***Notes from the field***

# Pin the tail on the user:

# Locating accountability in decentralized social media

**Christina Dunbar-Hester**, Annenberg School for Communication, University of Southern California, Los Angeles, California, USA. *dunbarhe@usc.edu*

# Pin the tail on the user:
# Locating accountability in decentralized social media

## Abstract

**(English):** This paper explores the potentials and perils of alternative social media, through a firsthand account of targeted harassment on a prominent decentralized social media network, Mastodon. It illustrates how both network architecture and norms place the onus on users for their own safety. Though singular in content, this case conforms to patterns for which minoritized users of the network have sought remedy for years. This matters because abusive behavior online is common and its burden falls heavily on women, racial, ethnic, gender and sexual minorities, and the like; the democratic potential of noncommercial, decentralized social media cannot be realized if enhancing accountability to users is not a priority. The paper argues for foregrounding accountability in the network, spanning sociotechnical relationships between and amongst users, moderators, and architects of the network. It suggests that relations of production and participation on decentralized social media be oriented towards "meshy accountability," invoking both consciously woven connections and the gaps and spaces between them.

**(Spanish):** Este artículo explora el potencial y los riesgos de las redes sociales alternativas a través de un relato directo de acoso selectivo en Mastodon, una prominente red social descentralizada. Ilustra cómo tanto la arquitectura como las normas de la red responsabilizan a los usuarios de su propia seguridad. Si bien su contenido es singular, este caso se ajusta a patrones que los usuarios minoritarios de la red han buscado solución durante años. Esto es importante porque el comportamiento abusivo en línea es común y su carga recae considerablemente sobre mujeres, minorías raciales, étnicas, de género y sexuales, entre otras. El potencial democrático de las redes sociales descentralizadas y no comerciales no se puede materializar si no se prioriza la rendición de cuentas a los usuarios. El artículo aboga por priorizar la rendición de cuentas en la red, abarcando las relaciones sociotécnicas entre usuarios, moderadores y arquitectos de la red. Sugiere que las relaciones de producción y participación en las redes sociales descentralizadas se orienten hacia una "rendición de cuentas mezquina," invocando tanto las conexiones tejidas conscientemente como las brechas y espacios entre ellas.

*Keywords:* alternative social media; free/libre and open source software (FLOSS); networked harassment; Mastodon; fediverse; content moderation; located accountability

## Introduction

In July 2023, I refreshed a social media app and noted with alarm that someone appeared to be trying to instigate a "pile-on" directed at me. Over a few weeks, the episode moved through a predictable repertoire of online harassment: embellished and fabricated accusations; sexist and racist language; "doxxing" me; reporting me at my workplace; an antagonist emailing me at my workplace demanding I "debate him"; and other attempts at public shaming. Conflict of this sort on social media is commonplace, and antagonism is more likely to be targeted towards: women; racial, ethnic, religious, or gender minorities; and especially people embodying intersections of these categories (Bailey, 2021; Marwick, 2021; Banet-Weiser & Miltner, 2016; Beard, 2015). In these regards, absolutely nothing about this episode stands out as unusual, and no new lessons can be derived from it.[1]

What is unique here is not the *content* of this episode but its *context*: an alternative social media network, called Mastodon, which is unique in being *decentralized* and *noncommercial*.[2] It is a FLOSS project with a foundation headquartered in Germany, begun in 2016. Such alternatives hold promise and urgency because their architecture and ownership structure are bulwarks against platform owners' ability to influence elections, amplify or suppress social movements, or even stoke genocides. This paper does not address networked communication's role in these larger social phenomena. Rather, it foregrounds user experience, which matters considerably for the potential of alternative social media networks at scale through wide adoption by a variety of users.

Merely offering *access* to alternative social media is insufficient. It is not empowering for users to adopt alternative networks unless access is accompanied by careful attention to potential burdens for users, especially those which may be distributed unevenly (see Eubanks, 2007). This paper reveals meaningful features of the network, which are not apparent to an outsider to the network, *or even to a user who is not experiencing problems*. Both norms and features on Mastodon contrast in various ways with those familiar from centralized, commercial social media, with attendant consequences for user experience; unfortunately, a user will likely only become familiarized with safety and moderation features in a moment of need.

Distributed harassment on a decentralized network introduces challenges, for both users and analysts. Multiple studies have demonstrated that moderation in a decentralized network operates quite differently (Melder et al. 2025; Zhang et al. 2024; Rozenshtein, 2023; Ermoshina & Musiani 2022; Spencer-Smith & Tomaz, 2025; Bono et al., 2024; see also Gehl & Zulli, 2023). Yet harassing behavior per se has not received much attention. To the extent that it has, researchers have tended to examine what is ostensibly permitted *at the level of servers* (which, linked together, constitute the network) and how different tools such as blocklists

[1] 41 percent of Americans have personally been subject to abusive behavior online, and one in five have been the targets of particularly severe forms such as sexual harassment, stalking, revenge porn, physical threats, and sustained harassment over time (Pew Research quoted in Salehi, 2019).

[2] Mastodon is the flagship of the "fediverse," a "federated universe" of software applications that can "talk" across services running on the Activity Pub protocol.

might manage problematic speech at that scale, *not what actually happens in individuals' interactions with the network* (see Gehl, 2024). Methodologically, there are some good reasons for this (discussed below).

However, given that people undergo targeted harassment and seek its remediation as individuals, attending to this scale in this sociotechnical space has value. When people fled (then) Twitter in droves in late 2022 after Elon Musk acquired it (Nicholson, 2023), technologist, researcher, and artist Caroline Sinders wrote of Mastodon (where she had had a presence much longer than many making the shift):

> My harassers are also on Mastodon; some have multiple accounts. The blocking feature is like horror house anxiety game- I block when I see their new account, hoping I've now blocked all of them but knowing I probably never will. Because it's a federated system, and you can have accounts on multiple servers, it means there's multiple accounts I have to block to create some digital safety and distance. … [Mastodon's] lack of built-in safety tools makes it an impossible place for me to be full time (2022).

Around the same time, former fediverse moderator Ginny McQueen commented: "The [harassing and threatening] scum and villainy are in a lot of cases *even harder to deal with in decentralized spaces* than just a [centralized] silo like Twitter or Facebook," underscoring the toll moderation takes on the volunteers who do it (2022, emphasis added).[3] Users' and moderators' perspectives both indicate that decentralized harassment is a known, and challenging problem. This should be foregrounded in evaluating, and ideally building more convivial decentralized social media (Bonini & Maria Mazzoli, 2022; Gehl, 2015, p. 9).

This paper explores how decentralized harassment is experienced and managed in Mastodon, from the perspective of a user, with glimpses into moderators' perspectives. It argues that "distributed accountability" in Mastodon largely *places a burden on individuals for their own safety in the network*.[4] By "distributed accountability," I mean a tendency to seed accountability within the network in such a diffuse way that responsibility for antisocial behavior becomes disjointed, ephemeral, shift-able onto others, not pinpoint-able. Here I invoke feminist Science & Technology Studies scholar Lucy Suchman's notion of *located accountabilities*, a move to non-universalize relations of production and participation in a sociotechnical environment (2002; see also Dunbar-Hester, 2014b). Locating accountability does not in itself preclude or remediate harm, but it provides a framework to name and ultimately realize alternative technical features and social relations. In noncommercial, decentralized social media, we might think of an alternative to distributed accountability as *meshy accountability*. *Meshy* here refers to both *consciously woven connections* and the *gaps and spaces between them*; it is decentralized without being disjointed (see also Solomon, 2019).

[3] In McQueen's words, "I'm here to tell you that the current landscape of open source social media is a hellscape filled with shitty white men that flows the darkest, most violent corners of the internet" (McQueen, 2022).

[4] A discussion of what "safety" is or might be is out of scope (and word count) for this paper (but see Salehi, 2019).

This argument bears some resemblance to David Gray Widder and Dawn Nafus' notion of "dislocated accountability" in the artificial intelligence "supply chain" (2025). They write, "We were struck by the deeply dislocated sense of accountability, where *acknowledgement of harms was consistent*, but nevertheless another person's job to address, *always elsewhere*" (p. 1, emphasis added). But autonomous social media is more contextually bounded than a quasi-metaphorical AI supply chain; all handoffs of accountability that I trace occur within a people-powered, federated, protocologically-bounded network that is governed locally and endogenously, amidst and across users, server administrators and moderators, network architects, and software. While accountability is distributed, it is not fully dislocated.[5]

This paper illustrates how longstanding norms from free and open source software (FLOSS) culture reveal themselves *in practice* in instances of targeted harassment; and posits that it is beneficial to identify and challenge these norms in order to maximize the convivial and public interest potential of decentralized social media (Bonini & Maria Mazzoli, 2022). Specifically, while users have some degree of control over *what they see* on the network, the network embodies many choices that subtly privilege *others' speech maximalism*. The case presented here is idiosyncratic, approaching frivolous. But the problems I recount index safety issues that minoritized users have been pointing out for years, so in that sense it is not trivial or frivolous at all.

Users have many good reasons to choose reach and visibility. These include hailing members of a community in common, or to participate in movement communication; both of these uses may be particularly important for members of non-dominant social groups (Jackson et al., 2020). As Rianka Singh and Sarah Sharma write, "Platforms are an enduring centerpiece in popular feminist movements. Platforms both mobilize a gaze and direct an audience" (2019, 302). Mastodon does not encode algorithmic visibility; the primary way to establish visibility in the network is through hashtagging (more below).

Accountability to users is important in any social media context. These matters are especially consequential at present, as some argue that dominant companies that shape much of people's experience online are "at risk of losing their relevance" (Dash, 2023). Alternatives to "Big Social" are urgently needed; Twitter/X's more democratic functions fracturing and the platform turning into a breeding ground for networked fascism under Elon Musk's ownership is a prime example. Yet more democratic, noncommercial social media networks cannot flourish nor maximize social good if they cannot be accountable to users, especially users most likely to be targeted for harassment and abuse.

## Research Activities, Methods

[5] My argument also contrasts with Widder and Nafus' in that Mastodon's bespoke networked environment does not privilege technical modularity. What is salient is that bespoke nodes constituting the network tend to converge around dominant FLOSS norms unless they are consciously scripted otherwise.

Though I am a scholar of infrastructure and technical communities, I joined Mastodon as a "regular" user, without any research agenda. In certain ways, such lack of forethought is not ideal. On the other hand, while it would have been unethical to provoke a dispute leading to retaliatory harassment, it was ethical and serendipitous (in a perverse way) to learn about user safety on Mastodon through firsthand experience. Indeed, it may be potentially extractive or traumatic to ask people to revisit and recount instances of harassment they have experienced (especially given the way that harassment falls disproportionately on minoritized users), so having skin in the game, as it were, is a way to place the researcher on similar footing to other users and glean insight into this topic without wading into sensitive territory for others (see Korn, 2017). In addition, this network has its own norms around researchers' presence: "[Mastodon's] structure and culture demands that researchers change their approaches to research ethics when it comes to consent and data collection. The privacy expectations of Mastodon users are quite different [from users of corporate social media]," write Roel R. Abbing & Robert Gehl (2024, p. 1).[6] I obtained consent (after the fact) to include all quotes and screenshots that appear, with the exception of the harasser's (whose right to be approached he waived by harassing me).

Going about my business as an active user of the network, the examples below were easy for me to collect, if unplanned. My data illuminate *the network as experienced by an individual.* It is a *resolutely partial* perspective of/on the network, mediated by individual follow/follower relationships and by instance-level decisions. Though I present isolated moments in a journey through harassment and moderation, I had been conducting an *immersive walkthrough* of the network from my own vantage point in it for several months, becoming even more attuned to Mastodon features as harassment unfolded (Dunbar-Hester, 2024; Light, *et al.*, 2018). When I eventually decided to analyze interaction on Mastodon from a scholarly perspective, I brought to bear both my firsthand experience, in an autoethnographic register, and my research expertise spanning participation in FLOSS projects and decentralized noncommercial media (Dunbar-Hester, 2020; Dunbar-Hester, 2014a). This felt appropriate because both my "personal" and "professional" self/s had been placed at the center of interactions in the network (albeit in ways I did not seek out) (Dunbar-Hester, 2024).

What led me to familiarize myself with moderation on Mastodon was a bizarre dispute with a combative donkeykeeper in Europe. Despite this idiosyncratic point of entry, disruption in social order was especially useful in revealing meaningful norms and features encoded on Mastodon (Garfinkel, 1967; see also Adjepong, 2019). Targeted harassment experienced by a relatively advantaged user in the network can illustrate problems that create a heavier burden for others; and therefore findings from this paper are more generalizable and more significant than my examples (see Melder et al., 2025). They are not, however, generalizable to all decentralized social networks. Rather, they exemplify how architecture choices and norms on Mastodon, many common to FLOSS in general and unique to Mastodon in particular, encode safety and accountability—or lack thereof—in this network.

[6] Markham & Buchanan (2017) provide a useful overview.

## On the internet, nobody knows you're an ass: Norms, rupture

I offer a condensed backstory of events that led to me orienting myself to moderation and safety on Mastodon.

First, though, I sketch two key features of Mastodon's architecture. One, individuals join the network through membership on a particular server, called an *instance*; the environments in which users interact are constituted within and across networked servers (Figure 1). Not all points on the network connect to one another; and there is no central or universal band of connection. Volunteers administer instances, with a division of labor that spans technical administration of the server and moderation duties.[7] Each instance has its own rules for conduct, content, and the like. Second, hashtags are particularly important as they are how the network seeds visibility across servers (for reasons that are beyond the scope of this paper, text-searching is unreliable).

Figure 1. A user-generated illustration of federation, described as "many small boats bobbing in a harbor. Everyone on a given server is in one boat together. Some boats have communication lines with some other boats, so the people on one boat can pass messages to the other connected boats. Those communication lines do not form on their own automatically, but depend on individuals from each 'boat' following people on other boats, and they can be severed by the boat owner at either end of the line" (image and description courtesy of Moss Wizard, November 2024).

Arriving in the October-November 2022 tidal wave of new users prompted by Musk's Twitter purchase, I and a fellow newcomer had while "kicking the tires" of our new

[7] The largest, "flagship" instance, mastodon.social, pays moderators. Spencer-Smith & Tomaz note that on Mastodon, "large generic instances [experience] the highest level of content moderation stress where workload significantly [outstrips] resources" (2025).

environment created a hashtag, “asstodon,” appended to pictures of donkeys (see Dunbar-Hester, 2024).[8] It was a self-consciously silly exercise in cultivating sociality while experimenting with the features of our new environment; a typical post of mine can be seen in Figure 2.


Figure 2. Donkey grazing in a field, post by author, 19 January 2023. #BeTheAlgorithm harkened back to the 1990s-era Indymedia slogan “Be The Media,” referring to a noncommercial, autonomous media environment.

Mirthful donkeyposting continued for several months, until a dispute arose. Rarely, people posted another kind of ass (human bottoms) on the tag, rather than donkeys. When a certain donkeykeeper spied these, he (unbeknownst to others) reached out to request posters edit out the hashtag so as not to “pollute” the donkey feed.[9] In July 2023, someone posted erotic furry art on #asstodon, and did not immediately respond private outreach from the donkeykeeper. So he blasted this poster, making a public request on the hashtag for others to call out this user, and their instance administrators as well. His “moral outrage” justified encouraging others to shame the furry-art poster (Marwick, 2021).

I was interested in repairing the rupture in our shared social space, but not in participating in public shaming. I commented on the hashtag around which we were assembled (see Jackson et al., 2020). My post winked tamely at the controversy, including a donkey looking back at the camera over its, well, ass (Figure 3). A couple of days later, I also mused

[8] “Asstodon” hewed to a Mastodon norm to create community hashtags punning on “Mastodon,” like “Mosstodon” for photos of moss; and “Monsterdon” for synchronously watching monster B-movies.
[9] Mastodon allows edits; an edit log shows original post and changes made.

about avoiding ambiguity by perhaps splitting the hashtag into two streams--a proposal the donkeykeeper rejected (Figure 5).


Figure 3. Author's post, 15 July 2023. (#Commodon hailed communication and media scholars on Mastodon. At this point, I didn't plan to write anything, but thought others might find the situation of literal academic interest.)

Other people weighed in too. One of my "mutuals" also disagreed with my proposal to split the tag: "The hashtag itself is sort of a joke… humor is fine but the controversy if there is any seems to verge on sex-not-positivism: the people who want to post human asses have just as good a reason to do so and it's not clear to me that it's very bad if the tags get mixed up. Maybe people can mute anything tagged #nsfw [not safe for work] if they are concerned" (Rich Puchalsky, 17 July 2023). He reminded people that usually on Mastodon, one was free to use one's account settings to filter content one did not wish to see, but one could not control others' posting.

The donkeykeeper regarded this "meta" discussion as further abuse of the hashtag, and blamed me for stoking discussion rather than keeping quiet or endorsing his puritanical pile-on of the furry-art poster. A few days later, I opened my app and found myself tagged by two unfamiliar accounts, in the guises of a donkey, and a carrot(!). Both politely warned me to step away from the hashtag; the "donkey" falsely stated that donkeykeepers had started the hashtag. Both posts included #asstodon, i.e. they hailed the hashtag community, attempting to incite "morally motivated networked harassment" (Marwick, 2021). After a moment of disbelief, I had a strong suspicion of who was behind the posts. I told him to knock it off, and "blocked" his main account, the carrot, and the donkey. I did not report him to moderators, because I (wrongly) imagined that the disagreement would blow over.

The donkeykeeper did not tolerate my pushback. He made a new sockpuppet account on a large server with a reputation for poor moderation (here called Friendly); there, in the persona of a Spanish fascist, he wrote a racist, sexist, and anti-semitic thread that he presented as a "satire" of the situation. His donkey account referred to me by my full legal name, which I did not use on Mastodon (thereby "doxxing" me). Both the donkey and Spanish fascist tagged posts "Commodon," a hashtag used by communication and media scholars for sharing papers or asking research questions, i.e. he wanted my professional community to witness his attacks (see Marwick, 2021, 2). This behavior crossed very bright lines, so people including myself began reporting his accounts; some observers also addressed the asstodon hashtag, urging its community to report his behavior: "If #Asstodon wants to tolerate … abusive hassling of women and others, it will just be known as a hashtag for assholes, not butts or donkeys" (Yehuda, 23 July 2023).

A few days later, the donkeykeeper tagged me again. I had him blocked and did not receive a notification, but someone alerted me. Since initiating contact after being blocked is rulebreaking, I screenshotted that post, posted it on my account, and again asked people to report him. An academic colleague, who'd had nothing to do with any of the dispute so far (but had probably seen the donkeykeeper's harassing posts on the "Commodon" hashtag), saw my post, and replied directly to the donkeykeeper to say "you are extremly [*sic*] rude" (1 August 2023).

At this provocation, the irate donkeykeeper tipped his hand: his vendetta had breached Mastodon. He replied with a screenshot of an email from administrators at the university where I work, logging a complaint he'd apparently lodged about *me* harassing *him* (Figure 4). I became for the first time genuinely alarmed, and found myself relieved that the donkeykeeper lived on another continent.


Figure 4. Donkeykeeper posts a screenshot of correspondence from my employer; his post includes both my surname and employer. 1 August 2023.

I contacted university administrators to explain. Within a couple of days the donkeykeeper essentially confirmed my version of events by emailing me at work demanding I "debate him"; and emailing administrators a second time to berate them for taking too long to address his complaints. Two weeks later, he wrote a public blog post venting his spleen and attempting to shame me (a pattern, one now notes, beginning with the furry-art poster). I learned of this because he posted a link on the asstodon tag, inviting another round of being told off.

This arc played out over a few weeks. As the donkeykeeper's harassment intensified, I shifted towards to trying to neutralize his attacks. Next I turn to moderation on Mastodon; and to features related to reach and access, with implications for accountability.

## Pinning the tail on anyone at all: Seeking accountability in a federated network

After being doxxed, I received a bewildering crash course in how moderation worked (or, more accurately, did not really work) in this decentralized network. Though I'd initially shrugged at his antagonism, as he escalated attacks on me and involved my employer, I became more motivated to try to ward him off. This was not straightforward; groping and stumbling, I felt a bit as though I'd been blindfolded and spun around.

Given the decentralized nature of the network, each server has local moderators, responsive (in theory) to their own instance members and to users on other instances who have problems with "home" instance members. Many servers now participate in a "server

covenant" with a core code of conduct they commit to enforcing (including silencing instances where hate speech and "free speech absolutism" flourish) (see Gehl & Zulli, 2022); yet in practice, there is variability across servers in terms of both abstract rules for enforcement and what will, in practice, result in a user being moderated. Moderator actions can range from a warning to the removal of a post to asking the user to depart a server to deactivating the user's account. (Moderators also can take action at the level of servers with tools like blocklists, as noted above.)

Within Mastodon's interface, there is a tool for reporting rulebreaking, which will call a post to moderators' attention. A reporter selects whether they are reporting only to their "home" instance, where they see the post, or also to the poster's server, where it originated. In theory, two sets of moderators will check up on a complaint (unless the users are on the same instance). Because the network runs on instance-to-instance connection, there is no omniscient perspective in it (refer again to Figure 1); and there is no moderation high court, as local servers abide by and enforce local rules. When I reached out to my home instance moderators after the first "doxxing" incident, they helpfully acted immediately to silence all of the donkeykeeper's accounts locally on our server, so no one on our instance would see his posts. Witnesses to his behavior blocked and reported him on their instances as well. This meant that many people, and some whole instances of varying sizes (mine had a few hundred users) would not see posts from the donkeykeeper's main or subsidiary accounts.

But crucially, the donkeykeeper's main account and two of his "alts" were hosted on very large instances (here referred to as Galaxy and Friendly, each with tens of thousands of users). This meant he was still free to roam around the network and address large audiences to spin whatever narratives he liked. Scattershot silencing on a handful of smaller servers seemed an insufficient response when he used the largest servers as home base and for alts, and persisted in doxxing me; I did not trust that his vendetta was necessarily dying down. I chose to directly email moderators for the servers that hosted his accounts, i.e. to not rely only on the reporting tool in the interface.

In an exchange I had with moderators on the donkeykeeper's initial main account's server, the large Galaxy instance, a moderator underscored that they would only act on posts on Galaxy: "We only moderate on posts on our instance, this is one of the downsides of federation. Or positive side actually as no person or organisation has full control" (email, 27 July 2023). In other words, no one had the power to take this user out of commission if he spread his behavior across many instances. The moderator framed this as a positive aspect of decentralization.

I reminded them that the "server covenant" Galaxy pledged to uphold forbade harassment, doxxing, and bullying, and if Galaxy *could see such behavior by their user*, even on other servers, it was potentially within their ambit to mete out consequences for the main account (with a few hundred followers) hosted on their server. I also pointed out that it defeated the point of a code of conduct to not enforce it (see Dunbar-Hester, 2020, chapter 3). My ability to document the rulebreaking across servers was unusually airtight: the

donkeykeeper's main account outright admitted that the "donkey" account (which had doxxed me, hosted on Friendly) was actually him.

I do not know what happened after my conversation with the Galaxy moderator, i.e. whether Galaxy moderators had a talk with the donkeykeeper and possibly invited him to depart Galaxy. But shortly thereafter he moved his account from Galaxy to a small European server, taking with him his few hundred followers. If Galaxy moderators had something to do with this move, that meant they ultimately took moderation more seriously than I initially sussed. I next emailed his new server's administrator to give an account of their new user's recent behavior; I recommended believing reports about him if any came in. I received no reply, but it seems this moderator took my warning seriously, as within a couple of days, the donkeykeeper migrated accounts yet again (still retaining his followers). I don't know if he did something to warrant swift enforcement; or if the new instance told him they had an eye on him. Next, the donkeykeeper's main account popped up on Friendly; a server even larger than Galaxy, with a reputation for insufficient moderation. From this new account, he tagged me, prompting my "Commodon" colleague to accuse him of rudeness; upon which he again doxxed me and displayed the screenshot of my employer's email.

Galaxy was remarkably responsive and transparent about their decisionmaking. Other moderators were harder to raise. Friendly, the very large instance that hosted the donkey, the satire author, and to which the donkeykeeper ultimately migrated, never got back to me affirming they'd reviewed the racist satire. Far worse, they never acknowledged my alert that their new user who'd departed Galaxy under suspicious circumstances *immediately doxxed me from Friendly*, and was *harassing me at work*. Friendly pledges to abide by the server covenant, but its user base is voluminous; in practice, its moderators are spread thin. (More triflingly, the small instance that hosted the "carrot" initially got back to me to say they'd have a look, and then went dark.) My experience demonstrates how moderation in this federated network places a perhaps unexpectedly large burden on the victim of harassment, as Sinders warned.

Some of this is an effect of Mastodon's decentralized architecture. Unlike a centralized platform, a moderator on a particular instance cannot view a deleted post created on another instance; they would not have direct access to that server's history; and they are (understandably) unlikely to take the time to coordinate, as volunteers, with the administrators on another server (also volunteers). (The tools they have in fact inhibit coordination; more below.) This leaves two ways to document harassing posts: screenshotting (which I, my allies, and my harasser all did); and creating links to archive posts that remain viewable on the web even if they are deleted (which I did not think to do, not having experience with being harassed online, let alone harassed on a decentralized network). Novices on the network lack the foresight and tacit knowledge to archive links. This resulted in Galaxy's moderator telling me, in response to screenshots I shared, that as screenshots can be altered, moderators tend to use them only "for context." It is true that it is relatively easy to manipulate a screenshot. This also meant that it was laughably easy for my harasser to evade appropriate disciplinary action, even when I presented (undoctored) screenshots of abusive behavior. My screenshots were *not* the only readily locatable evidence of his harassment. Yet even with a harasser *admitting in plain*

*view* to harassing me from multiple accounts, and leaving up doxxing posts for days on end, his home moderators were reluctant to act, because harassing posts were on other servers; and because some of his most offensive behavior was "only" up for a number of days, before it was eventually flagged by moderators or deleted by him (no way for me to know). This would leave a very wide opening for a more motivated or more sophisticated actor to harass quite broadly and perpetually (more on this below).

It was particularly troubling that doxxing, first by a sockpuppet hosted by Friendly, visible to Galaxy; and again by the donkeykeeper's main account, on Friendly after he'd moved there, never resulted in the donkeykeeper's main account being suspended by either instance. This could have been more consequential for me if more people had been persuaded to harass me on or off the network, which, one must imagine, was his intent given his repeatedly displaying my full name (and my employer). Even moderators willing to take some action against him limited those actions to deleting offending posts and seemingly inviting the donkeykeeper to move his account (with intact follow-follower relationships); so when he departed a server, he easily reestablished his intact account elsewhere. (Of course he could set up infinite new accounts even if his main one was deactivated; but it is inarguably more work to recreate a few hundred followers.[10])

Moderation tools also distribute, or perhaps sever, accountability amongst moderators. Moderators explained that the software tools do not facilitate moderators communicating across instances. When a user files a complaint, the reporting tool sends their complaint to multiple servers, but then each server is left to adjudicate separately and alone (Personal conversation, moderator, 23 July 2023; 11 September 2024). Even if moderators might choose to communicate with moderators on other instances, the moderation tool features inhibit collaboration. Yet one can imagine that if it were not burdensome for them to do so, moderators might in some cases choose to share information across instances.

This episode illustrates how the distributed nature of Mastodon creates opportunity for harassers. Users can establish accounts and seed harassing behavior across multiple servers, and no single instance moderator will have a full perspective on bad behavior, nor be motivated or necessarily even able to seek it. Unlike cases of "networked harassment" (what the donkeykeeper tried and failed to stoke, with a multitude of discrete users piling onto one person), this case of "dyadic" harassment should have been *particularly easy to adjudicate*, especially given that the donkeykeeper left breadcrumbs linking his accounts (Marwick, 2021, 10). Yet one large instance (Galaxy) felt their hands were tied if his most offensive behavior was on other servers (and it was: the donkeykeeper had some idea how to keep his main account an arm's length from his worst rulebreaking); and another large instance (Friendly) *was completely*

[10] In 2025, the donkeykeeper located my account on the Bluesky platform! I blocked him, but he began spamming user accounts with connections to my university, in an attempt to shame and defame me. In a now-familiar ritual, I enlisted mutuals to report him. Within a couple of days, moderators deactivated his account. Bluesky protocol and architecture is well beyond the scope of this paper, but in this case it functioned more like a centralized (and responsive) platform.

*unaccountable to racism and doxxing emanating from its user, and to off-network harassment clearly and visibly linked to its user*. (I never received any follow-up beyond an auto-reply from Friendly, in spite of having provided them with screenshots and a narrative account, and forwarding the harasser's emails sent to my workplace, which themselves contained Mastodon screenshots. This instance's reputation for disgraceful moderation is surely deserved.)

## "The reply button was a mistake": Reach, Visibility, Access

This essay illustrates that at core, many choices on Mastodon promote speech maximalism over user safety.[11] In other words, there is more going on here than just decentralized architecture. Despite the network being avowedly decentralized (so much so that Zulli and Gehl argue it should be characterized as *noncentralized*, 2020), and not all parts of the network connecting, *the network prizes reach, coupled with access, over consent*. Put differently, users whose posts can be *seen* by others cannot additionally or distinctly *consent* to interactions; if a post can be seen by other accounts, the network grants them access to interact and reply. Melder et al. observe this dynamic as well and explain it as "federation as a protocol encodes a negative form of consent, allowing users to interact by default" (2025, p. 021:1). This differs from Twitter/X and Bluesky, which allow posters to limit replies to followers, to people mentioned in a post, or to no one at all; which may be a preferable setting for a post about a potentially sensitive or polarizing topic.

A couple of additional examples show why this matters. Though the bulk of this paper concerns my dealings with a splenetic donkeykeeper, I gained additional insight when I occasionally posted hashtagged news stories about a particular topic. These posts invariably received combative replies from several accounts whose sole purpose was combative engagement on this topic. A "mutual" who was also being pestered by these accounts pointed out to me patterns in user handles and other similarities, indicating they were probably all the same person. Soon, they moved from replying to my hashtagged posts—i.e. the ones most visible to the network—to my posts on this topic that were not hashtagged. This indicated that they had begun monitoring *my account*, not only hashtags. I would block them, but before long, a new account would find me and reply. I wished mightily for the option to limit who could reply to publicly viewable posts—but I had no way to do so. I could block the individual accounts, and did, but this was fairly futile; another would soon appear. Many times when I posted a story or comment on this topic, they posted disagreeable (and once, threatening) replies likely to be seen by others reading my post. The only way to avoid receiving hostile replies was to not post in public view about this topic--which was, of course, their point. (I also witnessed them intimidate another user in exactly this manner.)

[11] Certain non-intuitive features are meant to *enhance* safety. Text search being inhibited was intended to keep would-be targeted harassers from term-searching. Mastodon also did not replicate "quote-boost," used in Twitter/X and Bluesky to call one's own audience's attention to another post (Bastian, 2023). Both of these choices were direct reactions to experiences people had had on Twitter.

It seemed as if one, in choosing to be *visible* in the network, was also consenting to potential harassment; visibility was a trap even in a decentralized network. Certainly, the problem of a hostile user setting up multiple throwaway accounts to bother people is not unique to Mastodon. But Mastodon's features that privilege *access to other users* over *users' consent to being contacted* intensify this problem. "Participants repeatedly emphasized that not all interaction was desirable or consentful," write Melder et al. in a study of "blocklists" as moderation tools (2025, p. 021:14). These patterns affect both individuals and collectivities: e.g., as Singh and Sharma note, "platform feminism straightens and whitens the movements because they privilege rising up [legibly, visibly]" (2019, p. 303), which not everyone can or will do.

Underscoring this is another feature, quite peculiar to Mastodon. As noted above, hashtagged posts set to the "public" setting are meant to be viewable across the network. There is another setting that makes a post *visible only to one's own followers*. Intended as a safety feature, this prevents posts from being boosted, where they might be read by non-followers or read out of context. In retrospect, I realized the donkeykeeper tried to intimidate me privately before he took to bullying me publicly. In poring over screenshots of our interactions for moderation purposes and eventually academic writing, I happened to notice that in a couple of our initial exchanges early in the dispute, he had used the private-reply setting to my public posts. In Figure 5, his reply switches the setting of my *public* post (the small globe icon) to the setting where only *his followers and me* can see his post (the small lock icon). I hadn't noticed the setting-change in real time, but it restricted visibility and replies from everyone but his followers, who were more likely to interact with him, to trust him, etc. than me. Using this setting curtailed public discussion; and it explicitly shut out *my* followers, who might be sympathetic towards me. (Though, above, my "mutual" disagreed with me! This was fine with me, as I was stoking a conversation, not a war.)

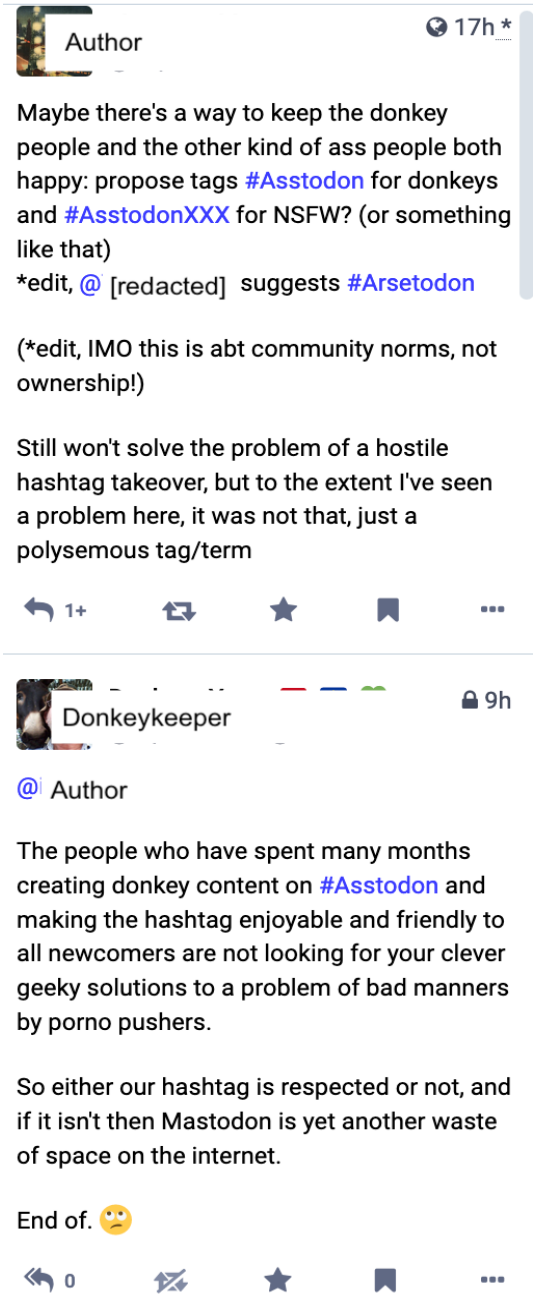


Figure 5. My post and the donkeykeeper's reply, in differently-viewable settings. 17 July 2023.

The content here is admittedly absurd. But implications are more dire. People using Mastodon for racist harassment have had an absolute field day with this feature, replying to Black users in particular with slurs *that are not publicly visible,* thus isolating victims from witnesses and allies (Pincus, 2024). While this setting is apparently intuitively clear to harassers, it is opaque to novice users; my own initial failure to notice that the donkeykeeper replied using

a setting where only his followers could engage was typical. (I much later wondered later if the donkeykeeper had spun lies or inducements to harass to his followers using this private setting.)

## Locating Users, Moderators, and Accountability

Above, I stated that "the distributed nature of Mastodon creates opportunity for harassers," but this is an oversimplification. The architecture of the network creates opportunity for motivated harassers, but problems I experienced with harassment on Mastodon were not only due to software. Equally important was some moderators' reflexive reluctance to discipline a user for bad behavior, perhaps especially so for behavior for which he had constructed the *flimsiest bit of plausible deniability*. His free speech just beyond the moderators' gaze was held to be more important than my freedom from harassment, in line with longstanding libertarian FLOSS mores (Reagle, 2013; Dunbar-Hester, 2020).

In a sense this was foreshadowed by an interlocutor's comment about Asstodon, "the people who want to post human asses have just as good a reason to do so and it's not clear to me that it's very bad if the tags get mixed up. Maybe people can mute anything tagged #nsfw if they are concerned." This reflects a mainstream viewpoint in FLOSS, where liberal individualism has reigned historically (Coleman, 2013). People are free to post as they choose; others can take steps to shield themselves from seeing, but they cannot intervene in how others post. This is actually how Mastodon handles far-right speech, which flourishes in some parts of the network, but from which many instances choose to entirely "defederate" (Gehl & Zulli, 2022, 12-13). Users are assumed to take responsibility for their own choices about what to view. This norm has much to recommend it for non-harassing and non-instigatory speech; in the context in which the poster raised it, it is entirely defensible.

Yet the norm that users are responsible for what they see has obvious pitfalls. The idea that "users are responsible" tends to universalize users, papering over how not all users are equally powerful, or equally vulnerable. Taken further, it potentially blames victims of harassment for their predicaments. FLOSS' tendency towards liberal individualism is strongly in tension with the fact that some users are more likely to be targets for abuse. Historically, "the ideas of freedom and openness can be used to dismiss concerns [of minoritized groups] and rationalize [the constitution of relatively monolithic FLOSS communities]" (Reagle, 2013).[12] Some FLOSS communities have agitated for "free speech" and "anti-harassment" to be understood as *compatible values*, not at odds; the emergence of codes of conducts and server covenants over the past decade or so reflects this discursive shift (see Dunbar-Hester, 2020, p.

[12] Early agitation about inclusion in FLOSS centered around gender and largely ignored race, as Reagle's account reflects. Unexamined whiteness is conspicuous in Mastodon (or as Liza Sabater, blogdiva@mastodon.social, once memorably called it, "Whitestodon," 16 September 2023) (see Hendrix & Flowers, 2022; Kiam & X, 2023; Dunbar-Hester, 2020, chapter 7).

86-87). However, though uniformly committed to being an alternative to corporate social media, Mastodon does not seek to harmonize irreconcilable positions on this spectrum, "solving" incompatibilities through decentralization; "free speech absolutism" reigns in parts of the network (Gehl & Zulli, 2022; Pincus, 2017).

Users being responsible for what they themselves see, but powerless regarding others' posting is highlighted in how "blocking" on Mastodon functioned its earliest days. In the words of Marcia X, an artist and writer from Chicago, and a moderator for the first dedicated instance for Black and people of color Mastodon users (which does not exist today): "if I blocked a user, it would block their content from being visible to me, [but] I was used to my content being blocked from being visible to them. This was never made clear as a user or moderator, and this is how we found out how 'blocked' users were able to squat on our timelines and mine our content" (Kiam & X, 2023). Though blocking as *currently* instantiated is supposed to limit the "blockee" from seeing, tagging, or replying to the user who has blocked them, the earlier architecture exemplifies how network architects left large holes in measures a user might take to revoke consent to others' access to them. The present architecture, which puts a user at risk of undesired contact from other users or even from one person's sockpuppets seeded across different instances, no coordination between instances, and no accountability at all from the largest instance (where newcomers are encouraged to onboard!) represents a fairly marginal improvement for user safety, and very little shift in the underlying norm. (Marcia X also innovated a user-safety practice, the "fediblock" hashtag, where users report abuse on that hashtag; this social practice exploiting the network-ability of hashtags is a meshy way to "hack" cross-instance coordination at the user and possibly moderator/administrator level [Kiam & X, 2023].)

On Mastodon, *as users become visible*, they may become subject to unwanted interaction, and even vulnerable to abusive interaction. Then, they find themselves betwixt and between, as moderators of instances will rarely if ever coordinate. Even if they did, "In decentralized networks, communications between instances whose admins have different norms about the definitions of and appropriate reactions to harassment add a level of complexity," writes Jon Pincus (2017). This onus falls hard, and immediately, on visible users who are in least-advantaged or multiply marginalized positions in the matrix of domination, like Black women (Collins, 1990; see Bailey 2021, 66; Francisco & Felmlee, 2022).

This sheds light on the Galaxy moderator's initial reluctance to discipline a Galaxy user who could be plainly seen harassing another user, but not from Galaxy's instance. Their statement, "This is one of the downsides of federation. Or positive side actually as no person or organisation has full control," subtly reveals that in terms of both norms and architecture, *users never really have control over who has access to them*, including people with evidently hostile intentions. Crucially, Galaxy's moderator presented this to me *as a property of the network,* downplaying their instance's ability to make choices. (I do not intend to single out Galaxy's moderator as a bad actor by any means; this viewpoint is representative of many on Mastodon, perhaps especially those in moderation positions unless they have made conscious choices otherwise. Furthermore, Galaxy's moderator was responsive to me, patient, and kind

interpersonally; taking time to enter into an email exchange, they were the most responsive moderator I have encountered in two and a half years on Mastodon, besides my "home" moderators. And community moderation is hard, often invisibilized work [Matias, 2019; Gilbert, 2023].)

"No person or organization has full control" is literally true--and in many ways politically preferable, as the moderator argued. But this move to *distribute* accountability can also *disavow* accountability. "Acknowledging and accepting the limited power of any actors or artifacts to control technology production/use" gestures towards accepting and sitting with responsibility over discrete parts of a system (Suchman, 2002, p. 101). By contrast, reminding a user that there is no centralized control may subtly shift a burden onto users. This belies moderators' empowered status: "Community moderators occupy a contentious role at the community level of power. On the one hand, they work within a system more powerful than themselves. On the other hand, they occupy positions of power relative to users," writes Sarah Gilbert (2023, p. 111:24).

Notably, the moderation problems I have chronicled have been pointed out by minoritized users of Mastodon *for years* (see Kiam & X, 2023; Pincus, 2022, 2023). And, as a tenured North American professor who is white, a native English speaker, and not a gender, sexual, or religious minority, I am not the most vulnerable user on this network (or other platforms) by any means.[13] Nevertheless, harassment I experienced on the network (and spilling off it) was not entirely trivial, and while parts of the network were responsive and acted admirably, the overall picture is not rosy. I had never been doxxed, reported at work, or stalked by motivated harassers watching my timeline and waiting to pounce before spending time on Mastodon; this may be coincidence, or there may be something special about Mastodon. At the network level, the network does a poor job guaranteeing anything like safety, which may embolden harassers. Rather, the network guarantees its opposite: *access to one's account from potentially hostile users is assured* if one posts in public view of the network (rather than only posting in the more-private followers-only mode, e.g.).[14] Reach and access are tightly coupled, and (admittedly peculiar forms of) visibility and reach are default settings.[15]

As currently configured, Mastodon values noncentralization, but it stops short of rewriting power relations to encode heterogeneity (Suchman, 2002). FLOSS universalism and FLOSS relations of domination pervade both the user base and many (certainly not all) "power users" who assume control of and take responsibility for many of the parts of the network through moderation and administration roles. This includes a commitment to liberal individualism, with hacker characteristics: "self-determined and rational individuals who use their well-developed faculties of discrimination and perception to understand the 'formal'

[13] See Banet-Weiser & Higgins (2023) on "economy of believability," which is broadly relevant here.
[14] Some instances use software that allows users to post to a feed that is *local to their instance*, meaning only instance-mates can see and interact with posts, but this is not commonly adopted.
[15] Thanks to Beadsland for discussion. (I say "admittedly peculiar" because of how federation makes for sometimes disjointed connection.)

world— technical or not—around them with such perspicuity that they can intervene virtuously within this logical system either for the sake of play, pedagogy, or technological innovation," as anthropologist Gabriella Coleman writes (Coleman, 2012, p. 7). In FLOSS in particular, this ethos has *not* historically been fertile ground for acknowledging let alone confronting structural problems like racism.

Moderators having more power than users may seem like a glaringly obvious revelation, but it is worth staying with the implications. To locate responsibility in the network would mean to acknowledge power relations, not only between users and moderators; but between users, moderators, and network architects; and users and other users (Matias, 2019; Gilbert, 2023; Hasinoff & Schneider, 2022). Sarah Gilbert provides an account of a subreddit community whose moderators regularly negotiate power self-consciously, "bring[ing] awareness to how moderation reinforces existing power structures at [individual, community, and systemic] level[s] so that moderation policies and practices can subverted from reinforcing oppression to supporting resistance" (Gilbert, 2023, p. 111:6; see also Gray, 2020).

Galaxy's decision to not intervene when their account-holder harassed me from another instance on the network was not an off-the-cuff decision, but in fact their considered policy. And to reiterate, this is a choice, *not* a property of the network. Despite disagreeing with it, I appreciated their explanation: "We have had several discussions in our moderation team about reports of users that have misbehaved on other instance or other social media… [as] it is difficult to prove a link between these accounts, we [do not] … suspend accounts over [them] or [take] similar strong actions," the moderator wrote (email, 29 July 2023).

Galaxy's considered stance stands in contrast to the stated code of conduct of a newer, vastly tinier instance: "the admins reserve the right to exclude people from [our instance] based on their past behavior, *including behavior outside [our instance] spaces and behavior towards people who are not in [our instance]*" (assemblag.es, n.d., emphasis added). This server's administrators do not view events outside its server as irrelevant to how they moderate; this network node enacts *meshy accountability* between and amongst the server, its administrator, users, and other instances. This unsettles familiar patterns of domination that pervade sociality in technical cultures and spaces, undercutting plausible deniability on which bad actors may lean. It also serves as a reminder that instance-level choices, and norms, are malleable and not determined; and for this reason, governance in this network can *potentially* be "much more focused on local norms and community needs than any large-scale centralized platform" (Kissane & Kazemi, 2024; Melder et al. 2025).

To cultivate meshy accountability would also implicate software developers who produce, update, and maintain Mastodon. Currently, another way in which accountability is distributed is that there are no direct consequences for the network architects when users encounter the same pattern of harassment over and over, to the point of driving many new users away, especially minoritized users. They too could bring greater awareness to how architecture and tooling reinforce existing power structures, and build to support resisting

rather than reinforcing oppression based on feedback from groups most affected by harassing behavior of the sort I have experienced. For me it has merely been aggravating, fortunately.

## Conclusion: From distributed accountability to meshy accountability

Mastodon as it currently exists grants users access to a noncentralized network, but it stops short of self-consciously implementing features and practices attuned to equitable power relations at the level of the network. However, distributed accountability as illustrated here is not an *inherent* characteristic of a decentralized network: the defining feature of a noncentralized network is that not everyone is or has to be connected to everyone. Just as rightwing servers have their own mostly-not-connected universe, a possibility for a more accountable decentralized social media network might *resemble* Mastodon, but prioritize intentional, self-organized choices about *equitable* online sociality that foreground social power (Eubanks, 2007). The idea is not to "control" how other users post, nor to disavow responsibility for remediating harm that interactions can cause; it is to cultivate *meshy accountability* in sociotechnical space such that there are preclusions and remedies for harm cultivated across the network's architects, moderators, users, and the network itself.

In a media and information environment dominated by commercial behemoths (both legacy media and platforms), with radical rightwing politics ascendant, it is demonstrably evident that people need something *like* Mastodon. Noncommercial alternative social media with public interest aspirations is one crucial aspect of repairing a dysfunctional media landscape; people deserve to "maintain strong social connections online while escaping the behavioral manipulation, pervasive surveillance, and capricious governance that characterizes large-scale centralized social platforms" (Kissane & Kazemi, 2024). But people urgently need something better than Mastodon as it currently exists.[16] For me personally, harassment on Mastodon has been merely annoying. Yet it also hinders my experience on Mastodon more expansively, as it indicates limits regarding who can comfortably use the network, and to whom I can in good conscience invite or recommend to join Mastodon. A sizeable missed opportunity of Mastodon lies here.

## Acknowledgements

Though this paper details an individual's experience, I didn't go it alone: I am grateful to have been supported by various "mutuals" and hashtag community members. Thanks also to Josh Braun and Jon Pincus for feedback on early and later drafts; to various decentralized

[16] Beyond the scope of this paper, there are: small Mastodon instances that admirably commit to curating good experiences for minoritized users; emerging interoperable ActivityPub alternatives, like GoToSocial; and interoperability with or bridging to other protocols like Bluesky's AT Protocol (itself a base for BlackSky, designed with community safety in mind).

interlocutors including moderators who both in real time and after the fact offered insight into reach, access, and moderation practices; and to Colin Rhinesmith for editorial feedback. Lastly, I include an (unsolicited) AI disclaimer: no so-called AI was consulted in process of producing this paper; mistakes and omissions are human.

## References

Abbing, R.R. & R.W. Gehl (2024, January 12). Shifting your research from X to Mastodon? Here's what you need to know, *Patterns* 5. https://doi.org/10.1016/j.patter.2023.100914, accessed 12 August 2024.

Adjepong, A. (2019). Invading ethnography: A queer of color reflexive practice. *Ethnography*, *20*(1), 27-46.

Assemblag.es. About. https://assemblag.es/about, n.d., accessed 11 November 2024.Bailey, M. (2021). *Misogynoir transformed: Black women's digital resistance*. NYU Press.

Banet-Weiser, S., & Higgins, K. C. (2023). *Believability: Sexual violence, media, and the politics of doubt*. John Wiley & Sons.

Banet-Weiser, S., & Miltner, K. M. (2016). # MasculinitySoFragile: Culture, structure, and networked misogyny. *Feminist media studies*, *16*(1), 171–174.

Bastian, H. (2023). Quote tweeting: Over 30 studies dispel some myths, *Absolutely Maybe (PLOS Blogs)* (January 12), at https://absolutelymaybe.plos.org/2023/01/12/quote-tweeting-over-30-studies-dispel-some-myths/, accessed 12 August 2024.

Beard, M. (2015). The public voice of women. *Women's History Review*, *24*(5), 809–818.

Bonini, T., & Maria Mazzoli, E. (2022). A convivial-agonistic framework to theorise public service media platforms and their governing systems. *New media & society*, *24*(4), 922–941.

Bono, C. A., La Cava, L., Luceri, L., & Pierri, F. (2024, May). An exploration of decentralized moderation on Mastodon. In *Proceedings of the 16th ACM Web Science Conference* (pp. 53–58). https://doi.org/10.1145/3614419.3644016

Coleman, E. G. (2013). *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press.

Collins, P. H. (1990). Black feminist thought in the matrix of domination. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*, *138*(1990), 221–238.

Dash, A. (2023, 30 December). The Internet is about to get weird again. *Rolling Stone.* https://www.rollingstone.com/culture/culture-commentary/internet-future-about-to-get-weird-1234938403/, accessed 15 February 2024.

Dunbar-Hester, C. (2014a). *Low power to the people: Pirates, protest, and politics in FM radio activism*. MIT Press.

Dunbar-Hester, C. (2014b). Radical inclusion? Locating accountability in technical DIY. In M. Boler & M. Ratto (Eds.), *DIY citizenship.* MIT Press (pp. 75–88).

Dunbar-Hester, C. (2020). *Hacking diversity: The politics of inclusion in open technology cultures*. Princeton University Press.

Dunbar-Hester, C. (2024). Showing your ass on Mastodon: Lossy distribution, hashtag activism, and public scrutiny on federated, feral social media. *First Monday*.

Eubanks, V. E. (2007). Trapped in the digital divide: The distributive paradigm in community informatics. *The Journal of Community Informatics*, *3*(2).

Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice Hall.

Ermoshina, K., & Musiani, F. (2022, November). Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation. In *Annual Symposium of the Global Internet Governance Academic Network (GigaNet)*. Addis Abbaba. https://hal.science/hal-03930548

Francisco, S. C., & Felmlee, D. H. (2022). What did you call me? An analysis of online harassment towards black and latinx women. *Race and social problems*, *14*(1), 1–13.

Gehl, R. W. (2015). The case for alternative social media. *Social Media+ Society*, *1*(2), 2056305115604338.

Gehl, R. W., & Zulli, D. (2022). The digital covenant: non-centralized platform governance on the mastodon social network. *Information, Communication & Society*, *26*(16), 3275–3291.

Gehl, R. (2024, March 14). Researching the fediverse: instances and individuals. *FOSS Academic,* https://fossacademic.tech/2024/03/14/Instances-and-Individuals.html, accessed 14 November 2024.

Gray, K.L. (2020). *Intersectional tech: Black users in digital gaming*. LSU Press.

Hasinoff, A. A., & Schneider, N. (2022). From scalability to subsidiarity in addressing online harm. *Social Media+ Society*, *8*(3), 20563051221126041.

Hendrix, J. & J. Flowers (2022). “The whiteness of Mastodon.” Tech Policy Press. https://techpolicy.press/the-whiteness-of-mastodon , accessed 11 December 2023.

Jackson, S.J., M. Bailey, & B.F. Welles (2020). *#HashtagActivism: Networks of race and gender justice*. MIT Press.

Kiam, R.I. & Marcia X (2023, 13 December). "Blackness in the fediverse: A conversation with Marcia X," *Logic(s)*, volume 20. https://logicmag.io/policy/blackness-in-the-fediverse-a-conversation-with-marcia-x/, accessed 15 February 2024.

Kissane, E. & D. Kazemi (2024, 20 August). Findings Report: Governance on Fediverse Microblogging Servers. https://fediverse-governance.github.io/, accessed 11 November 2024.

Kissane, E. (2024). Safer places, now. https://www.wrecka.ge/safer-places-now/ , accessed 11 November 2024.

Korn, J. U. (2017). Expecting penises in Chatroulette: Race, gender, and sexuality in anonymous online spaces. Popular Communication, 15:2, 95–109.

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New media & society*, *20*(3), 881–900.

Markham, A, & E. Buchanan (2017). Research ethics in context: Decision-making in digital research. In *The Datafied Society*, Van Es, Karin and Mirko Tobias Schäfer (eds.). University of Amsterdam Press, pp. 201–209.

Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media+ Society*, *7*(2), 20563051211021378.

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media+ Society*, *5*(2), 2056305119836778.

McQueen, G. (2022, 13 May). There Is No Safe Alternative To Twitter (Yet). *Ginny.Today,* https://ginnymcqueen.com/alternative-to-twitter/ , accessed 10 October 2024.

Melder, E., Lerner, A., & DeVito, M. A. (2025). "A Blocklist is a Boundary": Tensions between Community Protection and Mutual Aid on Federated Social Networks. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–30.

Nicholson, M. (2023, 5 October). An exploration of the Twitter to Mastodon migration. *CU Boulder Information Science*. https://medium.com/cuinfoscience/an-exploration-of-the-twitter-to-mastodon-migration-21c15c4336f2 , accessed 8 October 2024.

Pincus, J. (2017, May 10). Lessons (so far) from Mastodon for independent social networks. https://medium.com/a-change-is-coming/lessons-from-mastodon-for-independent-social-networks-ae2d4ccf8f72#:~:text=Federating%20with%20the%20Trouble , accessed 8 October 2024.

Pincus, J. (2022, 25 November). Mastodon: A partial history. https://privacy.thenexus.today/mastodon-a-partial-history/ , accessed 15 October 2023.

Pincus, J. (2023, 14 November). Mastodon and today's ActivityPub Fediverse are unsafe by design and unsafe by default. https://privacy.thenexus.today/unsafe-by-design-and-unsafe-by-default/ , accessed 15 December 2023.

Pincus, J. (2024, 13 August). 5 things white people can do to start making the fediverse less toxic for Black people. https://nexusofprivacy.net/start-making-the-fediverse-less-toxic/, accessed 11 November 2024.

Reagle, J. (2013). "Free as in sexist?" Free culture and the gender gap. *First Monday, 18*(1).

Rozenshtein, A. Z. (2023). Moderating the fediverse: Content moderation on distributed social media. *Journal of Free Speech Law*, *3*, 217-236.

Salehi, N. (2020, August 31). Do no harm. *Logic Magazine, 11*, https://logicmag.io/care/do-no-harm/ , accessed 11 November 2024.

Sinders, C. (2022, October 31). I'm @Sinders on Mastodon but I'm not giving up on Twitter, yet. *Medium*, https://medium.com/@carolinesinders/im-sinders-on-mastodon-but-i-m-not-giving-up-on-twitter-yet-5dcd4fb810e1 , accessed 11 November 2024.

Singh, R., & Sharma, S. (2019). Platform uncommons. *Feminist Media Studies*, *19*(2), 302–303.

Solomon, R. (2020). Meshiness: Mesh networks and the politics of connectivity. (Order No. 27837953, New York University). *ProQuest Dissertations and Theses* , 296.

Spencer-Smith, C., & T. Tomaz (2025). Labour pains: Content moderation challenges in Mastodon growth. *Internet Policy Review*, *14*(1), 1-21.

Suchman, L. (2002). Located accountabilities in technology production. *Scandinavian Journal of Information Systems*, *14*(2), 7.

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1), 20539517231177620.

Zhang, Z., Zhao, J., Wang, G., Johnston, S.-K., Chalhoub, G., Ross, T., Liu, D., Tinsman, C., Zhao, R., Van Kleek, M., & Shadbolt, N. (2024). Trouble in paradise? Understanding Mastodon admin's motivations, experiences, and challenges running decentralised social media. Proceedings of the Association for Computing Machinery, Human-Computer Interaction Vol. 8, No. CSCW2, Article 520.

Zulli, D., Liu, M., & Gehl, R. (2020). Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network. *New media & society*, *22*(7), 1188–1205.