaneousness" of online hate speech, see Brown, "What Is So Special about Online (as Compared to Offline) Hate Speech?"

16 Rini, "Fake News and Partisan Epistemology," 45.

17 Rini, "Fake News and Partisan Epistemology," 49.

18 See Tufekci, "It's the (Democracy-Poisoning) Golden Age of Free Speech"; Lynch, *Know-It-All Society*.

19 Tufekci, *Twitter and Teargas*, 241. See, also, Nguyen, "Echo Chambers and Epistemic Bubbles."

20 For a particularly dramatic example of the potential developments at the intersection of technology and harmful speech, consider "deepfakes," that is, videos made using

At the extreme, the crossover between online hate and real-life violence is hard to deny. After the New Zealand mosque shootings, *New York Times* writer Charlie Warzel wrote:

> It's becoming increasingly difficult to ignore how online hatred and message board screeds are bleeding into the physical world—and how social platforms can act as an accelerant for terroristic behavior. The internet, it seems, has imprinted itself on modern hate crimes, giving its most unstable residents a theater for unspeakable acts—and an amplification system for an ideology of white supremacy that only recently was relegated to the shadows.[21]

This pattern has repeated itself in other locales, most explicitly in Buffalo, New York, where a shooter once again posted their manifesto online and attempted to livestream their acts on an online platform.

It is undeniable, therefore, that harmful speech enabled by emerging technology poses new sorts of problems of urgent concern. Violence arguably caused by online propaganda and misinformation has been reported in many countries including the US, Myanmar, Germany, India, and Canada. The role of Facebook, YouTube, and other platforms in exacerbating regional conflict is a contested debate.[22]

It is noteworthy, however, that the bulk of this debate addresses online subordinating speech as it functions in its *propagandistic* mode—as outreach or as a source of hateful beliefs that later cause harm—rather than on cases where speech is directly targeting particular individuals.[23] This is apparent in the focus on online speech's ability to manipulate beliefs and otherwise poison the information environment.[24] This sometimes leads to discussions of the "potential" harms of online hate, disinformation, or deepfakes that focus on abstract values like "democracy" or "civility" as its main victims. However, it ignores those who have *already* been victimized by online hate. In what follows, I examine online abuse as a topic worthy of serious philosophical investigation.

---

machine-learning algorithms to create the illusion that someone has said or done something they never did. For analyses, see Rini, "Deepfakes and the Epistemic Backstop"; and Rini and Cohen, "Deepfakes, Deep Harms."

21 Warzel, "Mass Shootings Have Become a Sickening Meme."

22 See Barnes, "Online Extremism, AI, and (Human) Content Moderation."

23 Note that a single speech act can play both roles at once. For an overview of the distinction, see Langton, "Beyond Belief."

24 This is perhaps the result of the fact that social epistemologists have been most active in this area.

I aim to bring out its structural elements while also drawing attention to how it is experienced by those targeted.

2. ONLINE ABUSE

The previous section provided a broad overview of a few features that make online speech distinct from IRL speech, as well as some reasons to worry about novel types of *harmful* speech online. At the general level, I believe we need to understand the peculiar features of these speech acts—including the material, structural, and design affordances that enable them—in order to assess any threats they may pose and consider how we might mitigate their harms. To demonstrate this, the remainder of the paper will focus on the narrower topic of *online abuse*. Offline models of subordinating speech do not easily accommodate the online features of this type of harmful speech, so it calls for reconsideration. In this section, I will lay out some notable aspects of online abuse; in the next I will show how these pose a challenge for standard philosophical accounts on offer.

To begin, we need a better idea of what online abuse includes.[25] Media studies professor Emma Jane articulates the breadth of the problem well in her (aptly titled) paper, "'Your a Ugly, Whorish, Slut': Understanding E-bile." Jane coins the term "e-bile" to capture what she describes as the "extravagant invective, the sexualized threats of violence, and the recreational nastiness that have come to constitute a dominant tenor of Internet discourse."[26] Jane stresses how this e-bile is found in nearly all corners of the internet and displays impressive flexibility in terms of functional use, but also that its effect varies depending on factors like who is targeted and what particular speech acts are being performed. That is, noting first how common this type of vitriolic speech is and how and when it combines with other factors can help to pinpoint when it rises to the level of online abuse.

On its commonness, and flexibility, Jane writes that

25 For some first-person accounts that touch upon the varied features of online abuse in detail, see Koul, *One Day We'll All Be Dead and None of This Will Matter*; La, "Here's How Trolls Treat the Women of CNET"; Quinn, *Crash Override*; Valenti, *Sex Object*; West, *Shrill*. For journalistic pieces on the topic, see Bernstein, "In 2015, the Dark Forces of the Internet Became a Counterculture" and "The Unsatisfying Truth about Hateful Online Rhetoric and Violence"; and Jeong, *The Internet of Garbage*.

26 Jane, "'Your a Ugly, Whorish, Slut,'" 532. Note that the topic under discussion goes by a few names: "e-bile," "cyberbullying," "online harassment," and more. I go with "online abuse," partly to follow internet safety activist Zoë Quinn, who suggests that "the term 'online abuse' is far more accurate because it perpetuates the dynamics of real-life abusive situations" (*Crash Override*, 50).

> hyperbolic vitriol—often involving rape and death threats—has become a *lingua franca* in many sectors of cyberspace. It is a commonsensical, even expected, way to, among other things: register disagreement and disapproval; test and mark the boundaries of online communities; compete and create; ward off boredom; prod for reaction; seek attention; and/or simply gain enjoyment.[27]

And yet, despite being put to many uses in so many contexts, Jane notes that "the rhetorical constructs of individual e-bile texts are strikingly similar in terms of their reliance on profanity, *ad hominem* invective, and hyperbolic imagery of graphic—often sexualized—violence."[28] She concludes that e-bile is found in nearly all corners of the internet and is used to perform a wide variety of speech acts, but at the same time has a uniformity across these usages, with expressions of sexual violence being a prominent trope.

Interestingly, Jane says that in many cases "e-bile appears to be a pleasurable—albeit competitive—game, in which players joust to produce the most creative venom, break the largest number of taboos, and elicit the largest emotional response in targets." It is a sort of commonplace online derogatoriness, and for this reason, she suggests that "what looks like hate speech might better be classed as 'boredom speech' or 'gaming speech.'"[29] However, as Jane is quick to note, while this may reflect the intentions behind many of these utterances, this does not capture the range of effects the *targets* of e-bile may experience, which can, in some cases, be very serious. She says that some of those "who have been targeted by e-bile generally report … emotional responses ranging from feelings of irritation, anxiety, sadness, loneliness, vulnerability, and unsafeness; to feelings of distress, pain, shock, fear, terror, devastation, and violation."[30] This is particularly the case, moreover, when what the target receives is not a mere one-off message, but an abundance of vitriolic and violent utterances. Their email inbox, Twitter mentions, DMs, etc., become flooded with horrendous comments and threats from a large number of strangers.

And here is where we can begin to narrow from the more general rhetorical patterns common to e-bile down toward the phenomena of online abuse. What in some contexts may be a type of expected, consensual—though misogynist—mutual banter, in other contexts can constitute a type of verbal attack. That these utterances share similar rhetorical styles—and that they are undeniably common in online communities—should not distract us from the fact that

27 Jane, "'Your a Ugly, Whorish, Slut,'" 542.

28 Jane, "'Your a Ugly, Whorish, Slut,'" 533.

29 Jane, "'Your a Ugly, Whorish, Slut,'" 534.

30 Jane, "'Your a Ugly, Whorish, Slut,'" 536.

their power to harm varies relative to the types of actions they are used to perform. Below, I will explain more fully how the utterances of online abuse function to harm their targets. Here, I simply aim to delineate the topic, noting that online abuse is partly characterized by its sheer scale and volume. This widespread circulation is what we refer to when we mean something has "gone viral." While I maintain that a particularly determined individual can inflict online abuse through "cyberstalking" or "cyberbullying," my focus will be on cases where the harassing language comes from multiple speakers.

Moreover, there is also the plain fact that if one is a member of an oppressed group offline, then that identity affects how likely they are to suffer abuse online and, of course, what form that abuse will take. Research from the Women's Media Center Speech Project confirms that women are more likely to be victims of online abuse, and the content of that abuse is overtly misogynistic.[31] Men and women of color often receive racist comments in response to mundane posts, especially if they are public figures.

For targets, these messages often form a pattern, and that pattern maps onto and is a part of broader structures of oppression. It is these two features that raise these individual pieces of e-bile from one-off oddities to become the harmful, indeed abusive, speech acts they are. However, before explaining how these utterances harm in the way they do—and how that poses a challenge to philosophical accounts of harmful speech—I first want to address the issue of motivations behind these utterances in more detail as this helps clarify my approach.

That is, as Jane and many others note, the functions and motives behind abusive rhetoric are more diffuse than might be expected. Even when directing messages toward out-group members as part of what we may call an overall abusive campaign, individual posters of vile content may do so for wildly varying reasons. This leads some commentators to suggest that they are not *really* engaging in a type of hate speech, so it is best to just "ignore the trolls." However, it is worth highlighting that many emotions besides hate motivate hate speech. As Jeremy Waldron puts it, "hatred is relevant not as the motivation of certain actions, but as a possible *effect* of certain forms of speech," that is, what this speech aims at or is likely to incite.[32]

So, while it is true that the motives and superficial purposes of online abuse might vary—one-upping, building solidarity, etc.—a more insidious function plausibly sits just below the surface: the intimidation of outsiders in order to exclude, and the reification of existing hierarchies of domination. And

31 For a brief overview of relevant survey data, see https://womensmediacenter.com/speech-project/research-statistics.

32 Waldron, *The Harm in Hate Speech*, 35. See also MacKinnon, "Pornography as Defamation and Discrimination," 808; and Smith, "Fighting Hate Is a Losing Battle."

this function, I argue below, is achieved partly through the *group activity* that online abuse becomes. It is by recognizing the red herring that is the individual speaker's—or "shitposter's"—underlying psychology, and in particular the irrelevance of their (stated) motives, that we are led to put the focus back on the act the speech *performs*, along with its expected *effects*—that is, its illocutionary and perlocutionary dimensions, to use the speech act theory terms of Austin.[33] In the next section, I turn to the philosophical literature on subordinating speech in part to demonstrate why it is not up to the task of assimilating online abusive speech into its (offline) apparatus before describing how online abusive speech attains its subordinating force.

3. THE AUTHORITY PROBLEM FOR ONLINE ABUSE

If subordinating someone through speech is a type of power that only some speakers have, then a natural question to ask is who holds this power and how do they acquire it. This is the authority problem for subordinating speech, and the question of what authority conditions enable different types of subordinating speech acts is a topic that has received sustained analysis.[34] Many compelling answers to this authority problem have been developed, including the claim that, in fact, speakers do not require any special authority to subordinate with their words or, if they do, all that is needed is a type of informal authority within a given domain. Other models show how speakers can come to gain the authority they lacked prior to speaking through processes like licensing and accommodation.[35]

This last approach has proven quite powerful, and it will be my focus as I leave the others largely aside.[36] The basic idea of accommodation is that while a speaker can lack the necessary authority to subordinate before they make their utterance, their speech act can nonetheless contain a *presupposition* of authority. If their audience fails to block the speaker's speech act by remaining silent, then this presupposition of authority is successfully added to the speech

33 Austin, *How to Do Things with Words*.

34 For helpful articulation of the problem as well as some of the main moves in the debate, see Maitra, "Subordinating Speech"; Witek, "How to Establish Authority with Words"; Bianchi "Asymmetrical Conversations."

35 See McGowan, "On Covert Exercitives"; Langton, "Speech Acts and Unspeakable Acts" and "The Authority of Hate Speech"; Barnes, "Speaking with (Subordinating) Authority."

36 I do so partly because accommodation is, in my estimation, the most popular account on offer, but also because I believe considering its faults leads us toward a better account. Quickly, I will note that an account that relies on an informal conception of authority—e.g., one that picks up on parameters of privilege like race, gender, and class—will have a harder time in online contexts, in part because of the prevalence of anonymous speakers and others whose only physical presence might be a cartoon avatar on the target's screen.

situation, understood as the "score" (following a Lewisian framework) and/or the "common ground" (following a Stalnakerian framework). The thought, explained by Rae Langton, is that speech acts, "including directives generally, and hate speech specifically, can acquire authority by an everyday piece of social magic: authority gets presupposed, and hearers let it go through, following a rule of accommodation."[37]

But online speech poses problems for accounts of licensing and accommodation. In particular, the role of *silence* in online spaces is not straightforwardly analogous to offline spaces. For this reason, Alexander Brown argues that "it can be harder to infer assent, licensing, or complicity from silence in the face of hate speech when that hate speech occurs online as opposed to offline."[38] The upshot of his analysis is that the standard story of how speech (or speakers) may be licensed to achieve subordinating authority is importantly incomplete for online speech. If bystander silence is required for licensing, but *online* bystander silence is notably different from offline silence, licensed authority may be harder (or impossible) to come by for hate speakers.

Moreover, according to the standard picture of accommodating authority, blocking—where an audience member rejects or challenges the speaker's utterance, including its presupposition—should be sufficient to cancel the authority from being accommodated. As Langton describes it: "A hearer who blocks what is presupposed, also blocks the *speech act* to which the presupposition contributes. . . . That is why blocking a presupposition can make the speech act fail."[39] It is worth emphasizing that Langton is here referring primarily to the *illocutionary* success of a speech act, not its perlocutionary effects (though it can affect this too), and this is because blocking prevents—or rather undoes, by her account—the acquisition of authority.[40] "A successful blocker," she says, "changes a past utterance from the unactualized way it would have been to the way it actually is. If a speaker's presupposed authority is blocked by a hearer . . . that blocking changes the past."[41]

37 Langton, "Blocking as Counter Speech," 152. For a more full account of the specific harm that bystander silence can contribute, see Ayala and Vasilyeva, "Responsibility for Silence."

38 Brown, "The Meaning of Silence in Cyberspace," 221.

39 Langton, "Blocking as Counter Speech," 145.

40 To see both sides of this, Langton says that "besides interfering with persuasion—with 'perlocutionary' success, in Austin's terms—blocking can interfere with the speech act itself, its 'illocutionary' success" ("Blocking as Counter Speech," 149). And later: "Blocking prevents illocutionary accommodation, tracked by score, *and* perlocutionary accommodation, tracked by common ground, achieving the latter because it achieves the former" (155).

41 Langton, "Blocking as Counter Speech, 156." As she further explains this: "Blocking can disable, rather than refute, evil speech. It can make speech *misfire*, to use Austin's label for a speech act gone wrong. It offers a way of 'undoing' things with words (to twist his

But in cases of online abuse, this does not seem to be what happens, or so I argue. Consider how, in cases of online abuse, a target might receive hundreds of messages, including *some* that are supportive and do the job of challenging the speech of harassers, right alongside comments that encourage suicide or worse. Here, there is no single, linear conversation to map the score or common ground onto. I believe this goes some way to explaining why counterspeech standardly fails to render these speech acts nonsubordinating. While it can help, it does not do the job of "blocking" or "canceling" a move in a language game as Langton hopes.

Why is this the case? I argue it is because the conversational dynamics of online speech are very unlike IRL conversations, where—paradigmatically—there are two parties who engage in a back and forth. Even when we add more participants, the image is still of a single, continuous thread where each new contribution builds upon and is constrained by what preceded it. Blocking makes sense in this context, as it is itself a contribution that future moves must acknowledge. But if you have ever looked at the replies under someone's viral tweet, you will know that this is not what is going on. Some comments get traction while others are ignored. Multiple, overlapping conversations all occur at once, playing out in a manner whose progression is hard to track. And when you add reply or quote functionality, the ability to call back to a specific moment in the exchange is enhanced. This all leads to a sort of branching of multiple conversations—if we even want to call them that—whose IRL parallel is hard to find and that do not share a single, easily traceable common ground.

Another answer to why blocking moves typically fail online emerges from considering the speech acts being performed here in more detail. As Jane notes about e-bile, "the point is rarely about winning an argument *via* the deployment of coherent reasoning, so much as a means by which discursive volume can be increased—e-bile is utilized, in other words, to out-shout everyone else."[42] Seen in this way, it becomes clearer why more speech—blocking speech—often will not work. Recognizing that its point is not to add new content to the conversational score—content that might be contested—but instead to inundate its targets with a barrage of hurtful words and imagery, shows the limits of this standard approach when the assailants number in the dozens, hundreds, or

title)—and this 'undoing' has, I shall suggest, a *retroactive* character, which Austin himself described. It offers a ticket to a modest time machine, available to anyone willing and able to use it" (145–46). For a different account that explores the potential to "undo" the past, see Caponetto, "Undoing Things with Words."

42 Jane, "'Your a Ugly, Whorish, Slut,'" 534.

even thousands.[43] Seeing this speech for what it is thus explains the question about blocking online—that is, why counterspeech cannot effectively do the blocking work it is supposed to do.

To be clear, this is not to say that counterspeech is pointless or serves no purpose.[44] It is simply to show its limits and how those limits expose conceptual problems within the accommodation framework. That is, this also demonstrates how the accommodation model relies on an overly rational, psychological picture of how content gets added to the common ground. Challenging messages that (in theory) *ought* to undo the initial speech act(s) often fail to do so (in practice), and the harm and subordination remain. We see this in how targets of abuse can still experience legitimate harms despite the presence of "blocking" utterances from others, as well as how harassers often do not acknowledge that any counterspeech even occurred and instead carry on *as if* it had not.

So, considering the apparent inability of the accommodation account to explain the force of abusive speech performed online, I believe we need to look elsewhere. Specifically, what is needed is an alternative account that can explain how seemingly powerless and often anonymous speakers can attain subordinating authority, even in the face of counterspeech. Rather than the somewhat passive model accommodation offers, I believe a much more active process is in play. Online abuses, I will argue, are best understood as cases of *collective subordinating speech acts*, as they are backed up by a collective authority attained by a chorus of speakers. In the following section, I explain how the sort of anonymity of the crowd made possible in online spaces, along with coalescence around shared language, enables a mass of speakers to attain a type of authority that impacts the force of their speech acts.

4. Online Abuse and the Construction of Collective Authority

When considering the type of online abuse I am directing us toward, it can seem obvious—trivial even—that much of the power that lies behind these utterances emerges from sheer numbers. This is part of the story, to be sure. The impact of a large number of speakers directing their hostility at a single target is not something that can be ignored. And online abuse harms in the way

43 The important role of graphic sexual and violent imagery in online abuse is, unfortunately, one aspect I mainly leave aside for this paper.

44 As Lynne Tirrell says (about IRL speech): "Challenges tend to push the game backward—they cannot undo the move but they can revoke a license. . . . Over time, enough challenges or challenges of the right kind might kill the viability of the move, depending on how local or global the challenge becomes" ("Toxic Speech," 143).

it does in part because of how awful it can be to find oneself in an unwanted spotlight, particularly when this means one is bombarded by racist, sexist, and/or transphobic commentary. However, there are additional features beyond mere numbers that come into play and give online abusive speech the particular force it has. That is, there is more to the authority conditions that enable online abuse than simply scale. Below, I describe two features that each contribute to the authority that underlies abusive speech online and, in doing so, explain the subordinating force it has.

4.1. Anonymity and the Force of (Veiled) Threats

As we have already seen, threats of physical and sexual violence are not rare online. Indeed, one of the common tropes of e-bile highlighted above is the ubiquity of violent misogyny:

> E-bile targeting women commonly includes charges of unintelligence, hysteria, and ugliness; these are then combined with threats and/or fantasies of violent sex acts which are often framed as "correctives." Constructions along the lines of "what you need is a good [insert graphic sexual act] to put you right" appear with such astounding regularity, they constitute an e-bile meme. Female targets are dismissed as both unacceptably unattractive man haters and hypersexual sluts who are inviting sexual attention or sexual attacks.[45]

And while direct threats do occur, more common is violent aggression expressed in the form of "hostile wishful thinking, such as 'I hope you get raped with a chainsaw.'"[46] While this indirect phrasing allows abusers to avoid legal trouble and skirt terms of service, it does not make these statements any less threatening to their targets. It is often, I claim, an escalation, as it seems to imply a coordinated group effort with a division of labor.

That is, veiled threats of this sort are only properly understood when we consider them in their full context, where they tend to imply a larger network of harassers. First, if the threat comes from an anonymous or unknown account—a nonfollower, for instance—that might suggest that it was directed

45 Jane, "'Your a Ugly, Whorish, Slut,'" 533.

46 Jane, "'Your a Ugly, Whorish, Slut,'" 533. Sarah Jeong calls this "colorably threatening harassment," which is: "Harassment that is not overtly threatening, but is either ambiguously threatening such that an objective observer might have a hard time deciding, or is clearly intended to make the target fearful while maintaining plausible deniability" (*The Internet of Garbage*, 33).

there by others, as the coordination of abusive campaigns is more common than many realize.[47] As Sarah Jeong reports,

> [The] examination of sustained harassment campaigns shows that they are often coordinated out of another online space. In some subcultures these are known as "forum raids," and are often banned in even the most permissive spaces because of their toxic nature. In the case of the harassment of Zoë Quinn, Quinn documented extensive coordination from IRC chat rooms, replete with participation from her ex-boyfriend.[48]

Even if there is no explicit coordination, there is often an implicit type that works just as well. One common pattern in online harassment is for an account with a large number of followers to quote tweet—a type of retweet where the retweeter can add further commentary—another user, mock them, and subtly suggest that their own followers pile on. The dynamics of social media, which reward engagement, can often lead to an escalation in harassment as users encourage each other in their shared goal of belittling the person singled out. As legal scholar Danielle Citron puts it, "online harassment can quickly become a team sport, with posters trying to outdo each other. Posters compete to be the most offensive, the most abusive."[49]

Second, and a bit more subtly, the way these utterances are given *uptake* reveals something important about how speakers accrue authority. As Lynne Tirrell argues, "our speech acts also undertake a meta-level *expressive commitment* about the very saying of what is said. Expressive commitments are commitments to the viability and value of particular ways of talking."[50] These expressive commitments can shift the boundaries of what counts as acceptable discourse in a community. And, in the case of online abuse, given that harassing speech in this medium often receives "likes" from other users, these commitments to the value of this discourse take *tangible* form. This helps shift the boundaries of permissibility.[51] Alexander Brown gestures toward this idea

47 Again, I adopt a low threshold for what counts as anonymity as I am mostly concerned with how these speakers appear to their audience. For this reason, I consider the perceived anonymity of crowds to be sufficient for anonymity in this sense.

48 Jeong, *The Internet of Garbage*, 74. While this is only one instance, further evidence suggests this practice is not as uncommon as some presume. For further examples, see Tufekci, *Twitter and Teargas*; Gray-Donald, "Canada's Right-Wing Rage Machine vs. Nora Loreto"; and Phillips, "The Oxygen of Amplification."

49 Citron, *Hate Crimes in Cyberspace*, 5.

50 Tirrell, "Toxic Speech," 144.

51 For another account on the shifting bounds of permissible speech, see Saul, "Racial Figleaves."

when he says that "the process of licensing hate speakers online could require more in the way of positive engagement with the hateful content … [such as] clicking the heart icon … or adding a supporting comment *via* the 'Reply' function."[52] I agree, and I aim to make this explicit. As I am putting it, in online contexts we can often see the shifts in the normative terrain—resulting from speech acts backed up by subordinating authority—by noting the numeric value in the "likes" and retweets harassment receives.

So, what might at first glance seem like a one-off message from a single individual can, in fact, reveal a message from a group of like-minded people. It is in this sense that it is a mistake to view the speech acts typical of online abuse through an individualistic lens. As Citron says, "when cyber mobs attack victims, individuals each contribute little to the attacks. The totality of their actions inflicts devastating harm, but the abuse cannot be pinned on a particular person."[53] This poses a problem for criminal law—Citron's focus—but, in general, taking this perspective is not too difficult; it simply amounts to listening to those who have experienced this harm. As Jeong says, "targets of harassment, particularly members of marginalized groups, may view a single comment differently than an outsider might, because they recognize it as part of a larger pattern."[54]

For those targeted by such speech, then, what is noteworthy is that online abuse can be read as a glimpse into the in-group speech of others, where marching orders are being given, are well-received, and might then be carried out by any one of the many anonymous figures on the other end of the internet. This takes a very real toll on its targets. As Lindy West says of her own experience with online harassment, questions like "Am I safe? Is that guy staring at me? Is he a troll?" easily flood your mind in public spaces.[55]

So, while anonymity poses challenges for the description of online abuse—namely, by foreclosing some standard explanations for the authoritative force of subordinating speech—it, in fact, provides a powerful tool for those who wish to inflict harm on their targets. It is the combination of anonymity and apparent group solidarity—"likes" instead of condemnation—that is a dangerous mix for targets of abuse, and, I claim, an important source of the authority these speech acts rely on to subordinate their targets.

This is evident in Quinn's description of her own experience with online abuse: "I read many of the threats in my ex's voice.… But this was somehow

52 Brown, "The Meaning of Silence in Cyberspace," 125.

53 Citron, *Hate Crimes in Cyberspace*, 24.

54 Jeong, *The Internet of Garbage*, 32.

55 West, *Shrill*.

more insidious—he wasn't just continuing his abuse; *he was crowdsourcing it*."[56] This vivid account is supported by media researcher Eden Litt's suggestion that "without being able to know the actual audience, social media users create and attend to an *imagined audience* for their everyday interactions."[57] That is to say, when we cannot directly perceive our audience, we create it in our minds. This calls back to one of the defining features of online communication I described earlier, and here we see how it impacts the force of online abuse. With this in mind, Kathryn Norlock notes that advising someone to "ignore the trolls" is beyond pointless. . . . The advice to ignore the social community as it lives in one's head is more than ineffective—*it's missing the force*."[58]

I believe this is exactly correct, that the force of online abuse—which is determined in part by the authority that sustains it—is dependent on the unique features of online communication. By seeing how anonymous avatars can become a monolith in one's mind, we can recognize a conception of subordinating speaker authority that, in fact, requires something like anonymity. It is in leveraging the target's own cognitive resources—namely, their capacity for imaginal relationships, which are necessary given text-based communication—that large-scale online abuse campaigns become more than the sum of their parts. Beyond affecting the force of individual messages, anonymity creates the semblance of cohesion where there might not, in fact, be any, thereby uniting different speakers who might not have anything in common aside from their hostile speech directed at the same individual.

Moreover, it is *through* this speech that they become united (at least in the mind of the target). It is for this reason that I refer to these as *collective subordinating speech acts*, whose subordinating authority—its capacity to harm in the distinctive way it does—is constituted by the active participation of an *ad hoc* community of speakers. Through repetition and endorsement, signaling support and solidarity, individual speech acts acquire authoritative standing in relation to a target, enabling them to harm. Each new utterance adds to the strength of the overall practice. Like accommodation, then, audience uptake secures authority for speech that, absent that uptake, would have a different pragmatic force. But as I have emphasized, in these cases, the practices that do the heavy lifting here are *active*, not passive.[59] In all these cases, speech plays an active role in solidifying the collective authority that strengthens their words,

56 Quinn, *Crash Override*, 51.

57 Litt, "Knock, Knock," 333.

58 Norlock, "Online Shaming," 194.

59 For a different but related adaptation of the concept of accommodation, see Adams, "Authority, Illocutionary Accommodation, and Social Accommodation."

turning it into the genuinely subordinating speech that it is. This is done in part through the construction of in-groups and out-groups. Herbert and Kukla point out that the recognition of insider status is something that comes into being through social practices that, in fact, constitute that status. They say that

> being recognized as an insider by insiders is not just the recognition of a separate fact; rather, this recognition plays a constitutive role in having that insider status. Part of being an insider is being recognized as one. Crucially, the relevant sort of recognition is not mere passive, conscious acknowledgment, but the kind of recognition that is built into practice.[60]

In online abuse, this takes the form of harassers cheering on other harassers through "likes," "retweets," and one-upping one another, along with other practices like coordinating on targets and sharing information.

So far, I have argued that, in cases of online abuse, anonymity—or at least, the anonymity one finds in the crowd—can contribute to the active construction of a group identity that may be wielded to inflict great harm. But anonymity is only part of the explanation I want to offer; shared, insider language is the other. I turn to this next.

4.2. Shared Language and Solidarity

To start, it is useful to note that the affordances of social media make it clear how a user's speech act is always tied to their (ever-shifting) socially constituted position—even when it is anonymous. Whether via a profile picture, short bio, hashtag, or emoji, social media brings new means of signaling identity. I want to emphasize, however, how this just amplifies what has always been the case offline. Mary Louise Pratt articulates this thought well when she writes:

> Once you set aside the notion of speech acts as normally anchored in a unified, essential subject, it becomes apparent that people always speak from and in a socially constituted position, a position that is, moreover, constantly shifting, and defined in a speech situation by the intersection of many different forces. On this view, speaking "for oneself," "from the heart" names only one position among the many from which a person might speak in the course of her everyday life.[61]

On social media, these implicit features of offline life are made fully explicit, often purposely so. Including a rose emoji or #MAGA, for example, can instantly situate a speaker as part of a wider community and communicate their broader

60 Herbert and Kukla, "Ingrouping, Outgrouping, and Peripheral Speech," 584.

61 Pratt, "Ideology and Speech-Act Theory," 63.

allegiances. These aspects of online speech allow individuals to actively construct and manage the version of themselves they present. This allows for a lot of variety, freedom, and play, including the inconsistencies that Pratt describes—e.g., concealing or emphasizing distinct parts of oneself for different platforms.

What is relevant for my purposes is how, on social media, this type of signaling often occurs by parroting the speech acts of another. "Speaking for oneself," in this context, often means speaking with the voice of another. While this is not an uncommon feature of speech, it is heightened and made explicit online, most obviously so through the use of hashtags, which allow one to visibly connect their own utterance to those of (usually) many others.[62] This feature of social media has proven powerful, as large social movements can galvanize around a hashtag that, in essence, consists in joining with the voices of others.[63] This can bring out both good avenues for effective solidarity and bad ones, as practices like the co-opting and appropriation of the words and voices of the more marginalized are all too common. For example, the phrase and hashtag "Black Lives Matter" has been taken up, twisted, and put to use for all sorts of ends, including by opposing forces.

So, while I am talking about a more general phenomenon, here I want to focus on how this can contribute to the group authority at issue in online abuse. Namely, hashtags (and related rhetorical constructions) help unify the voices of many into an *ad hoc* collective. As I will describe it, hashtags are *explicit ventriloquisms* and are a vivid example of language's role in constituting a group identity. That is, as Quinn puts it, how "the same techniques that people have used to organize important grassroots movements like Black Lives Matter can be used by people trying to destroy someone."[64]

In the course of building his account of slurs, the linguist Geoffrey Nunberg describes ventriloquisms:

> In a particular context, a speaker pointedly disregards the lexical convention of the group whose norms prescribe the default way of referring to *A* and refers to *A* instead via the distinct convention of another group

62 For a pragmatic analysis of hashtags as well as other unique features of online speech, see Kukla, "'Don't @ Me!'"

63 For an analysis of the impact of social media and other digital communication technologies on progressive activism, as well as how repressive regimes have learned to clamp down on these groups, see Tufekci, *Twitter and Teargas*. And for the story of how social media played a key role in the growth of the Black Lives Matter movement, see Khan-Cullors and bandele, *When They Call You a Terrorist*.

64 Quinn, *Crash Override*, 52.

> that is known to have distinct and heterodox attitudes about *A*, so as to signal his affiliation with the group and its point of view.[65]

That is, when a speaker uses a ventriloquism, they are disregarding the standard term that convention dictates and are instead mimicking the voice of another. In doing so, they signal their allegiance to a specific community, at least in that moment. Nunberg uses the example of a university dean using *ain't* in place of *isn't* to implicate that the knowledge being communicated was more folksy than academic.[66]

While Nunberg's main goal is to argue that slurs are cases of ventriloquisms, I aim mainly to get at an interesting feature of language, and I believe this account helps get us there. As he summarizes his view: "In a nutshell: racists don't use slurs because they're derogative; slurs are derogative because they're the words that racists use."[67]

Crucially it is not just shared attitudes that are implicated, but shared group membership:

> As [Langston] Hughes tells it, the force of [the n-word] goes beyond anything the speaker believes or feels about blacks. . . . It also evokes the things such people have *done* to blacks—with the speaker pointedly affiliating himself with the perpetrators. The word can turn a bigot from a hapless, inconsequential "I" into an intimidating, menacing "we."[68]

Without committing to this account of slurs, I do want to suggest that this analysis clarifies the pragmatic force of online abuse. Namely, the conception of ventriloquisms on offer demonstrates the potential of constructing a collective identity through shared language and tropes, as well as the ability for such a collective identity to undergird harmful speech. As I see it, hashtags and other shared rhetorical constructions function as explicit ventriloquisms, and in doing so serve to strengthen shared group identity for harassers. Hashtags are the most visible in part because they are literally visible, and their pragmatic function is to tie one utterance to many others. At their most extreme, they generate utterances with a first-person plural speaker—resulting in speech acts

65 Nunberg, "The Social Life of Slurs," 267.

66 As Nunberg explains the case: "A dean at an Eastern university [said]: 'Any junior scholar who stresses teaching at the expense of research *ain't* gonna get tenure.' In the dean's mouth, the use of the demotic *ain't* rather than *isn't* implied that his conclusion wasn't based on expert knowledge or a research survey; it was as if to say, 'You don't need an advanced degree to see that; it's obvious to anyone with an ounce of sense'" ("The Social Life of Slurs," 265).

67 Nunberg, "The Social Life of Slurs," 244.

68 Nunberg, "The Social Life of Slurs," 286. Note that this, according to Nunberg, distinguishes his view from a similar one offered by Camp, "Slurring Perspectives."

spoken by a collective "we." They perform the function, in other words, of what Hughes (as told by Nunberg) said slurs were capable of, and as a result can bring a similar subordinating authority to bear on targets.

Renée Jorgensen Bolinger develops a similar pragmatic account of slurs that can add to this story. While not a perfect parallel, Bolinger's *contrastive choice* account of slurs can add to the idea of ventriloquism by explaining how marked expressions can carry important signals about their speakers *for their audience*. As Bolinger puts it, "When we use slurs, we communicate information about ourselves and our attitudes towards the targets."[69] This information is signaled, moreover, through a speaker's decision to choose a particular term over a non-marked alternative. As she explains:

> For signals based in *contrastive choice*, the relevant behavior is the free selection of a marked expression, and performance signals that the speaker endorses a cluster of attitudes associated with the term (or, more precisely, a high probability that the speaker shares some or all of the attitudes in this cluster).[70]

Using a hashtag, it is worth pointing out, involves choice. It is literally marked—in blue, generally—and in some situations, it communicates the choice of *affiliation* or *association* with other users. But I do want to suggest that this thought applies beyond hashtags as well, which, as I said above, were simply the most visible version of this phenomenon. Some phrases, I claim, play a similar role as hashtags—and so, function as ventriloquisms—without being so explicit. Most often this occurs when a hashtagged phrase gains so much prominence that it enters the lexicon as marked in this peculiar way. Some examples likely include BlackLivesMatter, MeToo, MAGA, GamerGate, and even longer phrases like "it's about ethics in journalism," which was a common trope in GamerGate.

Or consider the use of the term "SJW," particularly as it occurs online. This is, in most cases, used pejoratively, referring commonly to individuals who promote socially progressive views like feminism and anti-racism. Importantly, this term is used almost exclusively by those who *oppose* these goals. In using this term, then, whether prefixed by a hashtag or not, speakers pragmatically convey information about their own group membership to their audience.

Again, as Bolinger helpfully explains:

> The information content of signals based in contrastive choice is linked to how marked the term is: if α is a term that is used almost exclusively by speakers who embrace ϕ, and this fact is well-known, then a contrastive

69 Bolinger, "The Pragmatics of Slurs," 439.

70 Bolinger, "The Pragmatics of Slurs," 447.

> preference for a is a high-information signal, raising the probability of the speaker's endorsing ϕ nearly to 1. The more well-known the association between a and ϕ is, the higher the information content of the signal, and thus the more strongly the contrastive choice signals the speaker's endorsement of ϕ.[71]

Since "SJW" is used mainly by its detractors, and since this is well-known, using it carries a high-probability signal that the speaker endorses these views too. The act of signaling this information performs an important function for both insiders and outsiders. In short, what terms like this do when repeated so much as to be marked in this way is to express and solidify group membership. This is a dynamic process performed primarily through speech acts. And it is through this process, moreover, that the targets of online abuse come to recognize that they are being addressed not by a single speaker, but by a mob.

This interpretation makes sense, moreover, since it is often exactly what is occurring. And as Jeong reports, it is this interpretation that makes sense of the "really bizarre phenomenon" of "all the low-level mobbers, who have little-to-no real investment in going after the target, and would not manifest any obsessions with that particular target without the orchestrator to set them off." As she explains:

> Here they resemble the zombie nodes of spam botnets, right down to the tactics that have been observed to be deployed—rote lines and messages are sometimes made available through Pastebin, a text-sharing website, and low-level mobbers are encouraged to find people to message and then copy/paste that message.[72]

Here again we see how in online abuse the implicit is often made fully explicit. Speakers are literally copying and pasting their utterances from one another, and in doing so adding strength to the subordinating force of each speech act. Shared language, along with the technological features of online communication, make this possible.

More importantly, this shows vividly why an individualist approach to online abuse is inapt for describing the force of these speech acts. It is only when we see these speakers as part of a collective, and a collective, moreover,

71 Bolinger, "The Pragmatics of Slurs," 447. Moreover, on this view, this is not reducible to speaker intentions: "Signaling on this framework is factive: a speaker signals some content ϕ when her use of an expression satisfies the conditions, regardless of whether she intended to communicate ϕ, and independent of whether hearer uptake occurs" ("The Pragmatics of Slurs," 447).

72 Jeong, *The Internet of Garbage*, 68.

that is constructed in part through the active use of shared speech acts, that we capture the pragmatic impact of these speech acts. They are, as I put it, backed up by a distinctively collective subordinating authority and so are *collective subordinating speech acts.*

Seeing online abuse as a sort of group activity encourages us not only to reject a lingering individualistic lens but also, I claim, is necessary for devising solutions to the harm they present. Reporters Max Fisher and Amanda Taub put this succinctly when they write:

> It is becoming increasingly common for groups of people, whipped into a rage by influential people on social media, to single out targets for mass campaigns of online harassment and threats. . . . *The main problem seems to be that social media companies' guidelines tend to focus on content in isolation*. Because the accounts that instigate the hatred and rage don't necessarily participate in the mass harassment directly—often their followers are the ones who send the death threats or do the doxxing—this problem is a poor fit for that approach.[73]

As this shows, tackling this problem properly requires addressing the collective from which the speech draws its power, whether it is an organic *ad hoc* group, or a preexisting community with a clear (if informal) hierarchy. Seeing this bigger picture is helpful in explaining the damage it can do to a community and it paints the way toward effective solutions. Social media companies can track this behavior—like they do all of our behavior—and, rather than basing their moderation decisions on individual pieces of content examined in isolation, they could focus on these patterns: swarming, copy-pasting, mass movements in attention across platforms, and other group-based practices rather than content.

5. Conclusion

In this paper I have examined what I take to be some key features of online abuse. I have emphasized the role of anonymity in cultivating the appearance of coordination and a division of labor in online abuse—even if in fact there is none—and shown how shared language plays an important role here as well. That is, I argued that anonymity plays a key role in building a type of collective authority for online abusive speech acts and, moreover, that the construction and endorsement of this group identity through shared rhetorical constructs like hashtags further adds to the targets' sense that they are being addressed

73 Fisher and Taub, "Social Media Has a Mob Violence Problem."

by a collective rather than individuals. I argued that this combination of anonymity and apparent group solidarity—shared phrases and hashtags, likes, and retweets—is a dangerous mix for targets of abuse, and an important source of the authority these speech acts rely on to subordinate. It is in this way, I argued, that online abuse becomes more than the sum of its parts.

These and other features build collective authority for seemingly isolated speech acts and, as I will now suggest in closing, reveal aspects of IRL subordinating speech that are often overlooked. In other words, I believe that greater attention to features similar to those I have highlighted in the online case can help bring out underemphasized aspects of offline hate speech. Racist graffiti spray painted on college campuses, slurs yelled from passing cars, white-nationalist flyers displayed in public, all invoke a sort of anonymity and group activity in a similar way to create an overall environment of exclusion. Across mediums, the force of any individual subordinating speech act draws on many other instances of similar utterances made by similar speakers, and this should be made more explicit in our accounts of its pragmatic functions. This follows from the more general observation that accounting for the realities of subordinating speech—both online and IRL—demonstrates that speaker authority is dynamic and emergent, and often depends on the wider community in more ways than simple accommodation suggests. Passive bystanders play an important role, to be sure, but greater attention must be paid to those who *actively* back up the subordinating speech of others. As I argue, this sort of contribution leads us away from an individualistic understanding of oppression, as is necessary. Online abuse makes this vivid, but I claim this is a feature shared by IRL forms of subordinating speech, and one that must be kept in mind.[74]

Australian National University
michael.barnes@anu.edu.au

REFERENCES

Adams, N. P. "Authority, Illocutionary Accommodation, and Social Accommodation." *Australasian Journal of Philosophy* 98, no. 3 (July 2020): 560–73.

Adkins, Karen. "When Shaming Is Shameful: Double Standards in Online Shame Backlashes." *Hypatia* 34, no. 1 (Winter 2019): 76–97.

Austin. J. L. *How to Do Things with Words*. Cambridge, MA: Harvard University

74 I want to thank the editors and referees for this journal, as well as audiences at numerous conferences and workshops. And special thanks are owed to Matthew Shields, Heather Stewart, and Quill Kukla for written comments at earlier stages of this project.

Press, 1962.
Ayala, Saray, and Nadya Vasilyeva. "Responsibility for Silence." *Journal of Social Philosophy* 47, no. 3 (Fall 2016): 256–72.
Barnes, Michael Randall. "Online Extremism, AI, and (Human) Content Moderation." *Feminist Philosophy Quarterly* 8, nos. 3/4 (2022).
———. "Positive Propaganda and the Pragmatics of Protest." In *The Movement for Black Lives: Philosophical Perspectives*, edited by Michael Cholbi, Brandon Hogan, Alex Madva, and Benjamin Yost, 139–59. Oxford: Oxford University Press, 2021.
———. "Presupposition and Propaganda: A Socially Extended Analysis." In *Sbisà on Speech as Action*, edited by Laura Caponetto and Paolo Labinaz. London: Palgrave Macmillan, 2023.
———. "Speaking with (Subordinating) Authority." *Social Theory and Practice* 42, no. 2 (April 2016): 240–57.
Barney, Rachel. "[Aristotle], On Trolling." *Journal of the American Philosophical Association* 2, no. 2 (Summer 2016): 193–95.
Bartlett, Jamie. *The Dark Net: Inside the Digital Underworld*. New York: Melville House, 2014.
Bernstein, Joseph. "In 2015, the Dark Forces of the Internet Became a Counterculture." *Buzzfeed*, December 23, 2015. https://www.buzzfeed.com/josephbernstein/in-2015-the-dark-forces-of-the-internet-became-a-countercult?utm_term=.jcmLwKXlW#.hbPV3edYy.
———. "The Unsatisfying Truth about Hateful Online Rhetoric and Violence." *Buzzfeed News*, November 23, 2018. https://www.buzzfeednews.com/article/josephbernstein/does-hateful-speech-lead-to-violence.
Bianchi, Claudia. "Asymmetrical Conversations: Acts of Subordination and the Authority Problem." *Grazer Philosophische Studien* 96, no. 3 (September 2019): 401–18.
Bolinger, Renée Jorgensen. "The Pragmatics of Slurs." *Noûs* 51, no. 3 (September 2017): 439–62.
Brison, Susan J., and Katharine Gelber, eds. *Free Speech in the Digital Age*. Oxford: Oxford University Press, 2019.
Brown, Alexander. "The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech." In Brison and Gelber, *Free Speech in the Digital Age*, 207–23.
———. "What Is So Special about Online (as Compared to Offline) Hate Speech?" *Ethnicities* 18, no. 3 (June 2018): 297–326.
Camp, Elisabeth. "Slurring Perspectives." *Analytic Philosophy* 54, no. 3 (September 2013): 330–49.
Caponetto, Laura. "Undoing Things with Words." *Synthese* 197, no. 6 (June

2020): 2399–2414.
Cherry, Myisha. "Twitter Trolls and the Refusal to Be Silenced." In *The Real World Reader: A Rhetorical Reader for Writers*, edited by James S. Miller, 221–25. New York: Oxford University Press, 2015.
Citron, Danielle Keats. *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press, 2014.
Davis, Jenny L. *How Artifacts Afford: The Power and Politics of Everyday Things*. Cambridge, MA: MIT Press, 2020.
Fisher, Max, and Amanda Taub. "Social Media Has a Mob Violence Problem. Could Soccer Hooliganism Prevention Offer a Model for Solving It?" *New York Times*, June 6, 2019. https://static.nytimes.com/email-content/INT_14028.html.
Fogal, Daniel, Daniel W. Harris, and Matt Moss, eds. *New Work on Speech Acts*. Oxford: Oxford University Press, 2018.
Gray-Donald, David. "Canada's Right-Wing Rage Machine vs. Nora Loreto." *Briarpatch*, April 15, 2018. https://briarpatchmagazine.com/blog/view/canadas-right-wing-rage-machine-vs-nora-loreto.
Herbert, Cassie, and Rebecca Kukla. "Ingrouping, Outgrouping, and the Pragmatics of Peripheral Speech." *Journal of the American Philosophical Association* 2, no. 4 (Winter 2016): 576–96.
Jane, Emma A. "'Your a Ugly, Whorish, Slut': Understanding E-bile." *Feminist Media Studies* 14, no. 4 (2014): 531–46.
Jeong, Sarah. *The Internet of Garbage*, rev. ed. New York: The Verge, 2018.
Khan-Cullors, Patrisse, and Asha Bandele. *When They Call You a Terrorist: A Black Lives Matter Memoir*. New York: St. Martin's Press, 2018.
Koul, Scaachi. *One Day We'll All Be Dead and None of This Will Matter*. Toronto: Doubleday Canada, 2017.
Kukla, Quill R. "The Pragmatics of Technologically Mediated Speech Acts: Don't @ Me." In *The Oxford Handbook of Applied Philosophy of Language*, edited by Luvell Anderson and Ernie Lepore. Oxford University Press, forthcoming.
La, Lynn. "Here's How Trolls Treat the Women of CNET." CNET, July 11, 2017. https://www.cnet.com/news/cnet-women-hate-troll-comments/.
Lanchester, John. "You Are the Product." *London Review of Books* 39, no. 16 (August 2017). https://www.lrb.co.uk/the-paper/v39/n16/john-lanchester/you-are-the-product.
Langton, Rae. "The Authority of Hate Speech." In *Oxford Studies in Philosophy of Law*, vol. 3, edited by John Gardner, Leslie Green, and Brian Leiter, 123–52. New York: Oxford University Press, 2018.
———. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In

Maitra and McGowan, *Speech and Harm*, 72–93.
———. "Blocking as Counter Speech." In Fogal, Harris, and Moss, *New Work on Speech Acts*, 144–64.
———. "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs* 22, no. 4 (Autumn 1993): 293–330.
Levmore, Saul. "The Internet's Anonymity Problem." In Levmore and Nussbaum, *The Offensive Internet*, 50–67.
Levmore, Saul, and Martha C. Nussbaum, eds. *The Offensive Internet: Speech, Privacy and Reputation*. Cambridge, MA: Harvard University Press, 2010.
Liebow, Nabina. "Microaggressions: A Relational Analysis of Harms." In *Autonomy and Equality*, edited by Natalie Stoljar and Kristin Voigt, 195–219. New York: Routledge, 2021.
Litt, Eden. "Knock, Knock. Who's There? The Imagined Audience." *Journal of Broadcasting and Electronic Media* 56, no. 3 (September 2012): 330–45.
Lynch, Michael Patrick. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York: Liveright Publishing Corporation, 2016.
———. *Know-It-All Society: Truth and Arrogance in Political Culture*. New York: W. W. Norton & Company, 2019.
MacKinnon, Catharine. "Pornography as Defamation and Discrimination." *Boston University Law Review* 71, no. 5 (1991): 793–818.
Maitra, Ishani. "Subordinating Speech." In Maitra and McGowan, *Speech and Harm*, 94–120.
Maitra, Ishani, and Mary Kate McGowan, eds. *Speech and Harm: Controversies Over Free Speech*. New York: Oxford University Press, 2012.
Marsili, Neri. "Retweeting: Its Linguistic and Epistemic Value." *Synthese* 198, no. 11 (November 2021): 10457–83.
McGowan, Mary Kate. *Just Words: On Speech and Hidden Harm*. Oxford: Oxford University Press, 2019.
———. "On Covert Exercitives: Speech and the Social World." In Fogal, Harris, and Moss, *New Work on Speech Acts*, 185–201.
———. "On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racial Discrimination." In Maitra and McGowan, *Speech and Harm*, 121–47.
Meijers, Anthonie. "Collective Speech Acts." In *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, edited by S. L. Tsohatzidis, 93–110. Dordrecht: Springer, 2007.
Nguyen, C. Thi. "Echo Chambers and Epistemic Bubbles." *Episteme* 17, no. 2 (June 2020): 141–61.
Noble, Safya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

Norlock, Kathryn J. "Online Shaming." *Social Philosophy Today* 33 (2017): 187–97.

Nunberg, Geoff. "The Social Life of Slurs." In Fogal, Harris, and Moss, *New Work on Speech Acts*, 237–95.

Phillips, Whitney. "The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online." *Data and Society*, May 22, 2018. https://datasociety.net/output/oxygen-of-amplification/.

Pratt, Mary Louise. "Ideology and Speech-Act Theory." *Poetics Today* 7, no. 1 (1986): 59–72.

Quinn, Zoë. *Crash Override: How Gamergate (Nearly) Destroyed My Life, and How We Can Win the Fight against Online Hate*. New York: Public Affairs, 2017.

Rini, Regina. "Deepfakes and the Epistemic Backstop." *Philosopher's Imprint* 20, no. 24 (2020): 1–16.

———. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27, no. 2 supplement (June 2017): 43–64.

———. "How to Take Offense: Responding to Microaggression." *Journal of the American Philosophical Association* 4, no. 3 (Fall 2018): 332–51.

Rini, Regina, and Leah Cohen. "Deepfakes, Deep Harms." *Journal of Ethics and Social Philosophy* 22, no. 2 (July 2022): 143–61.

Ronson, Jon. *So You've Been Publicly Shamed*. New York: Riverhead Books, 2015.

Saul, Jennifer M. "Beyond Just Silencing: A Call for Complexity in Discussions of Academic Free Speech." In *Academic Freedom*, edited by Jennifer Lackey, 119–34. New York: Oxford University Press, 2018.

———. "Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump." *Philosophical Topics* 45, no. 2: (Fall 2017) 91–116.

Sherr, Ian, and Erin Carson. "GamerGate to Trump: How Video Game Culture Blew Everything Up." CNET, November 27, 2017. https://www.cnet.com/news/gamergate-donald-trump-american-nazis-how-video-game-culture-blew-everything-up/.

Smith, David Livingstone. "Fighting Hate Is a Losing Battle." *Boston Globe*, August 29, 2017. https://www.bostonglobe.com/opinion/2017/08/29/smith/j5F9Mf4ZqPohu40xC6stTP/story.html.

Stanley, Jason. *How Propaganda Works*. Princeton: Princeton University Press, 2015.

Suler, John. "The Online Disinhibition Effect." *CyberPsychology and Behavior* 7, no. 3 (July 2004): 321–26.

Tirrell, Lynne. "Genocidal Language Games." In Maitra and McGowan, *Speech and Harm*, 174–221.

———. "Toxic Speech: Toward an Epidemiology of Discursive Harm."

Philosophical Topics 45, no. 2 (Fall 2017): 139–61.

Tufekci, Zeynep. "It's the (Democracy-Poisoning) Golden Age of Free Speech." *Wired*, January 16, 2018. https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/.

———. *Twitter and Teargas: The Power and Fragility of Networked Protest*. New Haven: Yale University Press, 2017.

Valenti, Jessica. *Sex Object: A Memoir*. New York: Dey Street Books, 2016.

Waldron, Jeremy. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press, 2014.

Warzel, Charlie. "Mass Shootings Have Become a Sickening Meme." *New York Times*, April 28, 2019. https://www.nytimes.com/2019/04/28/opinion/poway-synagogue-shooting-meme.html.

West, Lindy. *Shrill: Notes from a Loud Woman*. New York: Hatchette Books, 2016.

Witek, Maciej. "How to Establish Authority with Words: Imperative Utterances and Presupposition Accommodation." In *Logic, Methodology and Philosophy of Science at Warsaw University*, edited by Anna Brożek, 145–57. Warsaw: Warsaw University, 2013.