SAFE FROM “HARM”: THE GOVERNANCE OF VIOLENCE BY PLATFORMS

Julia R. DeCook  
Loyola University Chicago

Kelley Cotter  
Pennsylvania State University

Shaheen Kanthawala  
University of Alabama

Introduction

Since the early days of the COVID-19 pandemic, and more recently, the January 6 U.S. Capitol Insurrection, a number of issues have emerged in regard to how platforms moderate and mitigate “harm” on their services. Although in recent years platforms have developed more explicit policies in regard to what constitutes “hate speech” and “harmful content,” the unintended consequences and side effects about how platforms define harm and how those definitions, in turn, affect users’ normative understandings of harm are understudied. Namely, 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,” “violence,” and “danger” within these platforms are complex and dynamic and represent entrenched histories of how control over these definitions extend to people’s perceptions of them (Arendt, 1970; Bourdieu, 1999).

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. Through this governance, they continue to control normative notions of harm and violence, effectively managing perceptions of their actions (Gillespie, 2017) and directing users’ understanding of what is “harmful” versus what is not. 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.

Data Collection and Method

We collected data in the form of public facing documentation (such as blog, policy, and help page documents) put forth by three of the major platforms (Twitter, Facebook, and YouTube)

that mentioned “harm” and its variants (e.g. violence, hate, etc.). Through keyword searches we
assembled a corpus of 270 documents across all three platforms.

Using Bourdieu’s symbolic violence framework (1999) and feminist technoscience critiques of
“unintended consequences,” (Parvin and Pollock, 2020) we conduct a critical discourse analysis
of how the three platforms define and police “harm” within their digital milieu. We paid close
attention to the power dynamics and asymmetries present (Lazar, 2007), focusing our attention
to policy documents about misinformation and hate speech to understand how the platforms are
categorizing “harm”, and their practices to mitigate it via human and machine-driven
interventions. Through an iterative process of open coding the documents that mention harm,
we tease out not only recurrent or similar practices but the discursive contours of defining and
classifying “harmful” content and behavior.

Initial Findings and Implications

There are three key findings from our initial review of platform policy documents. First, we
characterize the platforms’ use of the terms “harm” and “violence” as floating signifiers
(Mehlman, 1972). 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. As new harms emerge and
public outcry reaches a tipping point, the platforms adapt their policies accordingly, as in the
case of COVID-19 and election misinformation in 2020 (Coppins, 2020; Donovan, 2020). This is
apparent in the labels the platforms devise for emergent categories of harm. 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., QAnon, election interferences, etc.). While none of the three
platforms offered a clear or consistent definition of harm or violence, we did see a patterned
emphasis on normative notions of these concepts--namely, an emphasis on child safety,
cyberbullying, and terrorism, but less meaningful engagement with other significant forms of harm
and violence on their sites.

Second, 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, medium, high) to classify the severity of
“coordinated harmful activity,” which emphasizes the quantity of documentation of such activity.
Similarly, Facebook repeatedly refers to “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 minutes, that’s far worse than a piece seen 10 times in 30
minutes.” Metricizing harm and violence in these ways oversimplifies the complex ways harm
manifests and differently impacts different people at different times.

Finally, we saw a discursive positioning of harm and violence as physical and as something that
exists outside of the platforms in the “offline” or “real world.” In this sense, the platforms seemed
to be defining harm and violence narrowly in terms of materiality. Indeed, as mentioned, the
platforms consistently referred to child sexual exploitation and/or terrorism as prototypical
examples of harmful content. Such an emphasis could be a means of deflecting attention from
the less tangible but no less real psychological, emotional, and symbolic violence perpetrated
on their sites (Massanari, 2017; 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.” Such rhetoric urges the idea that “real” harm does not occur on the platforms, and, therefore, they should not be held responsible for it.

Our above findings suggest a reactive approach by platforms to defining and addressing harm and violence. By sticking close to normative understandings of harm, 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.

References


