SOCIAL MEDIA, MASS ATROCITIES, AND ATROCITY PREVENTION

2023 Sudikoff Interdisciplinary Seminar on Genocide Prevention

BACKGROUND PAPER

UNITED STATES
HOLOCAUST
MEMORIAL
MUSEUM

SIMON-SKJODT CENTER
FOR THE PREVENTION OF GENOCIDE

THE UNITED STATES HOLOCAUST MEMORIAL MUSEUM teaches that the Holocaust was preventable and that by heeding warning signs and taking early action, individuals and governments can save lives. With this knowledge, the **Simon-Skjodt Center for the Prevention of Genocide** works to do for the victims of genocide today what the world failed to do for the Jews of Europe in the 1930s and 1940s. The mandate of the Simon-Skjodt Center is to alert the United States' national conscience, influence policy makers, and stimulate worldwide action to prevent and work to halt acts of genocide or related crimes against humanity, and advance justice and accountability. Learn more at **ushmm.org/genocide-prevention.**

DANIEL SOLOMON, Associate Research Fellow at the Simon-Skjodt Center for the Prevention of Genocide.

TALLAN DONINE, Research Assistant at the Simon-Skjodt Center for the Prevention of Genocide.

COVER: iPhone in Hand on Screen Group of Social Media apps and Notification Badges. *Cristian Dina / Alamy Stock Photo*

CONTENTS

INTRODUCTION

In 2018, anti-Muslim extremists in Sri Lanka organized a series of attacks against Muslim civilians throughout the country (Taub and Fisher 2018). Extremist leaders used a misleading viral video to stoke fears that the country's Muslim minority was organizing a campaign to sterilize the majority Sinhalese population en masse. The video circulated widely on Facebook, and participants in the violence also set up private WhatsApp groups to coordinate the violence.

This is just one example of a disturbing pattern that is increasingly under the spotlight: social media seeming to fuel violence, including large-scale and deliberate attacks on civilians based on their identity. These examples have become widespread in recent years, with attacks on the Rohingya community in Burma, the Muslim community in India, and multiple marginalized groups in Ethiopia and elsewhere following a similar trajectory. Influential users appeared to use social media in these cases to dehumanize their targets, recruit participants, and coordinate violence.

Amid public outcry social media firms have begun to develop or revise some policies to limit the spread of harmful content. In the Sri Lanka case, Meta—which operates both Facebook and WhatsApp—acknowledged and apologized for the platforms' role in the violence (Facebook 2020), issued a series of human rights impact assessments of its role, and stated it would implement policies to address associated human rights concerns from its products, with applications to other high-risk contexts (Sissons 2020).

Across social media companies, responses to concerns about "offline harm" have largely focused on content moderation, modifying algorithms that promote different kinds of content, and restricting access to certain users. Many advocates argue that actions to date have not gone far enough, asserting that stronger action—including government regulation—is necessary to prevent social media from contributing to violence, including mass atrocities.

Social media use will only continue to rise, especially in countries at high risk of new mass atrocities. Beyond select cases, however, there is insufficient research about the specific relationship between social media platforms and the onset and escalation of mass atrocities—or how social media companies and government actors might limit these platforms' potential negative effects.

The goal of this paper is to stimulate and frame discussion during the Sudikoff Interdisciplinary Seminar on Genocide Prevention about the relationship between social media technologies and the risk and prevention of mass atrocities. Based on a review of relevant research, policy documents, and public statements by social media companies, the paper surveys current knowledge and identifies important gaps in understanding about (1) how social media platforms have contributed to the risk and occurrence of mass atrocities in the past and how they might do so in the future;[1] and (2) strategies to help prevent social media from fueling mass atrocities.

KEY CONCEPTS AND DEFINITIONS

Mass atrocities refer to “large-scale, systematic violence against civilian populations” (Straus 2016, 31). This conceptual definition overlaps with most instances of genocide and crimes against humanity, and some war crimes, as defined in international law.

Atrocity prevention, as defined by Straus (2016), is “the effort to prevent, contain, and/or mitigate violence against non-combatants either in or out of conflict.” Atrocity prevention can draw on a wide range of strategies and tools available to the governments–including diplomacy, foreign assistance, defense cooperation, and military action–and non-governmental actors.

Social media consist of digital platforms on which users generate and interact with information in textual, audio, visual, or hybrid formats, often described as “content” (Carr and Hayes 2015). Two fundamental characteristics distinguish social media from other communication channels or platforms.

- First, social media platforms rely on users—as opposed to paid professionals—to create media material. These platforms differ from media platforms such as newspapers, radio stations, or television stations or streaming services, in which corporate entities such as media editors, advertising professionals, or government communications officials exercise editorial control over content.
- Second, social media platforms allow users to communicate and interact with each other. This interactive communication involves multiple different formats, including textual, audio, or visual responses such as comment threads on a message board; emojis such as the Facebook Reactions button; and message forwarding such as retweets on Twitter.

Although several characteristics distinguish different social media platforms, we focus on two key user-facing attributes that are especially relevant to the link between social media and mass atrocities (Kietzmann et al. 2011). The first, *content sorting*, determines the media that users view and with which they interact. On some social media sites or applications, the platform sorts the content chronologically or thematically. These include messaging platforms such as WhatsApp, in which content only appears in the order in which users receive messages. On others—especially those whose business models rely on advertising revenue—platforms use statistical algorithms to promote content to users with specific marketing profiles or behaviors on the platform. The second, *platform access*, determines the other users and content with which users may interact. While some social media applications such as WhatsApp allow users to interact only with other users or closed groups in which they are members, others such as Twitter and the Chinese social media site Sina Weibo, feature open feeds or posts with which a broader universe of users may interact. Platforms with different access rules lead information to spread through different network "nodes" and channels.

SOCIAL MEDIA PLATFORMS AND MASS ATROCITIES[2]

Social media and atrocity risk factors

The research literature about mass atrocities identifies two main risk factors that social media may influence: (1) the presence of violent conflict or large-scale instability and (2) exclusionary ideologies (Straus 2016).[3]

Conflict and instability: The strongest predictor of mass atrocities is the presence of armed conflict. In war, groups have increased capacity to commit violence, the rule of law may be overlooked or suspended, and group leaders may have incentives to attack civilian populations. Similarly, in periods of major political instability, leaders may resort to extreme measures to obtain or retain power, or new leaders with violent ideologies may assume control.

Social media might influence the risk of violent conflict or political instability through a number of mechanisms, three of which we highlight below:

- Promoting polarization: Social media users tend to distribute information via self-perpetuating "echo chambers," which separate online groups over time and reinforce ideological distance between users with opposing views or identities (Del Vicario et al. 2016). To the extent these patterns of online polarization translate to offline relationships (Jones et al. 2013), increased polarization may contribute to a country's risk of large-scale conflict (Montalvo and Reynal-Querol 2005). Although almost all studies of the polarizing effects of social media center on the United States and, to a lesser extent, European democracies, these effects may be similar—or even larger—in countries at high risk of atrocities where polarization between groups is well-established (Barbera 2020).

- Coordinating protest and/or rebellion: Social media platforms may provide groups who seek to organize protests or rebellion against the state with a more effective or lower-cost means of coordinating their activities than non-digital platforms (Zeitzoff 2017). Evidence from multiple country contexts indicates that social media use is associated with higher levels of protest activity (Zhuravskaya, Petrova, and Enikolopov 2020). Although these actions are consistent with norms of free expression and protected under international law, rulers or security forces may perceive protests as threatening. Research on information-communication technologies (ICTs) and violent conflict provides some indication that mobile phone access—the principal means of accessing social media platforms in most contexts—is linked to violence, but the evidence is mixed and lacking as regards social media specifically (Zeitzoff 2017).

- Enabling repression: Even as social media lower barriers to participation in protests and violent rebellion, these platforms also provide governments with new tools for surveillance and repression (Earl et al. 2022) by making transparent information about a large universe of individual and collective beliefs that may lead to dissent (e.g., King et al. 2013). Despite the possibility of social media-enabled repression, the implications for mass atrocity risks are ambiguous: If governments succeed in discouraging the online coordination of protest or rebellion, risks of large-scale, systematic violence against civilians may be lower even as these successful repression strategies intensify other human rights violations (Gohdes 2020). On the other hand, risks of mass atrocities could be higher if governments use social media to identify and monitor the activities of targeted groups. Digital repression can also accompany large-scale violence: evidence from the early months of the Syrian civil war suggests that "digital repression" such as Internet blackouts occurred alongside large-scale physical violence (Gohdes 2015).

Exclusionary ideologies: The presence of an exclusionary or transformative ideology increases the risk of mass atrocities by providing perpetrators with a narrative that defines violence against specific groups of civilians as a "strategically and morally justifiable course of action" (Maynard 2022, 59; Kim 2018). These ideologies are especially important during political crises, when perpetrators use narratives to rally support for atrocities against political opponents from security institutions, members of the ruling elite class, and the general public (Straus 2015a; Nyseth Brehm 2016; Maynard 2022).

We highlight two main ways that social media may influence the relationship between ideology and mass atrocities:

- Normalizing violence: Like other forms of mass media (Straus 2007), social media provides perpetrators with an effective platform for disseminating information that justifies mass violence against targeted groups (Siegel 2020). Social media may normalize violence by encouraging the use of hateful language against targeted groups or spreading myths that promote resentment (Hook and Verdeja 2022). This speech may desensitize the general population against violence and worsen social stigmas against potential targeted groups (Staub 1989). The Burma case illustrates how social media may normalize group-targeted violence: anti-Rohingya extremist groups used social media to spread hate speech and encourage acquiescence to, and popular support for, the Burmese government's campaign of targeted violence, especially in Rakhine State (Liebowitz 2019).

- Directly inciting and encouraging participation in violence: Social media may play a more direct, individualized role in the relationship between exclusionary ideologies and mass violence by encouraging participation in violence, similar to the well-documented effects of the Radio Télévision Libre des Mille Collines broadcasts on participation in the Rwandan genocide

(Yanagizawa-Drott 2014). In these circumstances, regimes and groups use social media to recruit violent foot soldiers and broadcast information about informal violent activities (Dangerous Speech Project 2021). These direct effects are most likely to influence participation at a mass scale where social media provides a more "cost effective" means of encouraging participants in violence than other mobilization strategies.

Social media and atrocity triggers

Social media platforms may increase the likelihood or intensity of atrocity "triggers," which are events that catalyze the onset of atrocities or a sharp uptick in the scale, lethality, or systematicity of violence (Straus 2015b; Valentino 2016). Triggers may exert their effects by sharply increasing incentives for perpetrators to escalate attacks against their opponents or by providing hawkish elites new opportunities to encourage and promote participation in atrocities (Straus 2015b).

In addition to instances of conflict escalation, loss, or regime transitions (Valentino 2016), "symbolically significant violations," such as attacks on religious sites or transgressive behavior by a member of the targeted group (Straus 2015b), can be triggers of mass atrocities. Social media platforms may spread violence-triggering information about symbolically significant violations—whether by deliberate design of leaders or without their explicit intervention. In 2020 in Ethiopia, for example, false social media posts about the political allegiances of a popular Oromo musician preceded his assassination by an Oromo nationalist group, leading to a wave of targeted violence in Addis Ababa and the Oromia region (Gilbert 2020).[4] A simulation-based study suggests that "violence-promoting" rumors by popular extremist leaders are especially likely to spread among the broader population rather than fizzle out (Bhavnani, Findley, and Kuklinski 2009), as the anti-Rohingya leader Ashin Wirathu's influential rhetoric before and during the genocide against the Rohingya in Burma illustrates (Taub and Fisher 2018; Liebowitz 2019).

Key research gaps

Although a large and growing number of studies document the relationship between social media and discrimination, polarization, and violence in general, our review highlighted two main gaps about the relationship between social media and mass atrocities that merit additional research. First, as Barbera (2020) observes in the context of work on polarization, research on the interaction between social media and potential drivers of violence focuses overwhelmingly on the United States and a limited set of European countries. This leaves a significant gap in research about how social media contributes to polarization, contention and repression, and the spread of violence in higher-risk environments. This gap is especially pronounced for research about organization- and macro-level dynamics such as the capacity of government and non-state actors to mobilize violence.

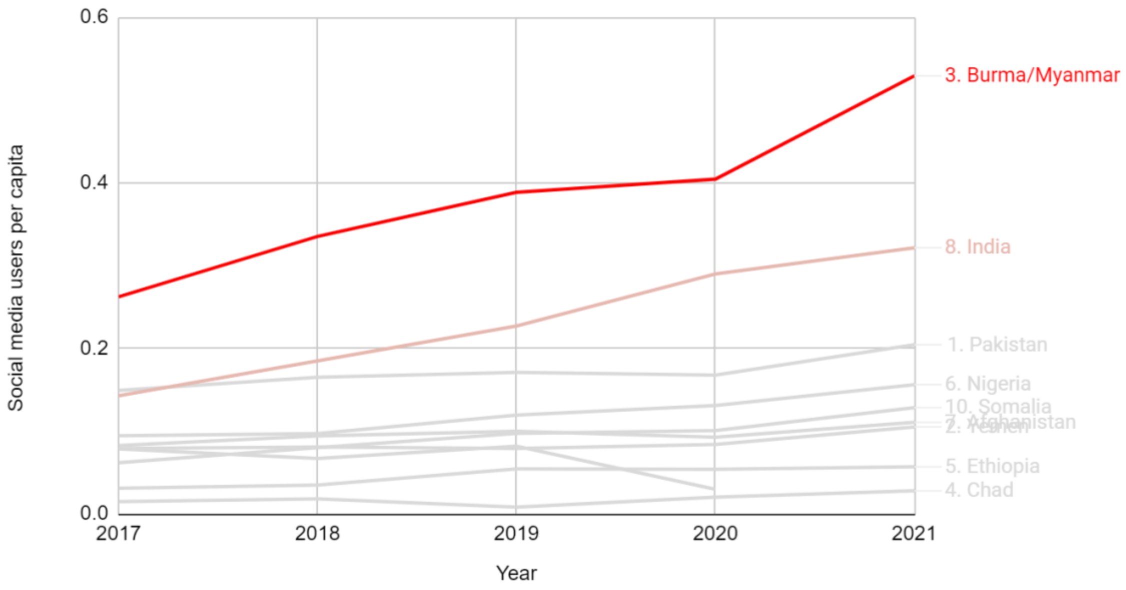
Second, research about the effects of social media on atrocity risks has typically focused on US- and EU-based platforms, with less work that examines the effects of platforms such as Sina Weibo and the Russian platform VK. Although researchers examining online hate speech (Rini et al. 2020) and nonviolent collective action have studied these platforms (Zhuravskaya, Petrova, and Enikolopov 2020), their effects on group-targeted violence and other atrocity risks merit additional attention.

Social media in countries at high risk for mass atrocities

In the last decade, social media platforms have experienced significant growth in their global user base and the content that these users produce (Ortiz-Ospina 2019). Users upload an estimated 3.7 million videos to YouTube every day (Hayes 2022), and more than one billion "stories" are posted daily on Facebook apps (Martin 2022), making these platforms two of the most active globally (DataReportal 2022).

However, this dramatic growth in use of social media has not occurred in many countries at relatively high risk of mass atrocities, based on existing risk assessment models. Figure 1 shows the trends in social media users from 2017–22 in the ten countries that ranked highest in the Early Warning Project's (EWP) 2022–23 assessment of countries' risk of experiencing a new mass killing event.[5] The chart suggests that the relative number of social media users in Burma and India—ranked third and eighth on the EWP assessment, respectively—has grown rapidly, to 52 percent and 32 percent of the country's population in 2022, respectively, while the number of users in the other eight countries has not.[6] For comparison's sake, social media users in the United States in 2022 amounted to 81 percent of the US population. If the impact of social media is greater where its use is more widespread, these different trends in social media use across cases suggest that the largest effects of social media uptake on mass atrocity risks may be felt only in the future for most high-risk countries.

Figure 1: Social media users in top-10 SRA countries



Although greater numbers of users or rapid changes may intensify the negative effects of social media (e.g., Whitten-Woodring et al. 2020), widespread use is not a prerequisite for these effects. For example, a recent International Crisis Group report on Chad—currently ranked fourth in the world for risk of mass killing[7]—shows that social media's use and perceived relevance for political discourse is growing at a critical time in the country's tense transition period (Crisis Group 2022), despite a consistently small number of social media users per capita in the country. Observers forecast that social media will play an increasing role ahead of the country's elections scheduled for 2024, with early indications that it could propel unrest and ethnic polarization in the absence of proactive action by social media companies, among other actors (Crisis Group 2022).

POTENTIAL STRATEGIES TO ADDRESS ATROCITY RISKS

This section discusses current efforts and new proposals to address risks from social media platforms divided into four categories, focusing on (1) content moderation; (2) modifications to the system or algorithm that promotes different kinds of content; (3) restrictions on user access; and (4) regulatory incentives associated with overall firm policies.

Content moderation

The conversation surrounding efforts to reduce harmful outcomes connected to social media has largely involved suggestions for expanding content moderation (Kornbluh and Goodman 2019).[8] Content moderation refers to "the organized practice of screening user generated content … posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction" (Roberts 2017a). As it relates to atrocity risks, advocates suggest that content moderation could prevent speech that would otherwise contribute to mass atrocities from circulating on social media platforms. The core challenge, therefore, is identifying, with sufficient accuracy and speed, content that would contribute to mass atrocity risks. This task is especially difficult because it requires assessing the speaker, the audience, and the context, in addition to the message itself (Dangerous Speech Project 2021).

In recent years, social media companies have taken some steps to discourage hate speech and restrict the use of social media by violent actors. These include:

Reporting content: Social media platforms generally allow users to report or flag content they perceive as violating its rules or that they find personally offensive (Common 2020). Users play a fundamental role "since almost every content moderation system depends on users flagging content and filing complaints" (Buni and Chemaly 2016). However, these systems do not always work to prevent negative effects. In Sri Lanka, where harmful content fueling targeted violence against Muslims circulated quickly on Facebook, advocates reported content clearly demonstrating calls for violence; and yet, Facebook ruled many of these posts were not in violation of its standards (Taub and Fisher 2018). Generally, these reporting systems are not widely used, the review process is usually not transparent, and social media users can manipulate the review process for political or personal reasons (Gillespie 2018). In addition to review by staff or contractors employed at social media firms, social media platforms such as Facebook and Instagram use an appeals process that allows an independent Oversight Board to perform this content-review function for a limited set of content (Instagram 2022).

Human and automated moderation: Social media platforms have sought to regulate content violating their policies through a combination of human content moderators and automated technologies. Major social media firms hire or contract human moderators to determine whether user-generated content violates their community standards.[9] Given the massive and ever-expanding quantity of social-media content and the economic costs associated with human moderators, companies devote increasing attention to artificial intelligence (AI)-based tools—such as natural-language processing (Hirschberg and Manning 2015)—that detect potentially harmful content (Gongane et al. 2022; Cambridge Consultants 2019). To a growing extent, algorithms are able to detect hate speech in languages associated with countries at relatively high risk of new mass atrocities, such as Indonesia (Ibrohim and Budi 2019). The parameters and performance of the detection algorithms that social media companies employ to track and mitigate hate speech on their platforms, however, are unclear.

Automated efforts typically still require some human intervention and oversight, particularly when determining whether the content violates standards after it is detected by AI tools (Gillespie 2018; Roberts 2017b; Vincent 2019). While AI tools may more quickly detect large amounts of harmful content, they can be prone to both “false negatives,” which incorrectly overlook harmful content, and “false positives,” which flag otherwise-harmless information as harmful (Cambridge Consultants 2019). These tools “[train] computers to perform specific analytical tasks based on repeated exposure to data,” for which “[s]uccess … hinges on a compilation of rich and large data sets" (Schiffrin et al. 2022). This makes it more difficult, for example, to sort out dangerous content in the absence of comprehensive and high-quality data sets on issues like hate speech that are largely context- or language-specific (Brown 2016; Hao 2021). Further, automated technologies currently maintain a limited understanding of specific social, cultural, and political contexts required for complex interpretation (Gillespie 2018). Although automated systems flag harmful content with growing accuracy, systematic reviews find that agreement between human coders and automated systems in identifying hate speech in a fixed corpus of text varies by language (Poletto et al. 2021) and the categories that the system uses to evaluate hateful content (Rini et al. 2020).

Newer ideas involve creating specific staff positions at social media companies to advance more effective moderation. For example, one analyst suggested “companies should convene groups of experts in various domains to constantly monitor the major topics in which fake news or hate speech may cause serious harm” and then create focused moderation strategies specifically for these fields (Yaraghi 2019). Social media companies can also expand their work with local actors and, with adequate resources, train moderators of the specific dynamics in focus countries, especially for high-risk contexts (Hook and Verdeja 2022).

Modifying content-promotion algorithms

Social media companies have recently instituted several system-wide changes to their platforms to mitigate risks that content will inflame violence. Adjusting how users receive potentially dangerous content may mitigate violence risks without removing the content altogether. This approach shares with content moderation the core challenge of identifying content that would increase atrocity risks, with accuracy and speed; it also relies on identifying specific ways to minimize risks associated with that content without removing it.

Downranking: Companies have altered their content-promoting algorithms to “downrank” and reduce the visibility of potentially harmful content (Byman 2022). To cite recent examples: Twitter’s community guidelines indicate that the company reduces the visibility of potentially harmful content “[d]epending on the potential for offline harm,” including by removing recommendations of the offending account’s Twitter posts (Twitter 2022a; Twitter 2022b). In 2019, YouTube announced changes to its recommendation algorithms to limit the spread of potentially harmful content (YouTube 2019). Instagram reports that “it may make it harder to find, rather than removing” content that could be “upsetting, offensive, or sensitive” but does not violate the platform’s Community Guidelines (Instagram 2022). In 2022, Facebook announced it would reduce “emphasis on shares and comments for political content” in user feeds globally (Stepanov 2021).

These downranking strategies can also be situation-specific. During "times of crisis" including conflict, Twitter limits or entirely disables engagement options—such as retweeting or liking a post—for content identified as harmful (Twitter 2022b). In response to its assessment of the impact of Facebook on ethnic violence in Sri Lanka, Meta set limits on how many times users could forward any message on WhatsApp to limit the spread of viral content (Facebook 2020). Additionally, others have proposed implementing "a tool to limit the exponential amplification of content, until human reviewers can determine whether the content violates platform policies or poses a risk to public safety" (Kornbluh and Goodman 2021).

Some observers have also recommended limiting visual displays of the virality of harmful information. This strategy would include "obscuring like and share counts so that individual pieces of content can be evaluated on their own merit" to prevent the spread of potentially harmful content as opposed to slowing its spread (Haidt and Rose-Stockwell 2019).

Labeling: Twitter and Instagram have recently introduced warning labels over forms of misinformation intended to slow its spread (Twitter 2022b; Instagram 2022). Observers have proposed broader and more transparent content labeling, such as including information on "who funds their outlet, where to find the outlet's standards, and whether the article claims to be news or opinion" (Kornbluh and Goodman 2021). Similarly, proposals have included allowing viewers to tag videos to "produce aggregate data by which users could filter their viewing experience" (Gillespie 2018). Additionally, suggestions have included labeling "[h]eavily flagged content, especially if by multiple, unconnected users" (Gillespie 2018). Twitter has introduced limited testing in the United States of a "Community Notes" feature where users "can write a note with additional information, to provide public context to the community on a Tweet they feel is misleading" (Twitter n.d.).

Counterspeech: Companies like YouTube have developed systems "to provide alternative information alongside the content with fake information so that the users are exposed to the truth and correct information" (Yaraghi 2019). Civil society organizations also use "peace messaging" campaigns in high-risk contexts to provide alternative narratives that discourage radicalization, promote reconciliation, and encourage resilience against atrocity risks (Brown 2016), as the well-documented case of messaging around Kenya's 2013 elections illustrates (Benesch 2014). Experimental evidence finds that these counterspeech efforts are especially effective in reducing hate speech when counter-speakers share a common identity with their audience (Siegel and Badaan 2020).

Restricting user access

Companies have also suggested limiting user access to restrict the ability of individuals to promote harmful information. This approach focuses on identifying potentially risky *users* rather than potentially risky *content*. The core challenge is analogous to those described above, but in this case about identifying users whose activity is likely to contribute to mass atrocity risks.

Deplatforming: Some social media platforms have regulated user access to limit the risk of harmful information spreading, such as through temporarily or permanently removing accounts. Twitter policy states it will "remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks" (Twitter 2022c). Facebook specifies that in times of "civil unrest," platform moderators "may also restrict accounts by public figures for longer periods of time when they incite or praise ongoing violence" (Meta 2022), as it did for state-run accounts before recent elections in Uganda and Iran (Satariano 2021).

Identity verification: Some scholars have proposed identity verification processes as a means for reducing the spread of misinformation and otherwise potentially harmful content (Galloway 2022). However, limited research explores the link between Identity verification and the spread of misinformation (Wang, Pang, and Pavlou 2021). Additionally, while some companies like Facebook have "real name" policies, these can be easily circumvented, and in addition, these policies pose concerns for individuals, such as activists, whose safety depends on remaining anonymous (Byman 2022).

User choice: Observers have proposed giving users more power to decide their individually curated content by choosing transparent algorithm options to help reduce the likelihood of harmful content amplification (Kornbluh and Goodman 2021).

Government regulation

Because social media platforms rely on user engagement to boost advertising revenues, companies have few direct financial incentives to address the proliferation of popular hateful content on their sites and applications (Gillespie 2018; Kornbluh and Goodman 2019). However, increasing concerns surrounding the dangerous use of social media globally have placed more pressure on firms to shift their practices. As journalist Tiffany Hsu explains, "increasingly, companies in and outside the tech industry see that abuse as a potentially expensive liability, especially as more work is conducted online and regulators and clients push for stronger guardrails" (Hsu 2022).

Many observers have pushed for government regulation that holds companies liable for violence-promoting content on their platforms (Schiffrin et al. 2022). In the United States, calls have centered on reforming Section 230 of the 1996 Communications Decency Act, which holds that "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" (47 U.S. Code § 230). While some have called for repealing these specific provisions, other experts have proposed revisions to the law focused on increasing accountability for tech companies and limiting harmful uses of social media. For example, proposed amendments have included revising Section 230 so that these companies must "[take] reasonable steps to prevent or address unlawful uses of its services" to be shielded from liability for content posted (Citron and Wittes 2017; Hamilton 2022).

Others have called to make these companies "liable for knowingly promoting unlawful content or activity that threaten or intentionally incite physical violence," among other dangerous uses (Kornbluh and Goodman 2019). Hamilton (2022) also observes that social media companies may enable atrocities regardless of the level of knowledge associated with the relationship between platforms and the onset of violence. As one report notes, "[a]lthough regulations are difficult to enforce, the mere awareness of them may incentivize tech giants to increase removals or down-rank mis/disinformation on their sites" (Schiffrin et al. 2022).

Concerns related to proposed solutions

Many observers have expressed concerns that strategies that involve removing or restricting certain content or users are in tension with free speech rights. As long as it is impossible to identify speech that will fuel violence with perfect accuracy, these strategies require judgments that some will see as overly restrictive and others will see as insufficiently restrictive. Some also suggest that pressure on social media companies to manage content on their sites will harm freedom-of-speech protections by giving companies more discretion over standards for permissible content (Bambauer 2020; Schiffrin et al. 2022) and granting already-repressive governments more control over the spread of dissenting information (Neo 2022). Additionally, reform conversations have focused mainly on potential US or Western European government responses to US-based tech companies, with little focus on other platforms or regulations in other contexts.

In addition, atrocity prevention advocates and other activists have raised concerns that "removing content … may lead to a lack of information access for prosecutors, investigators, and prevention policymakers, especially in international contexts not subject to legal warrants" (Hook and Verdeja 2022). This also raises worries that content meant to spread awareness may be removed and thus harm advocacy efforts (Solon 2020). To partly address this issue, proposals include creating a restricted-access archive to preserve posts for justice or advocacy efforts to draw on (Mooney et al. 2021) and formulating more careful, context-specific content moderation processes involving human moderators to differentiate educational from harmful content.

ENDNOTES

[1] Despite recent attention associated with cases such as the genocide against the Rohingya community in Burma, little direct research exists about the effects of social media on mass atrocities, specifically (Hook and Verdeja 2022). We synthesize research on social media and other forms of political violence and discuss how this research might apply to mass atrocities. We discuss how the mechanisms that link social media to violence may lead to violence of greater scale, lethality, and/or systematicity—that is, how social media may heighten risks of mass atrocities.

[2] In reviewing the relevant literature, we focus on evidence about the relationship between social media, mass atrocities, and atrocity prevention in contexts at relatively high risk of mass atrocities. Although the mechanisms that link social media to violence in lower-risk contexts may also translate to higher-risk environments, the absence of other, non-social media factors in the former contexts—for example, ongoing civil conflict or a recent history of mass atrocities—lowers the overall risk of mass atrocities. We focus on higher-risk contexts because we expect that social media will have different effects on risks of escalation where these broader social and political conditions are present than where they are not. It is important to note that social media may also contribute to risks of other violence, such as targeted hate crimes, in countries at otherwise-low risk of mass atrocities (Siegel 2020).

[3] Logically, social media platforms should not have an effect on the third major risk factor for mass atrocities for which there is consistent research evidence, a prior history of discrimination or violence against a targeted group. Although evidence about the risk factor centers on the cumulative effects of prior group-targeted violence, it is important to note that users who seek to mobilize against the targeted group may use social media to distort collective memories of previous episodes in a way that normalizes new violence. We thank Anita Gohdes for this observation.

[4] In December 2022, a group of Ethiopian and Kenyan advocates filed a lawsuit against Meta in Kenya accusing the firm of falling short of its responsibility to suppress dangerous content on its platform (Paul and Mersie 2022). Rohingya advocates filed a similar suit against Meta in US court in December 2021 (Culliford 2021).

[5] We compiled these data from DataReportal, a digital market research site. DataReportal defines *social media users* "as users that have logged into at least one social media platform, or made active use of at least one social media platform while logged in, during the past 30 days." DataReportal's measure of *social media use* includes data on, but not limited to Facebook, Instagram, LinkedIn, Snapchat, and Twitter. Additionally, DataReportal notes that country-specific data on *social media users* from its *Digital 2021* and *Digital 2022* country reports are not comparable with previous reports' data on social media users. For more information, visit https://datareportal.com/.

[6] The Early Warning Project's SRA uses publicly available data and statistical modeling to produce a list of countries ranked by their estimated risk of experiencing a new episode, or onset, of mass killing. For more information about the Early Warning Project and the annual country assessment, see https://earlywarningproject.ushmm.org/.

[7] According to the Early Warning Project, Chad has continued to increase in risk, landing at fourth with a risk estimate above nine percent for 2022–23 after being ranked tenth last year and 23rd the year before. Chad has consistently ranked in the high-risk (top-30) category, with fourth marking its highest ranking to date. For more information, visit: https://earlywarningproject.ushmm.org/countries/chad.

[8] Scholars have warned of the disproportionate attention placed on major social media platforms, especially Facebook, in conversations about content moderation. Gillespie and Aufderheide (2020) note: "The largest, US-based platforms do not provide a reliable guide for the entire social media ecology; innovative moderation strategies may emerge from smaller platforms, platforms outside of the US, and platforms that imagine themselves and their communities very differently than Facebook does."

[9] It bears noting that content-moderation contracting firms often expose content moderators to poor labor conditions, including low wages and regular (Stackpole 2022).

WORKS CITED

Article One Advisors. 2018. “Assessing the Human Rights Impact of the Facebook Platform in Sri Lanka.” San Francisco, CA: Article One.

Bambauer, Derek E. 2020. “How Section 230 Reform Endangers Internet Free Speech.” *Brookings* (blog). July 1, 2020. https://www.brookings.edu/techstream/how-section-230-reform-endangers-internet-free-speech/.

Barberá, Pablo. 2020. “Social Media, Echo Chambers, and Political Polarization.” In *Social Media and Democracy: The State of the Field, Prospects for Reform*, edited by Joshua A. Tucker and Nathaniel Persily, 34–55. SSRC Anxieties of Democracy. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/social-media-and-democracy/social-media-echo-chambers-and-political-polarization/333A5B4DE1B67EFF7876261118CCFE19.

Benesch, Susan. 2014. “Countering Dangerous Speech to Prevent Mass Violence during Kenya’s 2013 Elections.” Washington, DC: US Holocaust Memorial Museum. https://www.ushmm.org/m/pdfs/20140212-benesch-kenya.pdf.

Bhavnani, Ravi, Michael G. Findley, and James H. Kuklinski. 2009. “Rumor Dynamics in Ethnic Violence.” *The Journal of Politics* 71 (3): 876–92. https://doi.org/10.1017/S002238160909077X.

Brown, Rachel. 2016. *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech*. Washington, DC: US Holocaust Memorial Museum. https://www.ushmm.org/genocide-prevention/reports-and-resources/defusing-hate-a-guide-to-counteract-dangerous-speech.

Buni, Catherine and Chemaly, Soraya. 2016. “The Secret Rules of the Internet.” *The Verge*, April 13, 2016. https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

Byman, Daniel. 2022. “Content Moderation Tools to Stop Extremism.” The Digital Social Contract: A Lawfare Paper Series. Washington, DC: Lawfare.

Cambridge Consultants. 2019. “Use of AI in Online Content Moderation.” London, UK: Ofcom.

Carr, Caleb T., and Rebecca A. Hayes. 2015. “Social Media: Defining, Developing, and Divining.” *Atlantic Journal of Communication* 23 (1): 46–65. https://doi.org/10.1080/15456870.2015.972282.

Citron, Danielle, and Benjamin Wittes. 2017. “The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity.” *Fordham Law Review* 86 (2): 401.

Common, MacKenzie F. 2020. “Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media.” *International Review of Law, Computers & Technology* 34 (2): 126–52. https://doi.org/10.1080/13600869.2020.1733762.

Culliford, Elizabeth, and Elizabeth Culliford. 2021. “Rohingya Refugees Sue Facebook for $150 Billion over Myanmar Violence.” *Reuters*, December 8, 2021, sec. Asia Pacific.

https://www.reuters.com/world/asia-pacific/rohingya-refugees-sue-facebook-150-billion-over-myanmar-violence-2021-12-07/.

Dangerous Speech Project. 2021. *Dangerous Speech: A Practical Guide*. Washington, DC: Dangerous Speech Project. https://dangerousspeech.org/guide/.

DataReportal. 2022. “Global Social Media Statistics.” DataReportal. October 2022. https://datareportal.com/social-media-users.

Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. “The Spreading of Misinformation Online.” *Proceedings of the National Academy of Sciences* 113 (3): 554–59. https://doi.org/10.1073/pnas.1517441113.

Earl, Jennifer, Thomas V. Maher, and Jennifer Pan. 2022. “The Digital Repression of Social Movements, Protest, and Activism: A Synthetic Review.” *Science Advances* 8 (10): eabl8198. https://doi.org/10.1126/sciadv.abl8198.

Facebook. 2020. “Facebook Response: Sri Lanka Human Rights Impact Assessment.” Menlo Park, CA: Meta. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf.

———. 2021. “Reducing Political Content in News Feed.” *Meta* (blog). February 10, 2021. https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/.

Galloway, Scott. 2022. “ID.” No Mercy / No Malice. September 30, 2022. https://www.profgalloway.com/id/.

Gilbert, David. 2020. “Hate Speech on Facebook Is Pushing Ethiopia Dangerously Close to a Genocide.” *Vice* (blog). September 14, 2020. https://www.vice.com/en/article/xg897a/hate-speech-on-facebook-is-pushing-ethiopia-dangerously-close-to-a-genocide.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. 1 online resource (288 pages) : illustrations vols. New Haven: Yale University Press. https://www.degruyter.com/isbn/9780300235029.

Gohdes, Anita R. 2015. “Pulling the Plug: Network Disruptions and Violence in Civil Conflict.” *Journal of Peace Research* 52 (3): 352–67. https://doi.org/10.1177/0022343314551398.

Gohdes, Anita R. 2020. “Repression Technology: Internet Accessibility and State Violence.” *American Journal of Political Science* 64 (3): 488–503. https://doi.org/10.1111/ajps.12509.

Gongane, Vaishali U., Mousami V. Munot, and Alwin D. Anuse. 2022. “Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions.” *Social Network Analysis and Mining* 12 (1): 129. https://doi.org/10.1007/s13278-022-00951-3.

Hamilton, Rebecca J. 2022. “Platform-Enabled Crimes: Pluralizing Accountability When Social Media Companies Enable Perpetrators to Commit Atrocities.” *Boston College Law Review* 63 (4): 1349.

Hao, Karen. 2021. "AI Still Sucks at Moderating Hate Speech." *MIT Technology Review*, 2021. https://www.technologyreview.com/2021/06/04/1025742/ai-hate-speech-moderation/.

Hayes, Adam. n.d. "YouTube Stats: Everything You Need to Know In 2023!" Wyzowl. Accessed January 7, 2023. https://www.wyzowl.com/youtube-stats/.

Hirschberg, Julia, and Christopher D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349 (6245): 261–66. https://doi.org/10.1126/science.aaa8685.

Hook, Kristina, and Ernesto Verdeja. 2022. "Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities." Washington DC: Stimson Center. https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/.

Hsu, Tiffany. 2022. "Sympathy, and Job Offers, for Twitter's Misinformation Experts." *The New York Times*, November 28, 2022, sec. Technology. https://www.nytimes.com/2022/11/28/technology/twitter-misinformation-experts-hiring.html.

Ibrohim, Muhammad Okky, and Indra Budi. 2019. "Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter." In *Proceedings of the Third Workshop on Abusive Language Online*, 46–57. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3506.

Instagram. 2022. "How Do I Appeal Instagram's Content Decision to the Oversight Board?" 2022. https://www.facebook.com/help/instagram/675885993348720.

———. n.d. "How to Limit Sensitive Content That You See on Instagram." Accessed January 7, 2023a. https://help.instagram.com/251027992727268.

———. n.d. "Why Is a Post on Instagram Marked as False Information?" Accessed January 7, 2023b. https://help.instagram.com/388534952086572.

International Crisis Group. 2022. "Chad's Transition: Easing Tensions Online." Crisis Group Africa Briefing. N'Djamena, Chad: International Crisis Group. https://www.crisisgroup.org/africa/central-africa/chad/b183-chads-transition-easing-tensions-online.

Jones, Jason J., Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow, and James H. Fowler. 2013. "Inferring Tie Strength from Online Directed Behavior." *PLOS ONE* 8 (1): e52168. https://doi.org/10.1371/journal.pone.0052168.

Kietzmann, Jan H., Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons*, SPECIAL ISSUE: SOCIAL MEDIA, 54 (3): 241–51. https://doi.org/10.1016/j.bushor.2011.01.005.

Kim, Nam Kyu. 2018. "Revolutionary Leaders and Mass Killing." *Journal of Conflict Resolution* 62 (2): 289–317. https://doi.org/10.1177/0022002716653658.

King, Gary, Jennifer Pan, and Margaret Roberts. 2013. “How Censorship in China Allows Government Criticism but Silences Collective Expression.” *American Political Science Review* 107 (2 (May)): 1–18.

Kornbluh, Karen and Ellen P. Goodman. 2019. “Bringing Truth to the Internet.” *Democracy Journal* Summer 2019 (53). https://democracyjournal.org/magazine/53/bringing-truth-to-the-internet/.

Kornbluh, Karen and Ellen P. Goodman. 2021. “Immediate Bipartisan Fixes for Social Media.” *GMFUS* (blog). September 28, 2021. https://www.gmfus.org/news/immediate-bipartisan-fixes-social-media.

Liebowitz, Jeremy. 2019. “Hate Speech, Disinformation and Political Violence in Myanmar.” In . Bangkok, Thailand: Asia Centre.

Martin, Michelle. 2022. “39 Facebook Stats That Matter to Marketers in 2023.” *Social Media Marketing & Management Dashboard* (blog). March 2, 2022. https://blog.hootsuite.com/facebook-statistics/.

Maynard, Jonathan Leader. 2022. *Ideology and Mass Killing: The Radicalized Security Politics of Genocides and Deadly Atrocities*. Oxford, New York: Oxford University Press.

Meta. 2022. “Restricting Accounts.” October 4, 2022. https://transparency.fb.com/enforcement/taking-action/restricting-accounts/.

Montalvo, José G., and Marta Reynal-Querol. 2005. “Ethnic Polarization, Potential Conflict, and Civil Wars.” *American Economic Review* 95 (3): 796–816. https://doi.org/10.1257/0002828054201468.

Mooney, Olivia, Kate Pundyk, Nathaniel Raymond, and David Simon. 2021. “Social Media Evidence of Alleged Gross Human Rights Abuses: Improving Preservation and Access Through Policy Reform.” Mass Atrocities in the Digital Era Initiative. New Haven, CT: Yale MacMillan Center.

Neo, Ric. 2022. “A Cudgel of Repression: Analysing State Instrumentalisation of the ‘Fake News’ Label in Southeast Asia.” *Journalism* 23 (9): 1919–38. https://doi.org/10.1177/1464884920984060.

Nyseth Brehm, Hollie. 2016. “State Context and Exclusionary Ideologies.” *American Behavioral Scientist* 60 (2): 131–49. https://doi.org/10.1177/0002764215607579.

Ortiz-Ospina, Esteban. 2019. “The Rise of Social Media.” *Our World in Data* (blog). September 18, 2019. https://ourworldindata.org/rise-of-social-media.

Paul, Katie, and Ayenat Mersie. 2022. “Meta Accused in Lawsuit of Allowing Posts That Inflamed Ethiopia Conflict.” *Reuters*, December 15, 2022, sec. Africa. https://www.reuters.com/world/africa/lawsuit-accuses-meta-enabling-hateful-posts-ethiopia-conflict-2022-12-14/.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. “Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review.” *Language Resources and Evaluation* 55 (2): 477–523. https://doi.org/10.1007/s10579-020-09502-8.

Rini, Rini, Ema Utami, and Anggit Dwi Hartanto. 2020. “Systematic Literature Review Of Hate Speech Detection With Text Mining.” In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–6. https://doi.org/10.1109/ICORIS50180.2020.9320755.

Roberts, Sarah T. 2017a. “Content Moderation.” February 5, 2017. https://escholarship.org/uc/item/7371c1hf.

———. 2017b. “Social Media’s Silent Filter.” *The Atlantic*, March 8, 2017. https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.

Rose-Stockwell, Jonathan Haidt, Tobias. 2019. “The Dark Psychology of Social Networks.” *The Atlantic*, November 12, 2019. https://www.theatlantic.com/magazine/archive/2019/12/social-media-democracy/600763/.

Satariano, Adam. 2021. “After Barring Trump, Facebook and Twitter Face Scrutiny About Inaction Abroad.” *The New York Times*, January 14, 2021, sec. Technology. https://www.nytimes.com/2021/01/14/technology/trump-facebook-twitter.html.

Schiffrin, Anya, Hiba Beg, Juan Carlos Eyzaguirre, Zachey Kliger, Tianyu Mao, Aditi Rukhaiyar, Kristen Saldarini, and Ojani Walthrust. 2022. “AI Startups and the Fight Against Mis/Disinformation Online: An Update.” Washington, DC: German Marshall Fund.

Siegel, Alexandra A. 2020. “Online Hate Speech.” In *Social Media and Democracy: The State of the Field, Prospects for Reform*, edited by Joshua A. Tucker and Nathaniel Persily, 56–88. SSRC Anxieties of Democracy. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/social-media-and-democracy/online-hate-speech/28D1CF2E6D81712A6F1409ED32808BF1.

Siegel, Alexandra A., and Vivienne Badaan. 2020. “#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online.” *American Political Science Review* 114 (3): 837–55. https://doi.org/10.1017/S0003055420000283.

Sissons, Miranda. 2020. “An Update on Facebook’s Human Rights Work in Asia and Around the World.” *Meta* (blog). May 12, 2020. https://about.fb.com/news/2020/05/human-rights-work-in-asia/.

Solon, Olivia. 2020. “‘Facebook Doesn’t Care’: Activists Say Accounts Removed despite Zuckerberg’s Free-Speech Stance.” NBC News. June 15, 2020. https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110.

Stackpole, Thomas. 2022. “Content Moderation Is Terrible by Design.” *Harvard Business Review*, November 9, 2022. https://hbr.org/2022/11/content-moderation-is-terrible-by-design.

Staub, Ervin. 1989. *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press.

Straus, Scott. 2007. “What Is the Relationship between Hate Radio and Violence? Rethinking Rwanda’s ‘Radio Machete.’” *Politics & Society* 35 (4): 609–37. https://doi.org/10.1177/0032329207308181.

———. 2015a. *Making and Unmaking Nations: War, Leadership, and Genocide in Modern Africa*. 1st ed. Cornell University Press. https://www.jstor.org/stable/10.7591/j.ctt20fw633.

———. 2015b. “Triggers of Mass Atrocities.” *Politics and Governance* 3 (3): 5–15.

———. 2016. *Fundamentals of Genocide and Mass Atrocity Prevention*. Washington, DC: US Holocaust Memorial Museum.

Taub, Amanda, and Max Fisher. 2018. “Where Countries Are Tinderboxes and Facebook Is a Match.” *The New York Times*, April 21, 2018, sec. World. https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

Twitter. 2022. “Perpetrators of Violent Attacks Policy.” April 2022. https://help.twitter.com/en/rules-and-policies/perpetrators-of-violent-attacks.

———. n.d. “Crisis Misinformation Policy.” Accessed January 7, 2023a. https://help.twitter.com/en/rules-and-policies/crisis-misinformation.

———. n.d. “How We Address Misinformation on Twitter.” Accessed January 7, 2023b. https://help.twitter.com/en/resources/addressing-misleading-info.

Valentino, Benjamin. 2016. “Triggers for Mass Killing: Report on a Research Project for the Political Instability Task Force.” McLean, VA: Political Instability Task Force.

Vincent, James. 2019. “AI Won’t Relieve the Misery of Facebook’s Human Moderators.” *The Verge*, February 27, 2019. https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms.

Wang, Shuting (Ada), Min-Seok Pang, and Paul A. Pavlou. 2021. “Cure or Poison? Identity Verification and the Posting of Fake News on Social Media.” *Journal of Management Information Systems* 38 (4): 1011–38. https://doi.org/10.1080/07421222.2021.1990615.

Whitten-Woodring, Jenifer, Mona S. Kleinberg, Ardeth Thawnghmung, and Myat The Thitsar. 2020. “Poison If You Don’t Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar.” *The International Journal of Press/Politics* 25 (3): 407–25. https://doi.org/10.1177/1940161220919666.

Yanagizawa-Drott, David. 2014. “Propaganda and Conflict: Evidence from the Rwandan Genocide *.” *The Quarterly Journal of Economics* 129 (4): 1947–94. https://doi.org/10.1093/qje/qju020.

Yaraghi, Niam. 2019. “How Should Social Media Platforms Combat Misinformation and Hate Speech?” *Brookings* (blog). April 9, 2019. https://www.brookings.edu/blog/techtank/2019/04/09/how-should-social-media-platforms-combat-misinformation-and-hate-speech/.

YouTube. 2019. “Continuing Our Work to Improve Recommendations on YouTube.” Blog.Youtube. January 25, 2019. https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

Zeitzoff, Thomas. 2017. “How Social Media Is Changing Conflict.” *Journal of Conflict Resolution* 61 (9): 1970–91. https://doi.org/10.1177/0022002717721392.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political Effects of the Internet and Social Media." *Annual Review of Economics* 12 (1): 415–38. https://doi.org/10.1146/annurev-economics-081919-050239.

ACKNOWLEDGEMENTS

We thank Anita Gohdes, Thomas Zeitzoff, and our colleagues at the Simon-Skjodt Center for their helpful feedback on earlier drafts.

UNITED STATES HOLOCAUST MEMORIAL MUSEUM
SIMON-SKJODT CENTER
FOR THE PREVENTION OF GENOCIDE

100 Raoul Wallenberg Place, SW Washington, DC 20024-2126 ushmm.org