Article:

Gibson, A.D. orcid.org/0000-0003-4602-8810, Docherty, N. orcid.org/0000-0001-5945-008X and Gillespie, T. (2023) Health and toxicity in content moderation: the discursive work of justification. Information, Communication & Society, 27 (7). pp. 1441-1457. ISSN 1369-118X

https://doi.org/10.1080/1369118x.2023.2291456

Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rics20

Health and toxicity in content moderation: the discursive work of justification

Anna D. Gibson, Niall Docherty & Tarleton Gillespie

To cite this article: Anna D. Gibson, Niall Docherty & Tarleton Gillespie (12 Dec 2023): Health and toxicity in content moderation: the discursive work of justification, Information, Communication & Society, DOI: 10.1080/1369118X.2023.2291456

To link to this article: https://doi.org/10.1080/1369118X.2023.2291456

Published online: 12 Dec 2023.

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Health and toxicity in content moderation: the discursive work of justification

Anna D. Gibson [a], Niall Docherty[b] and Tarleton Gillespie[c]

[a]Massachusetts Institute of Technology, Cambridge, MA, USA; [b]University of Sheffield, Sheffield, United Kingdom; [c]Microsoft Research, Cambridge, MA, USA

ABSTRACT

Within academia, industry, and government, the terms 'health' and 'toxicity' are widely used to describe and justify decisions around online content and its removal. However, the meanings of these terms are assumed to be self-evident and therefore are rarely examined. This article turns a critical eye to the health and toxicity metaphor to unpack its hidden political work. We trace the metaphor through three different discourses: the historical political economy of the term, the usage by cultural elites in the last two decades, and finally through its contemporary instrumental usage by volunteer content moderators on Facebook. By linking these discourses together, we argue that the metaphor of health and toxicity serves as a means for justification and legitimacy under contemporary neoliberalized orders that typically chafe at modes of public intervention and the language of democratic statecraft. Rather than elucidating the challenges of online content, we find that the metaphor often serves to obfuscate or sidestep the hardest problems in democratic governance. This analysis therefore has practical significance for researchers, policymakers, journalists, and other speakers that publicly traffic in this discourse at large.

ARTICLE HISTORY
Received 18 July 2022
Accepted 13 November 2023

KEYWORDS
Health; toxicity; toxic; moderation; governance

> I'm a mod of this group and my job is to try to keep the conversation productive and healthy, as much as I can …
>
> (Facebook Group moderator)

Introduction

In the quote above, gathered during ethnographic fieldwork, one volunteer Facebook Group moderator deploys the metaphor of health to describe the kind of conversations they feel they should cultivate within their community. Accordingly, the metaphor works to justify their moderation interventions, such as deleting user content and steering discussions, lending credence to the very position they inhabit as a moderator. The quote is

CONTACT Anna D. Gibson adgibson@mit.edu Massachusetts Institute of Technology, Cambridge, MA, USA

representative of the kind of language that many volunteer moderators use to describe their work (Seering et al., 2022). Their choice of metaphor also sits comfortably with broader explanations of platform moderation itself, echoing Zittrain's (2019) assertion that the thinking around platform governance is shifting from a focus on individual rights toward a discourse more akin to public health. But what exactly makes a conversation 'healthy' is often left unspecified.

Nowhere is this discourse of health more obvious than in the now common use of the term 'toxicity' to describe all that is pernicious on social media. Harassment, hate, conspiracy, and misinformation are lumped together as 'toxic' by moderators, by platforms, and in the growing public debate about social media and its implications. 'Facebook knows Instagram is toxic for teen girls, company documents show' (Wells et al., 2021) stated a *Wall Street Journal* headline, in its 2021 'Facebook Files' project. Quite often – too often – 'toxic' is treated as an unproblematic term for these problematic behaviors, implying that they are easily defined and identified, and an obvious harm to the health of the user or the public.

In this article we argue that researchers ought to be skeptical of the comfort, translatability, and traction that terms like health and toxicity provide, and instead question how they have come to be so prevalent, what they allow us to say and what they do not, and, crucially, in whose interests circulating these terms serve. Academic research into content moderation and platform governance often uses the terms healthy and toxic unproblematically, as self-evident, empirically tractable terms. We believe that that to do so represents an error. Wielding the terms healthy and toxic as if they were measurable categories, rather than metaphors, overlooks that they are deployed purposefully, by particular actors, for particular reasons, and with particular ends. These terms were not applied to the practices of moderation, they were taken from them, and are already implicated in questions about how and why to moderate, and who has the right to do so.

Rather than simply accept the current framing of health and toxicity at face value, we argue instead that this metaphor is deployed to do important discursive work – interjecting a frame that asserts the legitimacy to intervene in spaces of public conflict. We are interested in the terms in which people justify themselves and their actions, the roots of these regimes of justification, and how people forge mutually intelligible understandings of legitimate governance (Boltanski & Thévenot, 2006). Accordingly, we do not define health and toxicity online, or intervene in debates of how to address it. Instead, we critically analyze the functioning of health and toxicity in popular discussions of content moderation, specifically examining their discursive force when used by moderators to make sense of the anti-social behavior they encounter, and to validate their work to expel it.

This paper traces and links three different discourses of content moderation – historical, ideological, and instrumental – to highlight the tangible limits this metaphor imposes on how we think of public governance at large. While the metaphor of health and toxicity may appear self-evident, we will show how it in fact provides a specific means of justification under contemporary neoliberalized orders that typically chafe at modes of public intervention and the language of democratic statecraft. This analysis has practical significance for researchers, policymakers, journalists, and anyone that publicly traffics in this discourse at large.

We begin by tracing the history and political economy of health and toxicity metaphor, establishing how its implicit delineation of morality is conjoined with economies of blame. We show that, in the service of an imagined Liberal digital public sphere, the metaphor of health and toxicity has been productively adopted and circulated by a range of actors invested in legitimizing the governance of others. This framing helps diffuse moral culpability, while simultaneously providing a broadly agreeable, not-yet-politicized, justification for the removal, reduction, or suppression of contributions to that public sphere. We demonstrate how this metaphor has been taken up in the past two decades by digital culture researchers, powerful corporate elites, and feminist critics alike to both justify – and sometimes sidestep – the hard problems of cultivating democratic discussion within current neoliberal hegemonies (Mirowski, 2014).

We then analyze interview data from an ethnographic study of volunteer moderators of Facebook groups. More than 70 million volunteers worldwide moderate Facebook Groups, Facebook's version of discussion forums (Facebook, 2020). For all of the grand debates about content moderation and public governance, much of the actual work of moderation is left to volunteers like these, who not only oversee group interactions, but also must regularly justify to that group the decisions they make and why they get to make them. Facebook Group moderators are not given any explicit training or guidance from the company to aid them in their work, and therefore necessarily must craft their own theories of justification for their interventions. While moderators are handed technical authority over the community they are a part of, including the ability to act arbitrarily (Schneider, 2022), to actually do so would profoundly risk their standing within that community. In practice, moderators must maintain their quite fragile legitimacy, even as they do intervene. As such, in this context we ask: how is the metaphor of health and toxicity regularly put to work in day-to-day decisions, offering a rationale for, and lending legitimacy to, the governance of others? These statements from moderators should not be taken as explanations of 'how things are', rather, following Lamont and Swidler (2014), we look to interview data with moderators to reveal the discursive tools they have 'available to think about a problem' (p. 161).

Ontological toxicity, relations of blame

Healthy and toxic are not simple opposites. But together as a single metaphor they demarcate a spectrum, aligned with the polarity of good and bad, that frames long-standing concerns for communities, publics, and the body politic in particular ways. While this metaphor has gained purchase in contemporary debates about content moderation and platform governance, it is by no means new.

Melenia Arouh (2020) traces the word 'toxic' to its Greek and Roman etymology, where it initially described an arrow that had been poisoned with an 'explicit aim to kill or incapacitate' (p. 69). Arouh argues that in current mediatized contexts, rather than describing a poisonous appendage, toxic has come to refer to a poisonous ontological state, 'symbolic of a certain cultural malaise, where people, identities, and communication are seen as harmful' (p. 69). Through this lens, human individuals, relationships, institutions, or ideas might be described as toxic, denoting something that is subtly damaging to others once it enters.

A second meaning of the term toxic draws on environmental discourses to describe something that was once healthy but has become spoiled. In the United States, this is often associated with ecological calamity, specifically 'inadvertent forms of chemical poisoning' (Wexler, 2013, p. 172) following unchecked industrialism. Buell (1998) argues that this ‘long standing mythography of betrayed Edens’ reveals ‘incipient anxieties about the techno-economic progress’ (p. 647) upon which the United States in particular premised itself.

The term toxic, then, can refer to toxic *beings* or *acts*, or to diagnose *relations* that have somehow soured. For Wexler (2013), the generative discursive function of toxicity ‘modifies what was once healthy, but now due to neglect, corruption, or indifference is sliding into deeper and more virulent trouble’ (p. 172):

> the responsibilities of those who were to have anticipated, prevented, and/or removed the threat of toxicity becomes a blameworthy issue. The old broom is found wanting, and at least in the eyes of those calling attention to the toxicity, change is urgently warranted. (p. 173)

Or if toxic is a diagnosis, a name for both the poisonous agent and the condition of a public or community poisoned by it, then who has the responsibility for keeping the public ‘healthy’?

The answers to this question, historically, lie at the heart of the ongoing, and likely unresolvable, debate about the proper governance of the public sphere. Since at least the Enlightenment, Western liberal democracies have worried over the character of the public sphere (Habermas, 1999; Hyland, 1995), grappling with the relationship between the citizen and the state, the body politic and the sovereign, and over what counts as a legitimate site of governmental intervention (Robbins, 1993). In modern times this has included questions about the role of private intermediaries that provide public information, public space, or frameworks for public association – from newspapers to television networks to social media providers. Scholarship in media history (Peters, 2005), democratic theory (Bohman, 1990), and public opinion (Splichal, 1999) has demonstrated how these debates often crystallize around the emergence of communication technologies, specifically those that unsettle existing patterns of public dialogue by enabling new types of speech, speakers, and social formations. Today, then, questioning where responsibility for ‘healthy’ or ‘toxic’ social media lands must be understood in relation to the social, cultural and political milieu within which those questions are currently being asked.

At least in the context of the United States, it is persistent neoliberal modes of corporate deregulation and psychologically individuated governance that help explain the metaphoric power of health and toxicity. Rather than constituting a simple endorsement of *laissez-faire* political economy, neoliberalism, following Gane (2012), is singularly concerned with addressing ‘the appropriate powers of the state and the role it should play in ensuring the freedom of the market’ (p. 625). Influential Chicago School economists like Hayek and Fiedman (Stedman Jones, 2012) rejected bloated welfare state models of government, instead favoring the market’s transcendent powers to provide individuals with everything they needed to flourish as autonomous citizens. This style of governance has persisted in the United States for the past 40 years, impacting education, healthcare, agriculture, telecommunication, and innumerable other policy domains (Davies, 2014). As Foucault (2010) argued, neoliberalism operates by discursively and materially displacing the authority to manage the social, economic, and interpersonal interactions of the citizenry from the state to the competitive market. Debates over the governance of US based social media such as Facebook must be understood against this political backdrop.

Just as neoliberalization shifted the responsibility for physical health from the state to the individual, amidst marketized lifestyle choices and for-profit care (Docherty, 2021; Pilkington, 2016), it shifted responsibility for the metaphorical health of the public body as well. Individuals, cast as makers of their own destiny, are atomized from their socio-political circumstance, yet fully responsible for navigating its dangers – which are often left to grow unimpeded by any meaningful state regulation. Compounding this in the American context is a long brewing skepticism of expertise, be it scientific, journalistic, political, or technological (Hofstadter, 1966; Nelkin, 1975). The profound mistrust of institutions, the association of higher learning with elitism, structural critiques of the knowledge professions, and the return of political populism, have rendered expertise a liability (Grundmann, 2017; Jewett, 2020).

In a political environment in which regulation by either the elected or the expert is in disfavor, and individuals are left to the whims of the market and the marketplace of ideas, which concerns manage to rise to the level of public intervention, to whom does that task fall, and under what notion of legitimacy? Someone must govern communities that simply do not govern themselves. But on what standing do they regulate when regulation itself is taboo?

The discursive work of toxicity for online communities, their critics, and social media platforms

Social media providers have, with much discomfort, taken on the role of governing this imagined public sphere (Gillespie, 2018), despite the legitimacy of their governance regularly being called into doubt. Having opened their platforms widely to users to contribute nearly anything, platforms then reactively imposed regimes of content moderation that govern what kinds of content are unacceptable and algorithmically manage what should be highlighted. Platforms apply enormous human resources and increasingly sophisticated automated tools to identify and assess content, and ultimately impose a techno-social apparatus for policing that content, while proclaiming some measure of accountability to users for having done so. How this is accomplished and according to what criteria differ by platform, and even within a single platform.

Platforms face two crises in this regard. The first, of course, is what to remove and why, something users, critics, and advocates can argue over endlessly. But the second is whether their interventions, their private governance of the public sphere, is seen as legitimate: what gives them the right to make these judgments at all? Social media companies face this question of legitimacy as they defend their content moderation efforts – and individual group moderators face it too, when they must explain to users why a specific post needed to come down or why a specific user was banned.

Even in the earliest days of the Internet, community managers, aggrieved users, and researchers began to note that anti-social behavior was proliferating in online communities. In trying to name these behaviors, many community managers and social scientists leaned first on terminology drawn from the online cultures themselves: griefing, flaming, flooding, stalking, trolling. Calling these behaviors 'toxic' was initially useful as a catch-all term, but the metaphor eventually gained prominence because of the specific rhetorical work it could do. Suler (2004), one of the first to argue that online anonymity had a disinhibiting effect, noted that disinhibition could sometimes be 'benign', allowing people to be more honest and vulnerable; or, it could lead to anti-social behavior, including profanity, mischief, threats, and anger – what he dubbed 'toxic online disinhibition'. (That Suler was both a research psychologist and an active member of an early online community makes sense: the metaphor of health and toxicity presents an analytical, almost epidemiological framework, but it aligns it with the practical labor of community governance.) As the term slowly gained traction, it continued to 'outline a vision of a healthy conversation and its other' (Thylstrup & Zeerak, 2020), setting up a clear dichotomy between the community experience Suler expected to have, and the version he warned could undermine it.

A decade later, feminist bloggers and gamers used the metaphor again to call out the growing harassment they faced, the sexist representations in the games themselves, and the callous attitude toward sexual violence in the gaming communities around them (Consalvo, 2012; Jane, 2014). Here, the health and toxicity metaphor helped critics describe not just behaviors, but a condition that had infiltrated online spaces. This 'toxic gamer culture', in Consalvo's words, built on the idea of 'toxic masculinity' – misogynist, tactical, and systemic. Importantly, platforms themselves were seen as implicated: in the aftermath of #Gamergate, Massanari (2017) analyzed the 'toxic technocultures' that plagued Reddit, arguing that perpetrators were leveraging these sociotechnical systems to harass and exclude women, and that the design and governance of Reddit 'provides fertile ground for these kinds of toxic spaces to emerge' (p. 2).

The expanding press coverage of online harms in Anglophone liberal democracies has increasingly embraced the word 'toxicity' and its connotations (as well as related metaphors of dirt and pollution) (Phillips & Milner, 2020; Thylstrup & Zeerak, 2020). In 2018, Amnesty International published a report titled *#Toxictwitter*, warning that harassment, violence, and abuse drives women from the platform, silencing them. Because 'the rhetoric of toxic discourse depends on a narrative of an idyllic space that has turned into a lethal one' (Risam, 2015), it was well-suited to a concern that Twitter itself had been poisoned.

For many of the same reasons, social media companies have also embraced the metaphor of health and toxicity when trumpeting new moderation efforts of their own. Calling some online behaviors or communities toxic allows platforms to demonstrate the depth of their concern, while positioning themselves as benevolently diagnosing the problem – rather than being framed as *responsible* for it, as both Massanari and Amnesty International argued. The scientific, quantifiable, connotations of 'toxic' also align neatly with the impulse of these companies to reduce sociotechnical problems to what can be measured and treated with the tools of data science and machine learning.

For example, in 2018 Twitter announced it would fund academic research, 'to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress' (Dorsey, 2018). The award description went further: 'Recently we were asked a simple question: could we measure the "health" of conversation on Twitter?' That question had come from Cortico Research, whose mission was to 'measure aspects of the health of the public sphere – in terms of communication exchanges between groups or tribes – grounded in data from public social media and other public media sources' (Roy et al., 2018). Twitter left it to the researchers they funded to define what conversational health was.

In 2017 Google's Jigsaw team released Perspective, a bundle of machine learning classifiers that assess whether a post is sexually explicit, threatening, profane, or 'toxic'. Unlike Twitter's more holistic sensibility, Perspective parses public participation down to its smallest components: the toxicity classifier assesses single posts, as opposed to users, communities, an entire platform, or the public sphere at large (Rieder & Skop, 2021). Jigsaw trained its toxicity classifier on example posts rated by humans, who were instructed to assess whether a comment is healthy or toxic, the latter defined as 'a rude, disrespectful, or unreasonable comment that may make you leave a discussion' (Jigsaw, 2017). The term 'toxicity' was chosen because of the apparent consensus it seemed to produce: Jigsaw said it settled on the word after finding that most reviewers agreed about what types of comments drive people away from a conversation (Wakabayashi, 2017).

Toxicity works as an umbrella term for the array of anti-social behaviors that plague online communities and social media platforms. Yet the examples above demonstrate the term does a great deal more for the different stakeholders concerned with these behaviors. What is labeled toxic is sometimes the poison, sometimes the poisoner, sometimes the body/environment that has been poisoned. Terms like 'healthy' or 'toxic' don't need to mean the same thing to different people, or research teams, or platforms – they merely need to seem to mean the same thing. At the same time, the metaphor of health and toxicity offers some rhetorical distance from blame or responsibility. A group moderator or a social media company can position themselves as delivering a serious diagnosis, but also as assuredly looking for a remedy. And it retains an implicit presumption of what healthy is or should be, justifying the taking of action on behalf of a social group, to address what ails it.

The discursive work of Facebook group moderators

In early 2017 CEO Mark Zuckerberg announced he was changing the mission of his company; rather than trying to make the world 'more open and connected', it would now 'give people the power to build community and bring the world closer together' (Wagner & Swisher, 2017). This meant centering the role of Facebook Groups on the platform, including moving the Groups icon to the center on the app's interface. By 2021, Facebook reported 1.8 billion people were involved in meaningful Groups globally, including 70 million admins (Facebook, 2021). These groups might be small, a few people who want to share pictures of an outing, or enormous, tens of millions of users gathering around political beliefs. While Groups have become key sites for broad anti-democratic and anti-vaccine organizing (Jankowicz & Otis, 2020), for many, Facebook Groups keep the site relevant (Petersen, 2022).

Importantly, Groups also helped Mark Zuckerberg move the onus of responsibility for moderating users' interactions from the company to the 'admins' who run these groups. Shifting the burden of regulatory responsibility from Meta to individual volunteer users, is indicative of the neoliberal atomization and responsibilization, the lack of willful state regulation, and the dual tropes of corporate triumphalism and community neglect that are manifest in Facebook's approach to Groups more generally.

Even as Zuckerberg has since seemed to have forgotten Facebook Groups, dazzled by his own visions of the 'metaverse', tens of millions of moderators continue to manage these groups, with minimal assistance, guidance, or political cover from the company. Volunteer moderators are left to solve a fundamental dilemma of the imagined public sphere on their own: on what legitimate basis can they intervene in the workings of the 'free' participatory space facilitated by Groups, and even potentially exclude people from that space? Ethnographic data suggests that taking on the role of Group admin or moderator actually poses a crisis of justification for many people, especially those who themselves identify with the ideology of democratic liberalism (Gibson, 2022).

Here, we analyze interview data from a larger ethnographic study about how volunteer content moderators on Facebook make removal decisions, to specifically examine when and how they invoke health and toxicity discourses. The first author conducted thirty semi-structured interviews[1] avoiding the use or mention of this metaphor herself. And yet, as the moderators confronted the problem of decision-making justification, they often drew on familiar discourses of health and toxicity.[2]

The semantic flexibility of healthy vs. toxic

The moderators we spoke to did not have a consistent definition of healthy and toxic, or any definition at all; rather they appeared to deploy the terms more functionally. 'Healthy' discussions seemed to be whatever led to good outcomes for the members of the group. In this usage, health suggests ideas of vigor and robustness. Like a strenuous workout, the actual discussion might be momentarily uncomfortable for participants. However, moderators believed that the benefits of this temporary discomfort would be outweighed by the communal benefits of members' increased understanding of each other and the world. The moment when discussions become unhealthy, or toxic, was when that net benefit disappeared; the exercise injures rather than strengthens. Moderators demonstrated the same semantic flexibility with 'toxic' as described above: toxicity could be a specific dynamic or interaction, a member of the group, or sometimes a condition of the group itself. The metaphor affords moderators flexibility, justifying allowing some level of conflict within the group without requiring them to allow all conflict, and to do so conditionally in different contextual situations.

The metaphor also offers moderators a more respectable language for decisions that are, in the end, subjective judgments of what is ultimately good or bad for a community. For Ida, a Black woman who runs a fan group for a US television show with racial themes, the important distinction was between self-expression and recklessness. She reported creating the group as a 'space for us to talk about the show [and] the nuances of racism, how it relates to today'. She wants group members to feel like they can express themselves: 'Let it out. Be you. Say what you want to say'. Because she wants to encourage people to express their feelings, she anticipates some arguments, and even thinks that they are productive. Disciplining members depends on her anticipation of whether this conflict will yield net positive results:

> So what I'll do is, it depends on how reckless you are … Good conflict is good. Healthy conflict is good. But when it's real negative and disrespectful, that's not cool.

Ida uses the term 'reckless' to describe behavior that will lead her to block members from the group. Such recklessness is dangerous because the conflict will no longer be 'good' or 'healthy' but harmful.

Simon moderates a very large podcast fan group and believes the normative value of individual user posts depends on the type of discussion they will generate. He described the main function of his role as 'approving [member-submitted] posts to make sure that they're actually good, are going to generate healthy discussion in the group'. While he did not define what 'healthy' specifically means for him, his later remarks indicate that good discussions allow for individual growth and education, especially when it means turning a critical eye on the hosts of the podcast. 'Frankly', he said, there are times where [podcast host] should be bullied, pointing to the host's history of casual racism and transphobia as reasons why fan critique was so important. Distinguishing between generative and destructive defines healthy and toxic as Simon sees it.

For Shady Boy, who moderates a large meme-posting group, the tension is between interesting contributions and irredeemable ones. She noted that her main rule for members is 'don't be a jerk'. She then described that the moderators of her group, 'are not here to police you', but also that they will 'just get rid of anybody that may engage in toxic behavior'. To her eyes, the term required no further explanation. Max, a moderator for a different large meme group, seemed to affirm the poison connotations of toxic content. He described a common situation in which most of the discussion will be 'really interesting, it feels joyful and good', but that one single comment will be 'horrible' and spawn a sub-thread where 'all the replies under that comment are really toxic or upsetting'. The group moderators will shut down an entire sub-thread because it is already too far gone.

Justificatory cover for earned but tacit expertise

Moderators develop expertise over time that allows them to better judge how to govern a community. For instance, many moderators described that certain topics come up again and again that they feel inevitably lead to overall undesirable outcomes. Some moderators had developed the practice of deleting certain kinds of posts because they were bound to lead to problems. Chuckie, a moderator of a role-playing game group, said he had to ban threads where members asked for help in naming characters. 'I cannot explain why this happens … there are certain posts that we just know will go bad so we automatically refuse them'. Knife described what has changed most for her since beginning her work as a moderator for a fan group is now 'knowing what's going to be bad' when posts come in. She immediately followed up with 'That sounds terrible to say, but genuinely it helps'. Both Chuckie and Knife could offer only pragmatic, rather than dogmatic, justification for these removals based on their experiences of the dynamics in their respective groups. Identifying 'bad' content as toxic justifies this pruning approach to moderating discussion, where the normative objection isn't necessarily the single post itself, but the knowledge that its subject matter or phrasing will generate *too much* conflict, so it is therefore poisonous and harmful to the group as a whole.

Such conditional, ad hoc, and tacit expertise may be effective, but is difficult to justify as fair and accountable governance. Framing it as being able to diagnose the difference between 'healthy' and 'toxic' helps to justify and cloak that deployment of expertise. These cloaking practices, we argue, become even more necessary due to the neoliberal political milieu of their expression, where the value of public-facing expertise to justify interventions runs counter to by the idealized sovereignty of autonomous, citizen decision-makers, acting individually, and only in terms of their own self-interest.

Toxicity was not just used to name problematic behaviors or threads of conversation; in the terminology of these moderators, individual members of the group may themselves be toxic actors. Knife, described above, helps run a very large fan group of a television show; both the show and its fandom have been criticized for their problematic portrayal of race in the United States, and Knife thinks that using the group space to criticize the fandom is 'healthy' and beneficial, in theory. However, she complained that certain members of the group felt personally attacked by such criticism, describing them as 'toxic little assholes'. When they made themselves known, she needed 'to remove them in order to restore health to the group'. Chuckie, also identified above, said that he has no qualms about removing people 'that do not belong' from the group because he sees it as necessary to 'maintain a healthy community'.

The gradually acquired and largely tacit expertise required to make a judgment as profoundly important to a community as 'who does not belong' benefits from being framed in the language of health and toxicity. The metaphor invokes other kinds of procedures that maintain bodily health or integrity through purge or excision. The body, here, is the community: a construct made up of all the relations of all members of the group to one another. The presence of a toxin in this body threatens the stability of the overall community, as it may be a source of persistent annoyance or corrupt relations amongst the group. If it were simply a disease, it could be cured; toxins, though, are fundamentally foreign incursions that must be flushed from the system. Describing a person as toxic both prefigures and justifies their removal, in one sweeping motion.

Cautionary tales of toxic communities

In this language, entire groups, too, can be healthy or toxic. If the 'little assholes' are left unchecked, the entire environment may itself become toxic. This usage invokes the analogies to pollution or disaster introduced above, in which the very air, water, or soil has become poisoned by external forces. When a group reaches this point, the only solution is to abandon or destroy it. None of the moderators we spoke with were currently grappling with profoundly toxic communities. (The sample we spoke with was presumably biased toward those who were reasonably happy in their current groups and therefore amenable to an interview.) When they did describe unhealthy groups they had previously been a part of, but since either left or were no longer active in, these stories were delivered as cautionary tales, to highlight how toxicity can spread and to justify early interventions.

Tess, who runs a group for writers, described a group she had moderated previously as having a lot of drama and anger: 'To me, it felt like there was no logic and I was getting so pulled into it … . I just kind of felt like, yeah, that's not how it should be and that's not a healthy group'. She offered this story as a contrast to her current group, which was not having any of those problems. Interestingly, she couldn't specify why the groups were so different: 'I don't really know why the other group I'm [in], that just doesn't happen in, I don't know what the reason is'.

Toxicity is useful for diagnosing a community that doesn't seem to have any one specific problem but is nonetheless unredeemable. Describing a community as toxic provides a useful justification for abandoning it. Ellis explained that when he was looking for groups to moderate, he specifically avoided one group because 'it's just a toxic cesspool'. As a moderator, he wouldn't be able to make a difference, and as an individual, it was a space he wanted to avoid.

By contrast, this means that keeping groups 'healthy' can serve as a powerful justification for any and all moderation decisions made in a given space. Knife said she wanted to be involved in cultivating a 'healthy environment' in her fan group:

> I want to be as involved in this as possible. I want to make it be a good space, because I know as a queer POC person who has lived through a lot of poverty and a lot of other experience, I can help make this space more accepting and safe for other people and a healthy environment, I guess, more than anything.

Finally, it's worth noting that many moderators also used the concept of health to describe their own relationships with the Facebook platform itself. In these cases, the antonym of health is addiction, rather than toxicity. Several moderators expressed concern for their own mental health and the mental health of their groups because of excessive Facebook use. For example, Carlos, who helps run a meme group, said he came close to deleting his Facebook account. 'I totally buy the whole "it's horrible for you and your mental health and addictive" whatnot', he said. 'I deleted it from my phone and I found that's a good balance'. These moderators use their own feelings, positive or negative, about using Facebook to decide whether their use of Facebook is healthy or unhealthy.

No universal metaphors

Even the broadest discursive justifications do not necessarily have universal purchase. Healthy and toxic can only do work for specific audiences once they have become a part of common discourse. The utility of the health discourse varied by participants' cultural backgrounds. Martin is a Latin American man from Mexico currently living in a large Asian city. For him, the idea of 'toxic' is tightly linked to 'toxic masculinity'; to discourage toxicity in his groups he leans heavily on stereotypically feminine symbols like pink hearts. Another participant, Larry, is involved in the culture of American men's sports, and was somewhat confused by the idea of toxicity, as he was only familiar with its usage with respect to romantic relationships.

While the metaphor of health and toxicity appeared in a variety of ways across many of the moderators we spoke with, it was by no means the only metaphor they drew on to justify their efforts to govern. A different paper could just as well document the way metaphors of civility, safety, or fairness are also deployed by moderators to justify their efforts. But for those that did use health and toxicity to describe and justify their work, it clearly solved particular problems concerning the right to privately govern a public space, particular to the time and place of their articulation.

Conclusion

Rather than identify a precise definition of health and toxicity on social media, if this was even possible, or desirable, we have instead revealed the discursive work of justification that the health and toxicity metaphor *does*. Facebook group moderators are dealing with a fundamental problem at the heart of the imagined public sphere: by whom, and by what right, can someone be excluded from public conversation? Thus, when deciding whether a post should be removed or a user suspended, it is useful to have a discursive framework that highlights the value of long-term community 'health', and that can pinpoint the peril of unchecked anti-social behavior to quickly poison the community and make it 'toxic'. The specific connotation of medical or ecological diagnosis, which justifies present assessments about future harms, helps frame these interventions as more than just subjective guesses about what may be good or bad. The discourse of health and toxicity allow moderators to authorize their own experience as the basis of their legitimacy to govern; it is their experience in handling the tricky issues of Group moderation that quietly grants them the right to continue to do so.

When deployed, the metaphor of health and toxicity helps make sense of not only what kinds of public contributions should be removed from certain spaces, but how this imposition of ostensibly arbitrary power should be understood as a legitimate effort to care for an imagined public. This metaphor has proven useful to group moderators for several reasons: for identifying the purveyors of abuse, narrating the psychology behind the behaviors in question, and calling attention to its consequences.

The early web practice of volunteer, light-touch community moderation has been largely overshadowed by the industrial logic of platforms like Facebook, where engagement, data collection, and advertising revenue throw the value of such subtle censure in doubt. Platforms, in the US at least, seem keener to invoke the First Amendment than to willfully intervene. Most problematic content issues on the rest of the Facebook platform are managed by thousands of outsourced workers, making rapid-fire removal decisions using proprietary guidelines built by Facebook largely to suit advertisers' needs (Barrett, 2020; Roberts, 2017). It is only in spaces like Groups, which Facebook has both elevated and somehow also ignored, that the problem of community governance explicitly lingers as a lived concern. Here, Facebook has passed the tough responsibility of moderating appropriate communication to volunteer moderators, with little to no direction, support, or renumeration in exchange.

Although the discourses and practices of healthy moderation seem to do discursive work for the Facebook Group moderators, we must also recognize that they also serve Facebook's economic interests to a significant degree: it is precisely this form of civic, *volunteer* engagement that makes Facebook Groups such an appealing space for sociality on the platform in the first place. The labor of group moderators not only compensates for the willfully anemic governance offered by Facebook itself, it also adds economic value to the platform at the same time. Facebook profits by offering advertisers the attention of a desired market segment, no matter how niche (Vaidhyanathan, 2018). While Facebook does not place advertisements in Group spaces specifically, advertisements in users' algorithmically assembled Feeds mingle with posts from their Groups. Therefore, users' continued interest and investment in their Facebook Groups helps Facebook flourish as a profit-driven data platform (Alaimo & Kallinikos, 2017).

We hope we have shown that the metaphor of health and toxicity functions as a means to discursively legitimate, but also de-politicize, the deeply political nature of content moderation. This is a useful tactic for journalists and policymakers, who do not recognize or may wish to cloak their own normative positions; and it is useful for moderators themselves, who need to keep their communities working and remain in a position to do so. However, this metaphor ultimately *serves the interests of platform companies*, by situating moderation dramas in terms that do not question their position as capitalist arbiters of the theoretically collective public sphere, or their shared responsibility for the strife that has been dubbed ‘toxicity’. Social media researchers, therefore, should reconsider when unthinkingly deploying terms like healthy and toxic as if they are empirical states of communities; doing so affirms a framing of social conflict and public governance most preferred by the platforms.

A discourse of health and toxicity elucidate battles over forms of life (Jaeggi, 2018) – tacitly normative cultural ensembles that function through the problematization, and solution, of human praxis and organization. Pfaffenberger (1992) argues that such ‘technological dramas’ take do not take rhetorical shape immaculately, but are always ‘projected into a spatially defined, discursively regulated social context’ (p. 291). Such frames are powerful because they do not emerge only inside specific debates about content moderation, but are already widely available as resonant discourses. The metaphor of health and toxicity used by volunteer Facebook moderators already entangle with long-standing, and controversial conceptions of public governance – of healthy bodies and healthy body politics; of the public sphere as a vital environment, which can be verdant or poisoned; and of the caustic effect of toxic behavior, masculinity, and relationships on particular ways of being in the world. Talking about and acting upon toxic behavior in the moderation of social media is thus always a situated proposition, revelatory of the hopes and fears of particular historical political struggles, and justificatory of specific forms of regulatory action as appropriate to their solution.

The perennial questions about the proper governance of the public sphere did not disappear with the emergence of the commercialized semipublic spaces constructed by social media companies. Rather, it was *how* these questions are posed that shifted. A tangible distaste for state regulation and expertise in neoliberalized political cultures set the scene for these crises of platform governance. We have shown how the metaphor of health and toxicity stands in for this deep justificatory abscess at the heart of democratic communication, made worse by the neoliberal times in which they are presently felt. Rather than relying on values like justice or equity, the metaphor the metaphor allows moderators to justify their interventions on behalf of others in the group using the language of diagnosis and care rather than policing, and thus grants them legitimacy when the very basis of that intervention is so shaky.

Notes

1. See Appendix A for information about study methods and participants. All names in this paper are pseudonyms.
2. Moderators must publicly demonstrate concern for a number of stakeholders, including their groups, other moderators, and the platform (Matias, 2019). Within an interview, the work of justification through thoughtful introspection must also be explicitly performed for the interviewer. These interviewees were therefore often concerned with articulating implicit and explicit justifications that guide their decision-making.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by School of Humanities and Sciences, Stanford University: [Grant Number Graduate Research Opportunity].

Notes on contributors

Anna D. Gibson is a Postdoctoral Associate in Comparative Media Studies/Writing at the Massachusetts Institute of Technology. She uses qualitative and quantitative methods to study the experiences of digital publics through lenses of labor and justice.

Dr *Niall Docherty* is a lecturer in Data, AI and Society at the Information School, University of Sheffield. His work examines the modalities of neoliberal governance, capitalism, and (bio)power expressed in digital well-being.

Tarleton Gillespie is a senior principal researcher at Microsoft Research, and an affiliated associate professor in the Department of Communication and Department of Information Science at Cornell University. His most recent book is *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale, 2018).

ORCID

Anna D. Gibson http://orcid.org/0000-0003-4602-8810

References

Alaimo, C., & Kallinikos, J. (2017). Computing the everyday: Social media as data platforms. *The Information Society*, *33*(4), 175–191. https://doi.org/10.1080/01972243.2017.1318327

Arouh, M. (2020). Toxic fans: Distinctions and ambivalence. *Ex-centric Narratives: Journal of Anglophone Literature, Culture and Media*, *4*, 67–82.

Barrett, P. M. (2020). *Who moderates the social media giants? A call to end outsourcing*. NYU Stern Center for Business and Human Rights.

Bohman, J. F. (1990). Communication, ideology, and democratic theory. *American Political Science Review*, *84*(1), 93–109. https://doi.org/10.2307/1963631

Boltanski, L., & Thévenot, L. (2006). *On justification: Economies of worth*. Princeton University Press.

Buell, L. (1998). Toxic discourse. *Critical Inquiry*, *24*(3), 639–665. https://doi.org/10.1086/448889

Consalvo, M. (2012). Confronting toxic gamer culture: A challenge for feminist game studies scholars. *Ada: A Journal of Gender, New Media, and Technology*, *1*. https://doi.org/10.7264/N33X84KH

Davies, W. (2014). *The limits of neoliberalism*. Sage Publications.

Docherty, N. (2021). Digital self-control and the neoliberalization of social media well-being. *International Journal of Communication*, *15*, 3827–3846.

Dorsey, J. (2018, March 1). We're committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress [Tweet]. *Twitter*. https://twitter.com/jack/status/969234275420655616

Facebook. (2020, October 1). We're launching new engagement features, ways to discover groups and more tools for admins. *What's New* (Blog). https://www.facebook.com/community/whats-new/facebook-communities-summit-keynote-recap/

Facebook. (2021, February 23). The power of virtual communities. *Facebook Community*. https://www.facebook.com/community/whats-new/power-virtual-communities/

Foucault, M. (2010). *The birth of biopolitics: Lectures at the Collège de France, 1978–1979*. Palgrave Macmillan.

Gane, N. (2012). The governmentalities of neoliberalism: Panopticism, post-panopticism and beyond. *The Sociological Review*, *60*(4), 611–634. https://doi.org/10.1111/j.1467-954X.2012.02126.x

Gibson, A. (2022). 'My Other Job': Volunteer content moderation as platform labor [Doctoral dissertation]. Stanford University.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Grundmann, R. (2017). The problem of expertise in knowledge societies. *Minerva*, *55*(1), 25–48. https://doi.org/10.1007/s11024-016-9308-7

Habermas, J. (1999). *The structural transformation of the public sphere*. Polity Press.

Hofstadter, R. (1966). *Anti-intellectualism in American life* (1st ed.). Vintage.

Hyland, J. L. (1995). *Democratic theory: The philosophical foundations*. Manchester University Press.

Jaeggi, R. (2018). *Critique of forms of life*. The Belknap Press of Harvard University Press.

Jane, E. A. (2014). "Your a Ugly, Whorish, Slut": Understanding E-bile. *Feminist Media Studies*, *14* (4), 531–546. https://doi.org/10.1080/14680777.2012.741073

Jankowicz, N., & Otis, C. (2020, June 17). Facebook groups are destroying America. *Wired*. https://www.wired.com/story/facebook-groups-are-destroying-america/

Jewett, A. (2020). How Americans came to distrust science. *Boston Review* 16. https://bostonreview.net/articles/andrew-jewett-science-under-fire/

Jigsaw. (2017). What do perspective's scores mean? *Medium*. Retrieved September 22 from https://medium.com/jigsaw/what-do-perspectives-scores-mean-113b37788a5d

Lamont, M., & Swidler, A. (2014). Methodological pluralism and the possibilities and limits of interviewing. *Qualitative Sociology*, *37*(2), 153–171. https://doi.org/10.1007/s11133-014-9274-z

Massanari, A. (2017). #Gamergate and the fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2), 205630511983677. https://doi.org/10.1177/2056305119836778

Mirowski, P. (2014). *Never let a serious crisis go to waste*. Verso.

Nelkin, D. (1975). The political impact of technical expertise. *Social Studies of Science*, *5*(1), 35–54. https://doi.org/10.1177/030631277500500103

Peters, J. D. (2005). *Courting the abyss: Free speech and the liberal tradition*. University of Chicago Press.

Petersen, A. H. (2022, February 13). This party sucks, why haven't we left [Substack newsletter]. *Culture Study*. https://annehelen.substack.com/p/this-party-sucks-why-havent-we-left

Pfaffenberger, B. (1992). Technological dramas. *Science, Technology, & Human Values*, *17*(3), 282–312. https://doi.org/10.1177/016224399201700302

Phillips, W., & Milner, R. (2020). *You are here: A field guide for navigating polluted information*. The MIT Press.

Pilkington, M. (2016). Well-being, happiness and the structural crisis of neoliberalism: An interdisciplinary analysis through the lenses of emotions. *Mind & Society*, *15*(2), 265–280. https://doi.org/10.1007/s11299-015-0181-0

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective API. *Big Data & Society*, *8*(2), 205395172110461. https://doi.org/10.1177/20539517211046181

Risam, R. (2015). Toxic femininity 4.0. *First Monday* 20. https://journals.uic.edu/ojs/index.php/fm/article/view/5896

Robbins, B. (Ed.). (1993). *The phantom public sphere*. University of Minnesota Press.

Roberts, S. T. (2017). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Roy, D., Yi, E., & Stevens, R. (2018, March 1). Measuring the health of our public conversations. *Cortico*. https://cortico.ai/news/measuring-the-health-of-our-public-conversations/

Schneider, N. (2022). Admins, mods, and benevolent dictators for life: The implicit feudalism of online communities. *New Media & Society*, *24*(9), 1965–1985. https://doi.org/10.1177/1461444820986553

Seering, J., Kaufman, G., & Chancellor, S. (2022). Metaphors in moderation. *New Media & Society*, *24*(3), 621–640. https://doi.org/10.1177/1461444820964968

Splichal, S. (1999). *Public opinion: Developments and controversies in the twentieth century*. Rowman and Littleford.

Stedman Jones, D. (2012). *Masters of the universe: Hayek, Friedman and the birth of neoliberal politics*. Princeton University Press.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Thylstrup, N., & Zeerak, W. (2020). Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3709719

Vaidhyanathan, S. (2018). *Anti-social media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Wagner, K., & Swisher, K. (2017, February 16). Read Mark Zuckerberg's full 6,000-word letter on Facebook's global ambitions. *Vox*. https://www.vox.com/2017/2/16/14640460/mark-zuckerberg-facebook-manifesto-letter

Wakabayashi, D. (February 23, 2017). Google cousin develops technology to flag toxic online comments. *The New York Times*. https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html

Wells, G., Horwitz, J., & Seetharaman, D. (2021, September 14). Facebook knows Instagram is toxic for teen girls, company documents show. *Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

Wexler, M. N. (2013). Rachel Carson's toxic discourse: Conjectures on counterpublics, stakeholders and the "Occupy Movement". *Business and Society Review*, *118*(2), 171–192. https://doi.org/10.1111/basr.12007

Zittrain, J. (September 23, 2019). Three eras of digital governance. https://ssrn.com/abstract=3458435

Appendix A: Study methods and participants

The data for this project comes from a study conducted by the lead author between 2020 and 2022 (Gibson, 2022) under approval #55290 from the Stanford University IRB. Participants were recruited through invitations sent to Facebook group admins and moderators, identified through Facebook's 'Discover' feature and snowball sampling, until theoretical saturation was achieved (Small, 2009). The 30 interviews represented in this dataset (see Table 1) are the total subset of all participants that verbally consented to an IRB-approved consent form including language that their anonymized data could be used beyond the scope of the original research project.

Interview protocols were structured as prompts rather than questions (Jiménez & Orozco, 2021) to gain information about participants' normal experiences as volunteer content moderators on Facebook. Interviews ranged from 30 min to 2 h and were conducted over Zoom or by phone. Participants were compensated USD$25 or local equivalent. Most were American, although others described themselves as Australian, European, Canadian, and Mexican.

The lead author used NVivo12 to iteratively code and memo the interview transcripts using grounded theory methods, which include coding data with emerging themes, grouping these codes to inductively identify emergent patterns, and then iteratively adjusting interview protocols with special attention toward concepts of interest over time (Charmaz, 2006). From among several hundred codes, emergent themes relevant to this study are discussed in the body of the paper.

Table 1. Demographic characteristics of participants.

	Frequency	Percent
Gender		
Agender	1	3.3
Gender fluid	1	3.3
Nonbinary	3	10.0
Woman	12	40.0
Man	13	43.3
Race/ethnicity		
Asian	1	2.6
Black	3	7.7
Hispanic and/or Latinx	4	10.3
Jewish	3	7.7
Mixed race	3	7.7
Native American	2	5.1
White, Caucasian or European-American	14	35.9
Age (years)		
20–29	13	43.3
30–39	8	26.7
40–49	4	13.3
50–59	1	3.3
60–69	3	10.0
70–79	1	3.3

Source*:* Interview data with volunteer Facebook Group admins and moderators.
Notes*:* Data comes from interviews with 30 participants.

Appendix references

Charmaz, K. (2006). *Constructing grounded theory*. Sage Publications.

Gibson, A. (2022). *'My Other Job': Volunteer content moderation as platform labor* [Doctoral dissertation]. Stanford University.

Jiménez, T. R., & Orozco, M. (2021). Prompts, not questions: Four techniques for crafting better interview protocols. *Qualitative Sociology*, *44*(4), 507–528. https://doi.org/10.1007/s11133-021-09483-2

Small, M. L. (2009). 'How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography*, *1*(10), 5–38. https://doi.org/10.1177/1466138108099586