Received: 29 September 2021 | Accepted: 23 February 2022

DOI: 10.1002/poi3.290

RESEARCH ARTICLE

P&I Policy & Internet PSO WILEY

Safe from "harm": The governance of violence by platforms

Julia R. DeCook[1] | **Kelley Cotter**[2] | **Shaheen Kanthawala**[3] | **Kali Foyle**[1]

[1]School of Communication, Loyola University Chicago, Chicago, Illinois, USA

[2]College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania, USA

[3]Department of Journalism and Creative Media, The University of Alabama, Tuscaloosa, Alabama, USA

Correspondence

Julia R. DeCook, School of Communication, Loyola University Chicago, 51 E Pearson Street, Chicago, IL 60611, USA.
Email: jdecook@luc.edu

Abstract

A number of issues have emerged related to how platforms moderate and mitigate "harm." Although platforms have recently developed more explicit policies in regard to what constitutes "hate speech" and "harmful content," it appears that platforms often use subjective judgments of harm that specifically pertains to spectacular, physical violence—but harm takes on many shapes and complex forms. The politics of defining "harm" and "violence" within these platforms are complex and dynamic, and represent entrenched histories of how control over these definitions extends to people's perceptions of them. Via a critical discourse analysis of policy documents from three major platforms (Facebook, Twitter, and YouTube), we argue that platforms' narrow definitions of harm and violence are not just insufficient but result in these platforms engaging in a form of symbolic violence. Moreover, the platforms position harm as a floating signifier, imposing conceptions of not just what violence is and how it manifests, but who it impacts. Rather than changing the mechanisms of their design that enable harm, the platforms reconfigure intentionality and causality to try to stop users from being "harmful," which, ironically, perpetuates harm. We provide a number of suggestions, namely a restorative justice-focused approach, in addressing platform harm.

KEYWORDS

discourse analysis, harm, platform governance, platform policy, symbolic violence

INTRODUCTION

Terms of service and moderation policies have always existed on platforms to some degree, particularly when it came to illegal content like child sexual exploitation or sale of regulated goods (drugs, firearms, etc.). But in recent years, platforms have scrambled to adjust their policies around what constitutes "harmful content" following a number of tragedies involving terrorism, mass shootings, and other acts of violence that could be traced to their platforms' affordances (Daniels, 2018; Odag et al., 2019). Indeed, after years of attempting to absolve themselves of blame or responsibility for what occurs on their services, platforms like Facebook, Twitter, Reddit, YouTube, and others have implemented new sweeping policies that have banned groups, hashtags, and other means of digital connection and coordination that these platforms afforded (Donovan, 2020a; Ganesh, 2018; Gillespie, 2018; Suzor et al., 2019). These platforms have had to reckon with their role in radicalization, extremist organizing, and rampant disinformation that threatens the social fabric of our society; in particular, the COVID-19 pandemic, violent white supremacist organizing during Black Lives Matter protests in the summer of 2020, and the January 6 U.S. Capitol Insurrection.

The repertoires of action platforms adopt in response to these problems—what has been called governance by platforms (Gillespie, 2017)—exhibit a variety of issues pertaining to platforms' commercial interests, isomorphic decisionmaking, influence over perceptions of their role and responsibilities, and sovereignty over decisions about what constitutes content or behavior worthy of remediation. In terms of their commercial interests, platforms must keep their users engaged, which often means ensuring popular accounts remain active and visible. This has led to inequitable moderation practices, wherein high-profile accounts receive greater leniency or complete immunity when it comes to enforcing platform rules (Caplan & Gillespie, 2020; Horwitz, 2021). Platforms also balance the concerns of their users with those of advertisers. However, as direct generators of revenue, advertisers' concerns in relation to moderation policies and practices tend to hold greater sway than that of average users. Beyond user concerns, platforms tend to develop and lean on AI and automation for identifying and managing harmful content to minimize costs associated with human moderation. Yet, these technologies frequently fail at detecting severe instances of harm that are nuanced and complex (e.g., hate speech; Gorwa et al., 2020).

Platform governance decisions also exhibit isomorphism (Caplan & Danah, 2018), as platforms adopt similar policy changes simultaneously or one after another. This suggests decision-making in response not only to public pressure but also peer pressure. We can see such peer pressure in, for example, decisions to remove figures like conspiracy theorist Alex Jones, former president Donald Trump, and many others who have been identified as central figures in these "harmful" communities that engage in offline violence (Fischer & Gold, 2021; Schwartz, 2019). The coevolution of platform policies helps platforms urge particular normative visions of how and when they should intervene in what their users post and share based on certain framings of what is "harmful" or "violent" enough to warrant removal and outright bans.

As platforms learn from each other, many strides have been made in updating policies around hate speech, harmful content, misinformation, and other forms of malicious content. Neverthless, it appears that platforms often use subjective judgments of harm that specifically pertain to spectacular, physical violence. Such judgments belie the complex political concepts and processes subsumed by the terms harm and violence (Arendt, 1970; Bourdieu, 1999; Giroux, 2017). The politics of circumscribing what constitutes "harm," "violence," and "danger" on these platforms are complex and dynamic and represent entrenched histories of how control over these definitions extend to people's perceptions of their significance and severity. Through these policies, platforms not only delimit their responsibility for what their users post and share, but reinforce certain political ideas and

frames of what constitutes violent or harmful behavior. They govern both discursive and material expressions of harm. Through this governance, they powerfully shape normative notions of harm and violence, effectively managing perceptions of their actions and directing users' understanding of what is "harmful" and what is not.

The purpose of this study is to examine the policies from three major social media platforms (Facebook, Twitter, and YouTube) to understand the ways that they conceptualize, implement, and enforce harmful and violent content. Through the lenses of Bourdieu's (1999) symbolic violence framework and Gillespie's conception of governance of and *by* platforms (Gillespie, 2017), we conduct a critical discourse analysis of how Facebook, Twitter, and YouTube define and police "harm" within their digital milieu. Specifically, we turn our attention to publicly accessible policy documents about misinformation, hate speech, and violent content from these three major platforms to understand how they are conceptualizing "harm" and their practices to mitigate it via human and machine-driven interventions. By open coding documents that substantively address harm and examining their underlying meaning often hidden in language (Fairclough, 1992, 2014; van Dijk, 1993), we tease out not only recurrent or similar practices but the discursive contours of defining and classifying "harmful" content and behavior.

With this analysis, we suggest that platforms' narrow definitions of harm, violence, and danger are not just insufficient, but result in these platforms engaging in ideological hegemony, imposing conceptions of not just *what* violence is and how it manifests, but *who* it impacts and by what mechanisms. We find that rather than changing the mechanisms of their design that enable harm, the platforms reconfigure intentionality and causality through these policy documents in an attempt to stop users from being "harmful," which, ironically, perpetuates harm.

LITERATURE REVIEW

Platform governance

Following the aforementioned events briefly touched upon in the introduction, the influence of these platforms on the political and civic process have received increased scrutiny (Gorwa, 2019). Although a number of interventions have been sought in Europe, the United States, and in other countries to address these issues, the tense relationship between external political forces and internal practices and actions have come into full view (Flew et al., 2019; Ganesh & Bright, 2020; Gillespie, 2018). Although initially portrayed and even marketed as beneficial and democratic technologies, the current political, social, and cultural situation points to these platforms as having deleterious effects on all of these spheres (Donovan, 2020a; Hao, 2021; Hemsley et al., 2018). Of course, the relationship between platform, person, and society is a complex sociotechnical one, and this relationship influences and is influenced by platform governance. How platforms should be governed is a fraught debate, and the term "platform" is in itself a vague and ambiguous term that allows these corporations flexibility in how they define their societal roles and responsibilities (Gillespie, 2010).

Governance, then, is "less a set of practices than a capacity," (Gorwa, 2019, p. 856) and is aligned with platforms' ability to make and enforce rules to moderate the behavior of users. For platforms and online services and communities that predated them, governance is often reflected in content moderation to prevent abuse and to provide mechanisms for enforcement like bans and censorship. In our current digital landscape, governance extends to algorithms, interfaces, as well as terms of service and content policies (Gorwa, 2019; Plantin et al., 2018). Platforms not only engage in governance at this individual user level,

but are political actors that engage and inform their own regulatory frameworks—making platform governance a "specific and complex network of interactions spanning different actors and behaviours" (Gorwa, 2019, p. 856) that includes actors in government, academics, nongovernmental groups, journalists, and many other stakeholders (Dijck et al., 2019).

Platform politics and the politics of platforms are enmeshed in our global politics—intervening in every aspect of ordinary life (Dijck et al., 2019; Gillespie, 2017; Plantin & Punathambekar, 2019). But platform governance becomes a greater issue when we consider the ways that platforms govern *themselves*, leading to the phenomenon that Gillespie noted as "governance *by* platforms" (emphasis added; 2017). The importance of these internal dynamics that then extend outward in the platforms' own policies and participation in initiatives like that of the Global Internet Forum to Counter Terrorism (GIFCT) understandably gives critical scholars pause—these actions not only reinforce platforms' self-governance, but influences users' perceptions of the processes that platforms uses to moderate content and influences their own behavior (Duguay, 2016; Gorwa, 2019; Myers West, 2018). This has not resulted only in perfunctory adherence and belief in platform policies and ability to self govern, referred to as "techlash" (a portmanteau of "technology" and "backlash") (Flew et al., 2019).

The issues with content moderation and policies that govern user content and behavior are global, and the uneven practices of moderation are deeply imbued with discriminatory beliefs regarding gender, race, ability, sexuality, and gender expression (Gerrard & Thornham, 2020; Gray et al., 2017; Zolides, 2020). Adding to this complexity, much of the responsibility of enforcing these policies relies on underpaid contractors (many in the Global South) (Roberts, 2019) and algorithmic moderation due to the sheer "scale" of content shared on these platforms daily (Gillespie, 2020; Gorwa et al., 2020; Zolides, 2020). Governance by platforms entails gendered, racialized, and even sexualized lenses that often harm the very people that these policies are meant to protect (Gerrard & Thornham, 2020). Hiding behind the "free speech" justification and democratic "platform values," platforms often engage in neoliberal interpretations of freedom of speech (Hokka, 2021) and claim to practice "neutrality" in their decisions (Hallinan et al., 2021). In addition, investigative reports and leaked memos from employees of major platforms point to all of these "free expression" values being at best hypocritical and at worst manipulative and negligent. For example, recently revealed internal documents at Facebook showed that high-profile users benefit from "freer" expression than ordinary users as they have been essentially whitelisted in moderation processes (Horwitz, 2021). Earlier reporting similarly documented inconsistency in YouTube's moderation policies—for example, at one time, instructing moderators to take down videos depicting drug-related violence in Mexico, but not violence related to political conflicts in Syria and Russia (Buni & Chemaly, 2016). This begs the question of what purposes moderation serves and who these practices and actions (or lack of action) actively harm through trying to prevent harm? As these platforms become more involved in public activities and everyday life, discussions around platform governance point to the need for "cooperative responsibility" between platforms, governments, and users rather than the current model where responsibility falls on the individual platforms (Helberger et al., 2018). As transnational corporations, the platforms are driven by private commercial interests, and not strictly by their claims of providing a public and democratic service (Gillespie, 2018; Helberger et al., 2018), which should be a key factor in examining their self-directed regulatory frameworks. Platforms' varying practices and actions to curtail "harmful content" reveal inherent issues with relying only on platforms to each govern themselves. These issues may also be visible but in platforms' definitions of what constitutes hate speech, violent content, and other forms of harmful content like mis/disinformation and platform manipulation.

Moderating hate speech, violence, and harm

Companies tend to take three approaches toward moderating content: artisanal (small scale, manual), community-reliant (relying on users to report content), industrial (large scale, often automated), or some combination of all three (Caplan, 2018). With the amount of content that is being posted, having clear definitions of what constitutes hate speech or violent/harmful content is crucial for having an equal moderation process. But this is complicated by factors such as cultural context, different languages, and moderators' inevitable subjectivity (Gillespie, 2018). Hate speech is often defined by hate crime laws in Western legal frameworks, particularly in the United States, that are often damaging and can exacerbate rather than help to alleviate the systemic issues that caused harm in the first place (Schoenebeck & Blackwell, 2021). Online harm, however, is more complicated and complex, taking on multiple forms, and as Schoenbeck and Blackwell noted includes but is not limited to psychological distress, physical violence, oppression and marginalization, and threats to free expression (2021). Harassment, ideological harm, and exploitation are also categories that need to be considered when conceptualizing "harmful content," (Banko et al., 2020). Although they may seem more straightforward, even content that seems "objectively" harmful, like terrorist content, are also abstract and often can fly under the radar of content moderators due to coded language and other ways of maneuvering around censorship (Gerrard, 2018; Murthy, 2021).

Harassment and hate speech are often the most highly contested forms of harmful content despite platforms' attempts to "clearly" define what constitutes harassment and/or hate speech on their platforms (Pohjonen, 2019). Hate speech, in particular, is a difficult term to define. Evidencing its harms in a way that meets a legal burden of proof is a significant problem that plagues its conceptualization and enforcement (Gelber & McNamara, 2016). Similarly, harassment in particular is a "harm" that platforms have long struggled to moderate or meaningfully enforce (Gray, 2020). As Schoenbeck and Blackwell (2021) noted, the focus of these platforms tends to lean towards removing individual content and accounts that violate their policies, but this prioritizes retribution over actual structural changes to the platforms themselves. How social media platforms recognize and repair harm, then, is a significant issue that still persists despite these companies' attempts to control hate speech, harassment, and other kinds of harm.

Although typologizing the kinds of harms that can occur is necessary to moderate it (Banko et al., 2020; Banks, 2010), what platforms often fail to do is meaningfully *repair* what leads to harmful behavior and content in the first place. Moreover, platforms also attempt to assign point values or quantitatively make decisions on what is or is not harmful content, which also can lead to overlooking or outright allowing certain harms to persist if they do not meet a certain numeric threshold, which perpetuates the notion that certain kinds of harms are prioritized over others (Scheuerman et al., 2021). Despite platforms implementing these policies and continuing to update what constitutes hate speech and harassment and their attempts to moderate this kind of content, inconsistencies in these definitions abound within platforms as well as across them (Pater et al., 2016).

Expecting platforms to enforce their own policies, particularly around hate speech and harassment, falls into the issues prevalent in the previous section about regulatory frameworks and platform governance. Despite evidence that the speech that circulates on these platforms leads to both online and offline harms, platforms are largely still left to self-regulate, and in doing so control more speech than any government (Benesch, 2020), often without the knowledge of users (Kalsnes & Ihlbæk, 2020). Due to the extensive global reach and sheer scale of content that these kinds of platforms have to moderate, the question of whether online hate is even governable is a valid one (Ganesh, 2018). Indeed, judging from platform actions and in many cases, inaction, platforms decide what constitutes harm,

violence, and harassment not just in the United States where many are headquartered, but globally.

Symbolic and cultural violence

Beyond physical violence, platforms often perpetuate symbolic and cultural violence. Symbolic violence, as defined by Bourdieu, refers to the ways that power, hierarchies, and inequalities are maintained less by physical force and more by forms of symbolic domination, particularly through language and discourse (Bourdieu, 1999; Morgan & Björkert, 2006). We adopt this framework of symbolic violence to better understand how social media platform policies perpetuate certain notions of what is legitimate and significant harm through the language presented in their own documents, because they are being imposed on users via these policies, and can influence what users perceive or understand to be "real" or "valid" forms of harm (Gray et al., 2017).

Symbolic violence is powerful because it is insidious and invisible, often because it is also internalized and silent (Morgan & Björkert, 2006). Social media platforms perpetuate symbolic violence, in particular, because they are a medium of discourse—they perpetuate systems of knowledge dictating what can and can not be said, and they are much less regulated than traditional communication channels (Recuero, 2015). It could be argued that platforms' consistent self-regulation is in and of itself a form of symbolic violence, because they dominate public discourse globally and control it in ways that are often opaque and unseen (Donovan, 2020b; Suzor et al., 2019). In a way, platforms act like gatekeepers for audiences by determining appropriate and inappropriate content. Platforms, then, control and construct reality in ways that tend to favor certain views and modes of power over others, perpetuating offline -isms and harms in the online world (Gray et al., 2017; Recuero, 2015).

We extend our framework by also incorporating what Galtung referred to as cultural violence, particularly as it refers to "those aspects of culture, the symbolic sphere of our existence … that can be used to justify or legitimize direct or structural violence," (Galtung, 1990, p. 291). As Galtung (1990) notes, cultural violence—similarly to symbolic violence—can make direct and structural violence look and feel natural. Cultural violence contributes to acts of violence that fall within the range of more conventional understandings of violence by "changing the moral color of an act from red/wrong to green/right or at least to yellow/acceptable" (Galtung, 1990, p. 292). Simply put, cultural violence is a helpful additional framework in understanding the insidious nature of symbolic violence, in that it helps us go beyond language to understand how these processes affect the lived experiences of users both online and offline. In the rest of this article, we come from these frameworks of symbolic and cultural violence to examine the ways that platforms conceptualize and govern harm. Further, we consider how these actions perpetuate structural and systemic violence, particularly in how they dominate perceptions and discursive notions of "harm," and how these discursive formations—as policy and enforcement actions—perpetuate cultural violence, and serve as a legitimizing force for direct or structural violence.

METHODS

For this analysis, we collected data in the form of policy documents put forth by three of the major platforms (Twitter, Facebook, and YouTube). Through keyword searches for "harm" and its variants (e.g., violence, hate, etc.), we assembled a corpus of 106 documents across all three platforms (22 for Facebook, 36 for Twitter, and 48 for YouTube). All documents

were collected between March 2021 and July 2021. It is important to note that these documents were retrieved after the January 6 U.S. Capitol Insurrection, which increased scrutiny toward platform policy among other ramifications. Nonetheless, most policy documents were not dated, but some offered a "last updated" date that fell within the range of 2 years.[1] Although platform policies are in a constant state of flux, we believe that this corpus of documents reflects the fundamental beliefs and values that guide platforms' notions of harm and violence that govern platform (in)actions. We included a full list of the policy documents used for this analysis in Appendix A, and refer to specific documents by their assigned number in the findings and discussion.

Using Bourdieu's (1999) symbolic violence framework and Galtung's (1990) concept of cultural violence, we conducted a critical discourse analysis (Fairclough, 2014; van Dijk, 1993) of how the three platforms define and police "harm" within their digital milieu. Per the recommendations from Lazar in conducting a *feminist* critical discourse analysis, we paid close attention to the power dynamics and asymmetries present within the language of the policies themselves (Lazar, 2007). Since we were also interested in how platforms perpetuate symbolic violence, this led us to focus our attention to policy documents about hate speech, harassment, and "violent/harmful content" (defined broadly by the platforms themselves) to understand how the platforms are conceptualizing "harm," and their practices to mitigate it via human and machine-driven interventions.

In particular, we view these policies not merely as governing documents but as discursive acts. Discursive acts, via the language of domination, reproduce certain relations of power, and we follow Fairclough's (2014) recommendations for critical discourse analy in that we not only look at the object of analysis (the policy documents) but also consider the processes in which these objects are viewed and perceived as well as the sociohistorical conditions that govern them (Fairclough, 2014).

To analyze the documents, we began by each independently conducting a close reading of a subset of documents, writing analytic memos on initial observations. In this, we focused on language used around references to "harm" and variant terms (violence, hate, etc.). Following this, we met to discuss our observations. From this discussion we synthesized initial codes to capture descriptive categories of harm observed, as well as discursive framings of harm–namely, the values and ideological presuppositions embedded in linguistic representations of harm in the documents. After coding additional documents based on these initial codes, we conducted a cycle of focused coding to tease out higher-level recurrent or similar practices across platforms that illuminate the discursive contours of defining and classifying "harmful" content and behavior.

FINDINGS

Defining harm

In our analysis, we found that none of the platforms provide a direct definition of harm. Instead, harm is described through types of specific harms or examples of what constitutes harmful content. For instance, Twitter provides a definition and examples for "physical," "psychological," and "informational" harms [TW01]. Facebook relies on the characteristics of harmful actions, including those with the "potential to incite violence" or "impact from bullying/harassment." YouTube offers the narrowest description of harm, seeming to limit potential instances of harm to "physical harm/violence," "death," and "inciting hostility." In Table 1, we attempt to provide brief descriptions of our observations of how these platforms seem to define and conceptualize harm based on our analysis of the policy documents. Of note, Twitter is the only platform that specifically had a clear definition of "harm."

TABLE 1 How platforms define harm

Facebook	Harm is defined by characteristics: "potential to incite violence," "efforts to silence/intimidate others," "civic harms," or "impact from bullying/harassment."
Twitter	Harm is described by its "forms": "physical," "psychological," "informational." Tiered approach to determine the severity of harm.
YouTube	Harm is based and defined by specific examples; as well as economic harm (for creators and demonetization) not much beyond "physical harm," and "death," and "inciting hostility."

While clear and consistent definitions of harm were elusive, we did see a patterned emphasis on normative notions of this concept—namely, an emphasis on child safety, cyberbullying, sexual content (i.e., pornography), and terrorism, but less meaningful enagement with other significant forms of harm and violence on their sites. Platforms also focused on harm as it existed in a legal framework, including but not limited to things like copyright, the sale of regulated goods (firearms, drugs, etc.), and even defamation (in the case of YouTube). Similarly, disinformation about "the civic process" was also included in documents from Twitter and Facebook as a significant harm, which is most likely in response to how these platforms were manipulated to sow discord about elections worldwide (Benkler et al., 2018; Woolley & Howard, 2018).

For all three platforms, harm is not treated as a discrete concept, but is flexibly qualified based on what they view as relevant offenses. As a result, platforms can mold their definitions and descriptions of harm according to any instances they deem it to have occurred. In this way, platforms' use of the terms "harm" and "violence" act as floating signifiers. According to Laclau, and as explained by Farkas and Schou, a floating signifier is "a signifier used by fundamentally different and in many ways deeply opposing political projects as a means of constructing political identities, conflicts and antagonisms," (Farkas & Schou, 2018; Laclau, 2005, p. 300). By avoiding clear and stable definitions of "harm," "hate speech," and "violent content," the platform can carefully navigate opposing political projects invoked by these weighty terms, while also instilling hegemonic notions of what constitutes these ills. And as floating signifiers, these definitions become involved in political and hegemonic struggles in who gets to define what constitutes harm and enforce policies to prevent it—in essence, it is absorbed into the matrix of domination that is language, according to Bourdieu's symbolic violence framework.

Through the ways they position "harm," the platforms also reinforce notions of who and what is important or most deserving of protection via these policies, and through enforcement, they reify a certain political order. Rather than sticking to "fixed" categories constitutive of harm or violence, the platforms seemed to use these terms as concepts that can be molded and interpreted flexibly to fit their needs at any given moment or within a specific context. By rendering "harm" a floating signifier, platforms can respond agiley to public opinion and outcry about emergent concerns around harm, rather than binding themselves to a pre-established definition, which would oblige them to proactively address these emergent concerns. As a linguistic act, the floating signifier of "harm" contributes to the symbolic violence that is so insidious because it exists primarily in language—the way that platforms define harm affects its perception, for both platforms and users. We can see this, for example, in how the platforms have adapted their policies to address public discussion of harms associated with COVID-19 and election misinformation in 2020 (Coppins, 2020; Donovan, 2020a). Moreover, the specific labels the platforms devise for emergent categories of harm also reflect flexibility. For example, all three platforms have in recent years begun referring to "coordinated influence operations" (YouTube), "coordinated behavior" (Facebook), and "coordinated inauthentic activity," as catch-alls for various harms

(e.g., conspiracy theories, election interferences, etc.). These terms are described in fairly abstract language in the documents, which precludes concrete conceptualization.

Hierarchizing harm

We observed an inclination to hierarchize and quantify harm and violence, presumably to accommodate the platforms' technical infrastructures. Operationalizing harm and violence in these ways assists automated tracking, identification and moderation of such content, which helps build towards decreased reliance on and investment in human labor. For example, in a blog post, Twitter described a three-tiered system (low, moderate, high) to classify the severity of "coordinated harmful activity," which emphasizes the *quantity* of documentation of such activity [TW01]. In addition to the quantity, this hierarchy also entails the assumption that *low harms*, if left unchecked, would most likely not cause additional harm, whereas *high harms* would almost certainly lead to additional harm.

Although Twitter utilizes this kind of "hierarchy" for determining the magnitude and severity of harm, Facebook similarly uses "prevalence" as a metric for gauging the magnitude of harm. For example, in a blog post, Facebook explained prevalence, writing: "If a piece of hate speech is seen a million times in 10 min, that's far worse than a piece seen 10 times in 30 min." Such hierarchies provide the platforms with a structured mechanism to manage what they consider harmful content. Evaluated content is often designated to one of these categories, and this categorization helps determine what actions they must take. However, metricizing harm and violence in these ways oversimplifies the complex ways harm manifests and differently impacts different people at different times. Further, it can lead to disproportionate reactions by the platform (either much less or more than what is needed).

And finally, harm is hierarchized based on its potential to "harm" what these platforms name as "protected categories" of individuals and/or groups, further illustrating how "harm" functions as a floating signifier and the ways that harm is racialized and gendered. The platforms categorize many groups of people into these protected categories, namely women, children, racial and ethnic minorities, and religious groups. However, in addition to these protected categories, which mimic protected groups in U.S. law, the platforms also emphasize the distinction between private and public figures. Facebook, for instance, claims they aim to "distinguish between public figures and private individuals because [they] want to allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience" [FB10].

This distinction is enforced by protecting private citizens with additional measures not applicable to public figures. In their policies, they name certain kinds of people (e.g., politicians) as being exempt from some of the platforms' own rules about content, and have admitted to this tiered system of moderation that goes beyond mere moderation but the creation of what Caplan and Gillespie (2020) refer to as "tiered governance." Disinformation about "the civic process" was also included in documents from Twitter and Facebook as a significant harm, which is most likely in response to how these platforms were manipulated to sow discord about elections in the United States and beyond. But the policies that platforms created around "harm" tended to focus on physical, tangible things rather than all the forms that violence can have, and neglected to address these more abstract conceptions of harm.

The material versus symbolic significance of harm

Finally, we saw a discursive positioning of harm as mainly tethered to impacts on the body or individual freedoms, rather than more abstract impacts of domination. For example, the

platforms commonly refer to harm in relationship to things like child sexual exploitation, terrorism, human trafficking, and gore. In this sense, the platforms often defined harm and violence narrowly in terms of threats to physical survival and wellbeing, or what Galtung (1990) called "direct violence," rather than considering how harm and violence exist on a spectrum. Platforms justified their policies primarily through references to physical and psychological impacts. As an example of the former, Facebook wrote in its Dangerous Individuals and Organizations policy: "In an effort to prevent and disrupt real-world harm, we do not allow any organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook" [FB03]. As an example of the latter, YouTube explained in its "Staying safe on YouTube" document: "YouTube is a place where people come to share their story, express an opinion, and engage with one another. We want to ensure creators and viewers feel safe doing so" [YT24]. To a lesser extent, the platforms also justified their policies through reference to law and threats to individual rights and freedoms. For example, in Twitter's Hateful Conduct Policy, the company explained: "Twitter's mission is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers. Free expression is a human right —we believe that everyone has a voice, and the right to use it" [TW02].

While the various kinds of harms the platforms referred to (e.g., sexual harassment, bullying, hate speech) are premised on systems of oppression and legitimated by cultural discourses, the platforms rarely acknowledged this. For example, in Twitter's Sensitive Media Policy, the platform stated: "We prohibit violent sexual conduct to prevent the normalization of sexual assault and nonconsensual violence associated with sexual acts" [TW04]. Here, we see recognition of the link between violent sexual conduct and the normalization of sexual violence, but no acknowledgment of the ways race, gender, class, and other structural inequalities pattern normalization. In this way, sexual violence is abstracted from its political context. While the platforms all routinely make references to special attention to protected groups (primarily children, racial and ethnic minorities, religious groups, and members of the LGBTQ + community), these guidelines seem to respond more to legal imperatives than a commitment to counteracting or undermining systems of power.

The platforms' emphasis on threats to bodies, psychological well being, and individual freedoms, could be read as a means of deflecting attention from the less tangible but more complex symbolic violence perpetrated on their sites (Massanari, 2015; Recuero, 2015). For instance, in a report on "harmful stereotypes," Facebook casually stated that the *direct causality* between such content on the platforms and "real world" violence is "uncertain."[2] Such rhetoric urges the idea that "real" harm does not occur on the platforms and, therefore, they should not be held responsible for it. In other words, they position the most significant harm as occurring "offline" or in the "real world" and implicitly disavow accusations that their sites contribute in any way to upholding white supremacy, sexism, ableism, and so on. Ultimately, the platforms take a pragmatic stance that protects them from legal trouble by bounding their responsibility for harm to surface-level matters which can be easily documented and have a direct link to activity on their sites, perpetuating the framing that they are not responsible for the content posted by its users (Gillespie, 2010).

DISCUSSION

Our above findings suggest a reactive approach by platforms to defining and addressing harm and violence, one that serves the platforms more than society. By sticking close to definitions of harm as (flexibly) visible and tangible, they give the appearance of attentiveness while avoiding controversy. Moreover, a positioning of harm and violence in

terms of quantifiability and physicality suggests a surface-level approach that underplays the impact and ignores the interrelatedness of different forms of harm and violence (physical, emotional, psychological, symbolic) and of violence and power. By rendering "harm" a floating signifier, the platforms preclude stable understanding of what kinds of harmful content they moderate. In some ways, this flexibility is necessary, given that platforms cannot anticipate all possible expressions of harm. However, it also challenges means of holding the platforms accountable because the rules they set are vague and transient. Our analysis included an analysis of only the English-language documentation and policies of these platforms, but we anticipate that many of these findings will translate in other contexts due to the fact that these are U.S.-based companies driven by certain cyberlibertarian ideals (e.g., Barbrook & Cameron, 1996).

As platforms have matured, harms they routinely engender or contribute to have come into focus. Like in the case of the Gamergate harassment campaign, symbolic violence is perpetuated on marginalized communities, particularly women and people of color, within these online and offline spaces (Gray et al., 2017). In these spaces and the technology sector, women and other marginalized groups are often rendered invisible, and face symbolic violence in regular, day-to-day experiences, with racist and sexist language permeating throughout these platforms (2017). Despite attempts by platforms to control hate speech, the very mechanisms by which they attempt to moderate them create sexist assemblages (Gerrard & Thornham, 2020), perpetuating normative and oversimplified ideas about gender identity.

Content moderation, despite its "best intentions," perpetuates inequality and symbolic violence as a result of this necessity of oversimplification that occurs due to reliance on automated moderation. But this means that some harms that are positioned as unforeseeable and unavoidable actually may be discernible and preventable with adequate degrees of attention and resources (Parvin & Pollock, 2020). External accountability for how platforms scope their work related to harm can be facilitated by greater transparency into the evolution of their policy documents. For example, while writing this article, Facebook migrated its community guidelines to a new website, where the company shared multiple versions of the policy documents over time, noting changes. This kind of transparency can help stakeholders better assess the integrity and intent of platform conceptualizations of harm. Though, we should be careful not to overlook the fact that such transparency reports are also usually written by the platforms themselves. As the platforms bound their responsibility for harm (though mutably), they also shift attention away from the most deep-seated and invariant harms, namely those that fall under the category of symbolic violence. As Recuero (2015) argued, platforms have granted symbolic violence "superpowers." Yet, in the policy documents analyzed in this study, this point is carefully buried, which tracks with the platforms' efforts to shape academic and public discourse about their societal impacts (Abdalla & Moustafa, 2021), for example downplaying and obscuring internal reports on political polarization (Horwitz & Seetharaman, 2020) and teen mental health (Seetharaman et al., 2021).

Overall, platforms frame their enforcement as a desire to protect the "well being" of their users. However, the abstract nature of *well being* presents another floating signifier—"well" stands in for a wide range of states. As noted above, these floating signifiers are intangible and hard to define which makes oversight and regulation challenging. In allowing themselves the flexibility to adapt to societal occurrences, platforms position themselves as accommodating public needs. However, such reactive tactics can present a slippery slope when needing to determine a trajectory for overarching wellbeing. These concerns are not assuaged, but rather heightened by platform's *increasing* self-governance, like Facebook's own "Oversight Board," established in 2018, which admitted to their failures to implement platform policies and that Facebook had concealed a number of things from

the public (Nover, 2021). Included in these series of reports was the revelation that after attempts to create a "healthier" community, the platform actually became more toxic because Zuckerberg repeatedly refused proposals to fix these issues (Hagey & Horwitz, 2021).

During the same time period, the CEO of Instagram Adam Mosseri appeared in a podcast and compared social media platforms to cars, saying: "Cars have positive or negative outcomes… We know that more people die than would otherwise because of car accidents. But by and large, cars create way more value in the world than they destroyed. And I think social media is similar," (Kafka, 2021). Facebook, in a similar vein, rather than correcting the harms that the platform has produced and strengthened, is opting to defend its image by pushing "Facebook-positive" stories to users' feeds (Mac & Frenkel, 2021). These acts of blatant dismissal and defense, and statements like Mosseri's, demonstrate the ways that, for platforms, defining "harm" represents an actuarial equation—the principal concern is harm to their profits and public image rather than harm to the possibility of a just society. The violence perpetuated in Mosseri's analogy justifies harm to serve an abstract "greater good," forcing people to accept the platforms' lack of moderation as a natural and unchanging reality, perpetuating the symbolic violence of social media. Yet, it is unclear whom this greater good serves.

Mosseri's quote also highlights the ways that the rhetoric of "unintended consequences" is used to defend the consequences of these technologies, and these phenomena are often deemed too difficult, costly, or complicated to deal with until they become problems for others (Parvin & Pollock, 2020). Facebook, similarly, is doubling down and refusing to apologize further but rather going on the defensive with its own PR strategy, as mentioned in the previous section. To claim that all of these platform ills are the result of "unintended consequences" further absolves the platforms of any blame (Gillespie, 2010; Parvin & Pollock, 2020) and makes it appear as if they were unaware of the adverse effects, which is far from the truth. By deflecting blame, platforms can continue business as usual, and not face any real accountability for the harms that they continue to produce despite their claims that they were attempting to correct these "harms" facilitated and afforded by their services. As a result, platforms get to control the language used and perception of "harm" on their services, rationalizing these harms. Symbolic violence is what occurs as a result, and upholds systemic and structural violence that occurs in offline worlds. By applying the framework of symbolic and cultural violence, we hope to demonstrate how an analysis of platform policies' language can continue to shed light on the ways that they perpetuate inequality and marginalization, and how it shapes social media governance.

Implications for policy and possible solutions

The formulation of harm-related policies put forth by platforms provides insights into their ideologies and philosophies surrounding their own role in the media ecosystem and society at large. As we observed above, platforms primarily seem to take reactive measures rather than pre-emptive action and their approach belies a central interest in self-preservation above repairing systemic harms. As noted by Schoenebeck and Blackwell, this is akin to Western frameworks of criminal justice that focus on identifying perpetrators of harm and punishing them, and which largely "overlook the needs and interests of targets of harassment and remove offenses and offenders from the community without any attempt at rehabilitation (Schoenebeck & Blackwell, 2021, p. 14). Such a paradigm equates unintentional rule-breaking with intentional acts of harm and leaves no space for reeducation, rehabilitation, or forgiveness. Moreover, it draws attention away from attempts to understand the impact of harm, what those experiencing harm might need, and how platforms could revise their systems accordingly (2021).

Similarly Baker et al. (2020) note the need to acknowledge the shortcomings of algorithms and automated processes while addressing the uncertainty of science and real-time data. They observe that, conceptually, *harm* is not neutral and content moderation can be manipulated to propagate values or cultivate doubt, suggesting a need to move beyond removal of *harmful content* on the basis of official advice (2020). Another possibility is through a restorative form of justice which includes "mediated conversations between those who perpetrate and those who experience harm, typically with mediators and community members actively participating" (Schoenebeck & Blackwell, 2021). Such a system includes perpetrators by having them acknowledge and express remorse for wrongdoing. Certainly, community-based restorative justice is more time-intensive, but would produce better outcomes by centering those experiencing harm, which would re-focus energy on more meaningfully addressing "harm" on these platforms. Policies that would come out of a restorative justice framework would require imagination, commitment, and accountability - as well as require platforms to no longer deflect blame by identifying their platform ills as "unintended."

And finally, platforms cannot self-regulate to the extent that they do currently, and must be more willing to share raw data with researchers—not strategically scoped or cleaned data, in the case of Facebook (Lyons, 2021)—allow external audits and investigations, and more government regulation of these platforms is needed. However, it must be acknowledged that governments and the state are holders and perpetrators of violence themselves, and so relying on them to regulate these platforms may further exacerbate cultural, symbolic, and material violence. As a result, it is not only important but necessary to coconstruct platform governance models through participatory action research, where the communities who are most harmed by these platforms contribute to research and policy interventions. Advocacy groups, academics, and journalists can help with highlighting and making others aware of these issues, giving a voice to these often overlooked but purportedly "protected" groups.

These suggestions are but a few in the wider area of platform governance research, and our analysis adds to these conversations by examining the policies of these platforms and how accountability is made elusive in the policies themselves. These discursive positionings perpetuate ideologies that reify symbolic and cultural violence, which often can and does become physical violence as we have seen historically and also in our current sociopolitical moment. Platforms can no longer hide behind the flexibility and slipperiness of these moderation policies to avoid doing anything more meaningful. In conclusion, they must be truly and radically transparent, and no longer operate "in the shadows" (Donovan, 2020a, 2020b) or deflect criticism and create policies through their own regulatory frameworks and "oversight boards." These pursuits can no longer be legitimized, and must consider ethical questions to guide their design rather than having "ethics" serve as a regulatory practice that is used to adjust their services after-the-fact (Parvin & Pollock, 2020). Platforms must consider restorative justice and ethics to be at the heart of their design rather than as ad-hoc elements, and no longer marginalize these kinds of frameworks and practices.

ENDNOTES

[1] After data collection ceased, Facebook re-launched its policy documents on its transparency website (transparency.fb.com), making multiple versions of the policy documents available by date.

[2] https://about.fb.com/wp-content/uploads/2018/11/PPF_08.11.2020_Harmful-Stereotypes.pdf

REFERENCES

Abdalla, M., & Moustafa, A. (2021). The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity. *arXiv*, *2009*, 3676v4.

Arendt, H. (1970). *On violence*. Houghton Mifflin Harcourt.

Baker, S. A., Wade, M., & Walsh, M. J. (2020). The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. *Media International Australia*, *177*(1), 103–107.

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, *24*(3), 233–239.

Barbrook, R., & Cameron, A. (1996). The Californian Ideology. *Science as Culture*, *6*(1), 44–72.

Benesch, S. (2020). But Facebook's not a country: How to interpret human rights law for social media companies. *Yale Journal on Regulation*, 38,1–26.

Buni, C., & Chemaly, S. (2016). The secret rules of the internet. *The Verge*. https://www.theverge.com/2016/4/13/11387934/internet-moderatorhistory-youtube-facebook-reddit-censorship-free-speech

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bourdieu, P. (1999). *Language and symbolic power*. Reprint edition. Edited by J. Thompson. Translated by G. Raymond and M. Adamson. Harvard University Press.

Caplan, R. (2018). *Content or context moderation?* (pp. 1–37). Data & Society. https://datasociety.net/library/content-or-context-moderation/

Caplan, R., & Danah, B. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 205395171875725.

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 1.

Coppins, M. (2020). The billion-dollar disinformation campaign to reelect the president. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2020/03/the-2020-disinformation-war/605530/

Daniels, J. (2018). The algorithmic rise of the "Alt-Right". *Contexts*, *17*(1), 60–65.

Dijck, J., van, Nieborg, D., & Poell, T. (2019). Reframing platform power. *Internet Policy Review*, *8*(2). https://policyreview.info/articles/analysis/reframing-platform-power

Donovan, J. (2020a). Social-media companies must flatten the curve of misinformation. *Nature*.https://www.nature.com/articles/d41586-020-01107-z

Donovan, J. (2020b). *Why social media can't keep moderating content in the shadows*. MIT Technology Review. https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/

Duguay, S. (2016). Trending this moment: Examining social media platforms as information gatekeepers through Facebook's Trending topics and Twitter's Moments. In *66th Annual Conference of the International Communication Association: Communicating with Power.*

Fairclough, N. (1992). *Discourse and social change*. Wiley.

Fairclough, N. (2014). *Language and power.* (3rd ed.). Routledge.

Farkas, J., & Schou, J. (2018). Fake news as a floating signifier: Hegemony, antagonism and the politics of falsehood. *Javnost—The Public*, *25*(3), 298–314.

Fischer, S., & Gold, A. (2021). All the platforms that have banned or restricted Trump so far. Axios. https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50.

Galtung, J. (1990). Cultural violence. *Journal of Peace Research*, *27*(3), 291–305.

Ganesh, B. (2018). The ungovernability of digital hate culture. *Journal of International Affairs*, *71*(2), 30–49.

Ganesh, B., & Bright, J. (2020). Countering extremists on social media: challenges for strategic communication and content moderation. *Policy & Internet*, *12*(1), 6–19.

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511.

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*.

Gillespie, T. (2010). The politics of "platforms". *New Media & Society*, *12*(3), 347–364.

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

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 205395172094323.

Gillespie, T. (2017). Governance of and by platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE.

Giroux, H. A. (2017). War culture and the politics of intolerable violence. *Symplokē*, *25*(1-2), 191–218.

Gorwa, R. (2019). What is platform governance?. *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794.

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

Gray, K. L., Buyukozturk, B., & Hill, Z. G. (2017). Blurring the boundaries: Using Gamergate to examine "real" and symbolic violence against women in contemporary gaming culture. *Sociology Compass*, *11*(3):e12458.

Hagey, K., & Horwitz, J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal*, 15 September. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Hallinan, B., Scharlach, R., & Shifman, L. (2021). Beyond neutrality: Conceptualizing platform values. *Communication Theory*. https://doi.org/10.1093/ct/qtab008

Hao, K. (2021). How Facebook got addicted to spreading misinformation. *MIT Technology Review*. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Hemsley, J., Jacobson, J., Gruzd, A., & Mai, P. (2018). Social media for social good or evil: An introduction. *Social Media +Society*, *4*(3), 205630511878671.

Hokka, J. (2021). PewDiePie, racism and Youtube's neoliberalist interpretation of freedom of speech. *Convergence*, *27*(1), 142–160.

Horwitz, J. (2021). Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. *Wall Street Journal*, 13 September. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

Horwitz, J., & Seetharaman, D. (2020). Facebook executives shut down efforts to make the site less divisive. *Wall Street Journal*, 26 May. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Kafka, P. (2021). Instagram boss Adam Mosseri on teenagers, Tik-Tok and paying creators. (Recode Media). https://www.vox.com/recode-media-podcast

Kalsnes, B., & Ihlbæk, K. A. (2020). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society*.

Laclau, E. (2005). *On populist reason*. Verso.

Lazar, M. M. (2007). Feminist critical discourse analysis: Articulating a feminist discourse praxis. *Critical Discourse Studies*, *4*(2), 141–164.

Lyons, K. (2021). Facebook reportedly provided inaccurate data to misinformation researchers. *The Verge*. https://www.theverge.com/2021/9/11/22668396/facebook-provided-inaccurate-data-misinformation-researchers

Mac, R., & Frenkel, S. (2021). No more apologies: Inside facebook's push to defend its image. *The New York Times*, 21 September. https://www.nytimes.com/2021/09/21/technology/zuckerberg-facebook-project-amplify.html

Massanari, A. (2015). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Morgan, K., & Björkert, S. T. (2006). 'I'd rather you'd lay me on the floor and start kicking me': Understanding symbolic violence in everyday life'. *Women's Studies International Forum*, *29*(5), 441–452.

Murthy, D. (2021). Evaluating platform accountability: Terrorist content on YouTube. *American Behavioral Scientist*, *65*(6), 800–824.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383.

Nover, S. (2021). *What if Facebook is lying to its own Oversight Board? Quartz*. https://qz.com/2064919/what-if-facebook-is-lying-to-its-own-oversight-board/

Odag, Ö., Leiser, A., & Boehnke, K. (2019). Reviewing the role of the Internet in radicalization processes. *Journal for Deradicalization*, *21*, 261–300.

Parvin, N., & Pollock, A. (2020). Unintended by design: On the political uses of "Unintended Consequences". *Engaging Science, Technology, and Society*, *6*(0), 320–327.

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of online harassment: comparing policies across social media platforms. *Proceedings of the 19th International Conference on Supporting Group Work New York, NY, USA: Association for Computing Machinery (GROUP ' 16)*. pp. 369–374.

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310.

Plantin, J.-C., & Punathambekar, A. (2019). Digital media infrastructures: Pipes, platforms, and politics. *Media, Culture & Society*, *41*(2), 163–174.

Pohjonen, M. (2019). Extreme speech|A comparative approach to social media extreme speech: Online hate speech as media commentary. *International Journal of Communication*, *13*(2019), 16.
Recuero, R. (2015). Social media and symbolic violence. *Social Media + Society*.
Roberts, S. T. (2019). *Behind the screen*. Yale University Press. https://yalebooks.yale.edu/book/9780300235883/behind-screen
Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A framework of severity for harmful content online. *arXiv*.
Schoenebeck, S., & Blackwell, L. (2021). *Reimagining social media governance: Harm, accountability, and repair*. Social Science Research Network.
Schwartz, M. S. (2019). Facebook Bans Alex Jones, Louis Farrakhan and Other 'Dangerous' Individuals, *NPR.org*. https://www.npr.org/2019/05/03/719897599/facebook-bans-alex-jones-louis-farrakhan-and-other-dangerous-individuals
Seetharaman, G. W., Horwitz, J., & Deepa, S. (2021). Facebook knows instagram is toxic for teen girls, company documents show. *Wall Street Journal*, 14 September. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.
Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.
van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, *4*(2), 249–283.
Woolley, S. C., & Howard, P. N. (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.
Zolides, A. (2020). Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society*, *23*(10), 2999–3015.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: DeCook, J. R., Cotter, K., Kanthawala, S., & Foyle, K. (2022). Safe from "harm": The governance of violence by platforms. *Policy & Internet*, 1–16. https://doi.org/10.1002/poi3.290