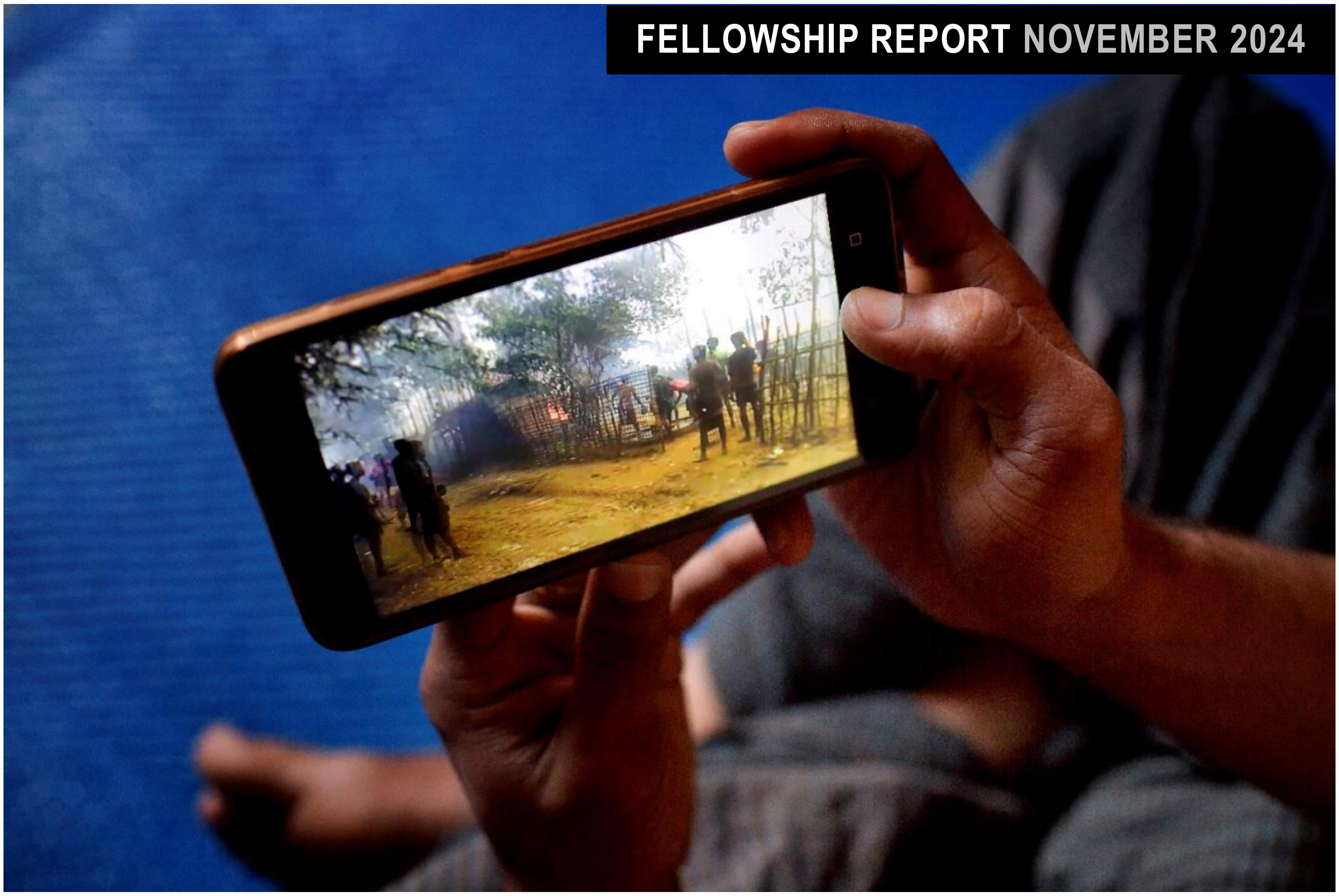
LEVERAGING SOCIAL MEDIA FOR GENOCIDE AND MASS ATROCITY PREVENTION

Understanding the Digital Toolbox



FELLOWSHIP REPORT NOVEMBER 2024

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 fulfill Elie Wiesel's challenge 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 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.**

SHANNON RAJ SINGH, Leonard and Sophie Davis Genocide Prevention Fellow (2023-24) with the Simon-Skjodt Center for the Prevention of Genocide.

The assertions, opinions, and conclusions in this report are those of the author. They do not necessarily reflect those of the United States Holocaust Memorial Museum.

This work was made possible by a grant from The Leonard and Sophie Davis Fund.

COVER: In this Dec. 23, 2017 photo, Mujibullah, 22, a Rohingya refugee watches a video, which he has shot in Myanmar before crossing over into Bangladesh, at Kutupalong refugee camp in Ukhiya, Bangladesh. For many Rohingya living in refugee camps in Bangladesh, all that remains of their old lives in Myanmar are memories captured in photos and videos on their cellphones. Since August, more than 630,000 Rohingya Muslims have fled to Bangladesh to escape attacks by Myanmar security forces. Few refugees had the chance to grab many belongings when they fled, but most took their cellphones. *AP Photo/A.M. Ahad.*

CONTENTS

FOREWORD

The Museum’s Founding Chairman and Holocaust survivor Elie Wiesel’s vision was that the Museum would do for victims of genocide today what was not done for the Jews of Europe: "Only a conscious, concerted attempt to learn from past errors can prevent recurrence to any racial, religious, ethnic, or national group."

The Museum's Simon-Skjodt Center was established to help fulfill that vision. The Center’s mandate is "to alert the national conscience, influence policy makers, and stimulate worldwide action to confront and prevent genocide." As a trusted resource for government officials, we strive to promote broad and enduring bipartisan commitment among policy makers to preventing genocide and related crimes against humanity.

As one way to carry out this charge, the Simon-Skjodt Center explores contemporary and future trends affecting mass atrocities and atrocity prevention. The growth of social media around the world is having profound and far reaching effects, including on the risk of mass atrocities. Much has been written about the potential for social media to spread antisemitism and other hate, mis/disinformation, and incite people to violence. Less is known about the other side of the coin: how social media can be used to help prevent mass atrocities. Therefore, in 2023, we recruited a fellow to research this topic, culminating in this report.

Shannon Raj Singh, an international criminal lawyer with expertise in the intersection of technology and mass atrocities, carried out this project during her time as a Leonard and Sophie Davis Genocide Prevention Fellow. She consulted current and former representatives of social media companies, academics, practitioners, and members of at-risk communities to understand the range of social media interventions that may support core atrocity prevention strategies.

The report’s findings affirm that social media tools and interventions offer an opportunity to expand the atrocity prevention toolbox to meet contemporary challenges. To seize this opportunity, social media companies first need to recognize the influence of their platforms in countries at significant risk of mass atrocities and their responsibility to use their tools to help prevent mass atrocities. The report discusses numerous specific ways that social media could be–and in some cases, has been–used to help prevent mass atrocities. Yet, the findings also underscore that these product, policy, and operational interventions require further research, including on their potential unintended consequences. One important, overriding challenge is how to reduce the reach or influence of social media content that might increase the risk of mass atrocities while protecting the right to free expression.

Alongside the Simon-Skjodt Center's “Strategic Framework for Helping Prevent Mass Atrocities” and Tools for Atrocity Prevention website, this report is meant to help people think through the range of potential options to help prevent mass atrocities. The project's main goal is to increase awareness of how tools deployed or developed by social media companies might reduce the risk of mass atrocities.

Preventing genocide is of course difficult. We know from the Holocaust what can happen when early warning signs go unheeded and responses fall short. We aim for this report to serve as a tool and a resource for social media professionals, policy makers, practitioners, and others interested in prevention. We hope it helps them think through the actions that can make the greatest impact in saving lives.

Naomi Kikoler
Director, Simon-Skjodt Center for the Prevention of Genocide
United States Holocaust Memorial Museum
November 2024

EXECUTIVE SUMMARY

Much of the existing literature discussing social media focuses on how it might fuel or incite mass atrocities, drawing from experiences in contexts such as Sri Lanka and Burma. But there is significantly less awareness of how tools deployed or developed by social media companies might *reduce* the risk of mass violence and contribute constructively to atrocity prevention efforts.

This report aims to address this gap by focusing on how social media tools can support two core atrocity prevention strategies:

(1) protecting vulnerable civilian populations at risk of mass atrocities, and
(2) degrading potential perpetrators' capacity to commit mass atrocities.

It provides a landscape assessment of the suite of social media product, policy, and operational interventions that may offer potential to support these strategies and articulates some of the associated limitations, risks, and important considerations when these tools are deployed.

This report is primarily aimed at those inside social media companies with authority to develop or deploy tools in moments of heightened atrocity risk (which may include trust and safety professionals, human rights or crisis response teams, and senior leadership), as well as atrocity prevention experts and policy makers who may be able to encourage or incentivize the use of digital tools to support atrocity prevention. Select tools may also be of interest or use for humanitarian and civil society advocacy organizations that operate in atrocity risk settings.

The objective of this report is to fill a gap by expanding the understanding of both policy makers and social media platform representatives about the available tools in the digital realm to support atrocity prevention efforts, to stimulate future research in this space, and to broaden our collective imaginations in designing modern atrocity prevention policy strategies that leverage digital tools and opportunities.

This report is based on a series of semi-structured expert consultations, held under the Chatham House Rule of non-attribution, with more than 30 current and former representatives of social media companies, academics and practitioners specialized in technology and atrocity prevention, and members of at-risk communities who lent their experiences and insights to support this project.

The report concludes that expanding the atrocity prevention toolbox to include digital tools and interventions offers an opportunity to develop more modern atrocity prevention strategies to meet the challenges of the moment.

It identifies the following categories of interventions as offering potential to support civilian protection:

- **Protecting online privacy**: tools or interventions aimed at restricting the visibility of digital content that may put civilians at risk in atrocity risk settings
- **Securing social media accounts**: interventions aimed at protecting social media users against hacking, impersonation, and account takeover efforts
- **Surfacing crisis resources and credible information**: interventions aimed at connecting social media users to crisis resources and/or amplifying credible information
- **Disseminating early-warning information**: interventions that make use of social media to communicate warnings about atrocity risks
- **Enhancing communication and coordination capabilities**: interventions that enhance civilians' ability to communicate and coordinate in atrocity risk settings

This report also identifies the following categories of interventions as offering potential to degrade the capacity of atrocity perpetrators:

- **Preventing perpetrators from gaining a foothold of platforms at scale**: interventions aimed at preventing perpetrators from setting up a large presence on social media platforms
- **Disrupting perpetrators from organizing and coordinating**: interventions aimed at disrupting perpetrators from using social media to coordinate and organize the commission of violence
- **Limiting the presence or visibility of dangerous content in atrocity risk settings**: interventions aimed at reducing the presence or visibility of potentially inflammatory digital content during periods of heightened atrocity risk
- **Contextualizing perpetrator content**: interventions aimed at providing additional information or context around inflammatory digital content
- **Preventing perpetrators from mobilizing bystanders**: interventions aimed at reducing the incentives for bystanders or third-party enablers to inadvertently contribute to narratives and ideologies being advanced by perpetrators
- **Implementing last resort or "break glass" measures**: interventions that temporarily and intentionally degrade or disable social media features in moments of heightened atrocity risks

For each of the preceding categories, this report sets out specific considerations and preliminary recommendations on how they might be developed and implemented. It also sets out the following as general recommendations to platforms seeking to constructively contribute to atrocity prevention efforts:

- Platforms should invest in building internal atrocity prevention capacity and expertise. They should ensure they have a dedicated crisis response function that can define and categorize potential atrocity risk situations according to a principled risk assessment process and should develop clear protocols on when various interventions and policies will be deployed.
- Platforms should invest in research and development on social media tools that hold potential to help prevent mass atrocities. The inventory of tools in this report offers a starting point for both deepening understanding of when and how different tools can address mass atrocity risks and expanding the range of available tools.
- Platforms should invest heavily in local partnerships that can support awareness of atrocity risk dynamics. These relationships should be established well in advance of moments of crisis, and platforms should explore providing training on relevant product and policy interventions so they can be rolled out more effectively in at-risk communities.
- Platforms should build their awareness on how their products are being used in atrocity risk settings to create a baseline for further assessment of risks and opportunities.
- Platforms should localize all resources to ensure accessibility and ease of use for affected communities. Any tools or interventions developed for use by individuals in at-risk communities must be made available in the relevant local languages of affected populations.
- Platforms should hold tabletop or scenario-based simulations to prepare for atrocity risk settings.
- Platforms should preserve digital evidence of mass atrocities and, where appropriate, share information to assist in the investigation and prosecution of atrocity crimes. They should also clarify their policies on data preservation in atrocity risk and conflict settings, and consult with civil society organizations (and, where feasible, affected communities) to identify content relevant to international justice and accountability efforts.

Finally, this report sets out recommendations to policy makers, urging them to assess both risks and opportunities to leverage the digital environment to address the risks of mass violence and to explore opportunities to incorporate social media tools and interventions into atrocity prevention strategies.

I. INTRODUCTION

In the wake of staggering international inaction during the Rwandan genocide, Samantha Power wrote scathingly of the limited imaginations of policy makers. Political leaders, she wrote, framed the choice before them as “one between doing nothing and sending in the Marines.”[1] This “all-or-nothing approach” to atrocity prevention, she argued, failed to capture the array of tools available to policy makers, who have a responsibility to “look at every tool in the toolbox” in the face of mass atrocities.[2]

Today, there is broad acceptance of the concept of an atrocity prevention “toolbox,” offering a diverse range of options that can support atrocity prevention strategies.[3] Indeed, the field of atrocity prevention has evolved significantly from the days when there was a perception that policy makers had only binary options of military intervention or inaction in the face of mass violence. The Simon-Skjodt Center for the Prevention of Genocide has observed that a “range of tools can be employed in both prevention and response including preventive diplomacy, peace messaging, condemnation, sanctions such as arms embargoes, travel bans and targeted economic sanctions, preventive deployment of peacekeepers or troops, accountability mechanisms, and, in rare instances, military intervention.”[4] And as the Center has previously articulated, the concept of a toolbox remains a “powerful way to counter the misconception that policy makers’ choices when facing a mass atrocity crisis amount to acquiescence or forceful intervention.”

Yet while military, diplomatic, and economic actions are widely accepted to be part of the atrocity prevention toolbox, relatively little is known about tools in the domain of social media. Much of the existing literature discussing social media focuses on how it might fuel mass atrocities, drawing from experiences in contexts such as Sri Lanka and Burma. But there is significantly less awareness of how social media tools might *reduce* the risk of mass violence and constructively contribute to atrocity prevention efforts.[5]

This report focuses on how social media tools might support two core atrocity prevention strategies: (1) protecting vulnerable civilian populations at risk of mass atrocities, and (2) degrading potential perpetrators’ capacity to commit mass atrocities. It provides a landscape assessment of the suite of social media product, policy, and operational interventions that may offer potential to support these strategies, and articulates some of the associated limitations, risks, and trade-offs when these tools are deployed.

Although the tools referenced in this report are primarily in the hands of social media platforms, greater awareness of their existence and potential impact can support both those inside and outside social media companies. Teams focused on human rights, trust and safety, and crisis response may benefit from a greater understanding of how the tools and interventions at their disposal may map onto atrocity prevention strategies, and how they can be developed with greater intentionality and preventive impact. In turn, policy makers focused on designing atrocity prevention initiatives in a given context may be able to encourage or incentivize the use of digital tools to support atrocity prevention, to engage in partnerships related to their deployment, or to factor in the possibility of their use as part of broader prevention strategies. As a result, this report is aimed at both those inside social media companies with authority to develop or deploy tools in moments of heightened atrocity risk (which may include trust and safety professionals, human rights or crisis response teams) and atrocity prevention experts and policy makers.

These new tools and interventions warrant consideration and study. For too long, many of these tools—their existence, their impact, and their consequences—may have been the subject of internal research by social media companies but were rarely discussed outside the walls of private sector actors. This has meant that the formal participation of the atrocity prevention community (experts, practitioners, and at-risk communities) in the development and deployment of these tools is sorely needed. The community's expertise is essential, too, to the work of critically assessing the impact of these tools in atrocity risk settings where they have been tested and deployed. As a result, part of the objective of this report is to bridge two distinct areas of expertise.

This report's goal is for both social media representatives and policy makers to come away with a better understanding of what social media tools have been tried in the field by platforms to date, a preliminary sense of which tools are viewed as offering potential to support meaningful prevention efforts, and an understanding of the theory of change for that tool or intervention. Where relevant, this report has also articulated some of the core considerations, limitations, and risks that should be taken into account when designing or deploying each tool, based on its use to date. The recommendations section at the end of this report provides a summary table of tools referenced, along with theories of change, core considerations, and examples.

Given that all of the tools referenced in this report require further research—including on their potential unintended consequences—the objective is not to conclusively recommend the use of specific tools at this juncture. Rather, the objective is to conduct a preliminary assessment of potential interventions, map them against atrocity prevention strategies, and clarify a possible theory of change through which these interventions may be able to contribute to prevention efforts. More broadly, the objective of this report is to expand the understanding of both policy makers and social media platform representatives about the available tools in the digital realm to support atrocity prevention efforts, to stimulate future research in this space, and to broaden our collective imagination in designing modern atrocity prevention policy strategies that leverage digital tools and opportunities.

II. THE RELATIONSHIP BETWEEN SOCIAL MEDIA AND MASS ATROCITIES

Although this report focuses on opportunities to leverage social media tools in support of atrocity prevention, it must be contextualized against what is understood about the broader relationship between social media and mass atrocities. To date, that relationship has been marked by a series of high-profile examples in which dynamics on social media have seemingly contributed to inciting or fueling mass violence, including in contexts such as Sri Lanka, Burma, and Ethiopia.[6]

As examined in a prior report by the Simon-Skjodt Center, existing literature identifies two primary risk factors that social media may influence.[7] First, social media may contribute to the presence of violent conflict or large-scale instability by promoting polarization, coordinating protest and/or rebellion, or enabling repression, including through tools for surveillance. Second, social media may contribute to the presence of exclusionary ideologies, including by normalizing violence through spreading myths or encouraging the use of hate speech, or by inciting and encouraging participation in violence. These themes were examined further during a series of interdisciplinary seminars convened by the Simon-Skjodt Center in early 2023 that brought together scholars, practitioners, policy makers, and social media company representatives for discussions on the relationship between social media and mass atrocities.

Existing research also links social media to radicalization and persuasion, as well as to inciting dangerous behavior.[8] In addition, misinformation on social media may not only contribute to the vulnerability of at-risk groups, but also may foster a state of "epistemic insecurity," in which charges of bias and conspiracy theories can undermine facts and evidence, presenting particular risk in fragile communities.[9] Further, as articulated in a recent policy brief by the Global Centre for the Responsibility to Protect, "the explosive growth of social media and digital messaging platforms have accelerated and contributed to the detrimental effects of information silos and disinformation and are increasingly used to contradict, distort or entirely deny past and ongoing atrocities or spread hateful messages that may influence or incite offline violence."[10]

This report looks at the different but related question of whether social media might also offer opportunities to support atrocity prevention. Participants in the 2023 Sudikoff Seminar raised the need for social media companies to act preventively, urging them to make greater investments in identifying and assessing atrocity risks before the onset of violence. They also suggested exploring opportunities to promote content that might de-escalate conflicts, to bridge the gap between social media companies and at-risk communities, and to strengthen responsible product design. This report explores several of these themes.

An emerging body of work focuses on the use of social media to support the prevention of violence.[11] Some of these initiatives build on pioneering efforts to make use of other forms of digital technology, such as the Eyes on Darfur campaign, which used satellite imagery, or the 2008 *Ushahidi* platform in Kenya and the 2009 Voix des Kivus project in the Democratic Republic of the Congo, both of which leveraged information submitted via text message.[12] More recent initiatives have explored how social media can provide information about emerging risks, which can similarly inform early-warning and prevention initiatives.[13] Other initiatives seek to make use of social media for peacebuilding, or to counter misinformation and disinformation.[14]

At the same time, the literature cautions against a sense of "techno-utopianism" around the use of digital technologies to support atrocity prevention and response.[15] While social media can help sustain public attention on mass atrocities, experts have flagged a series of limitations with overreliance on social media in early-warning initiatives.[16] Experts also lament the "missing conversations about trade-offs before tech deployment" and the risk of unintended adverse consequences related to the use of technology interventions in atrocity risk contexts.[17] Possible risks and unintended consequences of the tools discussed in this report have been articulated, but, as will be discussed, these interventions require more research before they are deployed in a given atrocity risk setting. Although in some cases, social media companies have led their own internal research on applicable interventions and tools, public-facing research on some of the interventions referenced in this report is scarce. This report should be considered an invitation for further study rather than a series of recommendations for specific tools.

III. DEFINING THE DIGITAL TOOLBOX

The existing atrocity prevention toolbox comprises several categories of tools. These include (1) diplomatic tools, such as mediation, naming and shaming, or public diplomacy; (2) informational tools, such as fact-finding or support for civilian self-protection; (3) economic tools, such as development assistance, investment incentives, or economic sanctions; (4) legal tools, such as protections for refugees, official amnesties, or prosecutions; and (5) military tools, such as arms embargoes, peace operations, or military intervention.[18] This report explores a sixth category: digital tools, or interventions available in the online information environment that can contribute to atrocity prevention strategies, either independently or in combination with other interventions.[19]

The focus in this report is on social media rather than on other forms of digital technology. Although the parameters of this report did not prescribe a fixed definition of social media for purposes of the consultations, the term "social media" is used in this report to refer to digital platforms on which users generate and interact with information in textual, audio, visual, or hybrid formats. As set out in the Background Paper for the Simon-Skjodt Center's 2023 Sudikoff Seminar, two fundamental characteristics are used to distinguish social media from other communication channels: (1) the concept of user-generated content to distinguish social media from legacy media and (2) the opportunity afforded by social media platforms to enable user communication and interaction.[20]

As noted previously, this report provides a landscape assessment of social media interventions or "tools" to support atrocity prevention. Tools, however, are *types* of actions, whereas strategies are *ways* in which a set of actions help achieve a stated goal.[21] Like any tools, social media interventions should be deployed pursuant to atrocity prevention strategies, so that response efforts are not "scattershot collections of discrete actions," but rather include a holistic assessment of *how* a set of actions will yield impact, as well as which tools should be used together.[22]

As a result, this report focuses on social media tools or interventions that can support two core atrocity prevention strategies, building on the Simon-Skjodt Center's previous work articulating general strategies for atrocity prevention.[23] The first is protecting vulnerable civilian populations, and the second is degrading perpetrator capacity to commit mass atrocities. These strategies have been selected for a few reasons. First, social media may offer unique opportunities to communicate with and protect vulnerable communities, many of which are today often highly reliant on the digital information space in crisis and conflict settings. Second, experience has shown that potential perpetrators of mass violence have appeared to use social media as a critical resource for organizing, coordinating, and inciting mass atrocities—indicating social media's relevance for those exploring opportunities to degrade perpetrator capabilities. Although overlap exists between interventions identified to support each strategy, each strategy presents distinct analytical considerations and objectives. In both cases, this report's primary focus was on downstream interventions that could support prevention efforts in moments of acute risk. While future research may present the opportunity to assess social media tools that can support other atrocity prevention strategies, such as dissuading potential perpetrators, these are outside of the scope of this report.

This report is based on a series of expert consultations, held under the Chatham House Rule of non-attribution, with current and former representatives of social media companies, academics and practitioners specialized in technology and atrocity prevention, and select members of at-risk communities who lent their expertise and insights to this project. Following a review of existing literature on the intersection of social media and atrocity prevention, the study team conducted a series of semi-structured interviews focused on two core issues: (1) How might social media contribute to *civilian protection efforts?* and (2) How might social media support efforts to *reduce the capacity of potential perpetrators* to commit atrocities? The team sought to understand the specific tools or interventions that might be deployed to support either strategy, as well as relevant risks, tradeoffs, and considerations. Because of the volume of interventions identified in the consultations, they have been aggregated and categorized according to their intended objective.

This report focuses not only on tools or interventions available to social media companies, but also on social media tools available to those outside companies, including but not limited to community leaders, civil society organizations, and human rights defenders, who may be able to leverage the digital environment for prevention efforts. Thus, while most of the tools identified are within the control of social media companies, others are available to a broader range of stakeholders.

IV. ALIGNING DIGITAL TOOLS WITH PREVENTION STRATEGIES

Research on the efficacy of these interventions—both individually and in combination with other tools—is vital. This can help identify not only which tools offer greatest potential, but also what unintended consequences arise when they are deployed. At the same time, all atrocity prevention tools carry risks and require thoughtful implementation and assessment of impact.

In addition, digital risks do not operate on a simple spectrum of magnitude. They differ not only in scale but also in form. The Simon-Skjodt Center's research "underscores that the effectiveness of atrocity prevention tools depends largely on factors related to the context in which the tool is used and the manner in which the tool is designed and implemented."[24] During these consultations, interviewees described how a range of socio-political contexts present varied risk profiles related to the digital environment. For example, highly authoritarian regimes may use the information space to unilaterally impose a top-down version of the truth, while in other contexts, a regime's objective may be not to impose truth, but to breed cynicism and division. Important distinctions may also exist between the risk profiles of authoritarian states with weak capacity and sophistication in digital spaces and those with sophisticated surveillance and cyber capabilities, presenting different considerations when deploying the same tools and interventions. Interviewees suggested that classifying contexts for their digital risk profiles could support efforts to inform the range and combination of appropriate digital tools to support atrocity prevention strategies.

Tools and interventions in the digital space are not deployed in isolation. They operate in ways that can interact with, undermine, or complement traditional economic, military, legal, and diplomatic atrocity prevention tools—as well as one another. Depending on the context, it may be necessary to consider a combination of digital tools that might be paired or sequenced with other, offline interventions to support a selected atrocity prevention strategy.

It is also necessary to keep in mind the limitations and inherent biases that may come into play when communities consider the use of social media tools and interventions to prevent atrocities. Most of the social media platforms discussed in this report are based in the United States, and tools developed to date may have had little input from those in at-risk communities around the world. In addition, vulnerable and marginalized groups face disproportionate risks when navigating the digital environment, including heightened risks of severe abuse and harassment. Moreover, levels of digital literacy can differ across gender, race, linguistic background, and other facets of individual identity. Those differences may meaningfully affect the accessibility and risks associated with the interventions and tools discussed in this report.

Further, social media interventions must be carefully timed and sequenced to have impact in an atrocity risk setting. Deploying digital interventions in the middle of a crisis is less likely to be effective and may set in motion unintended consequences; instead, social media interventions need to be developed and tested well in advance, so that members of vulnerable communities have an understanding of the range of tools available to them. The interventions should also be carefully calibrated to the circumstances and may be usefully paired or deployed in sequence with other tools in the atrocity prevention toolbox. As discussed later in this report, all interventions explored in this report should be accompanied by a clear road map for implementation and outreach, incorporating insights from those familiar with conflict zones and their dynamics.

V. THE DIGITAL TOOLBOX AND CIVILIAN PROTECTION

This section describes the landscape of tools and interventions discussed by interviewees that aim to support the protection of vulnerable civilian populations, as well as some of the benefits, risks, and trade-offs these tools may present. The tools may offer civilian protection by denying potential perpetrators the opportunity to attack a civilian population, by increasing civilians' capacity to defend themselves, and/or by mitigating harm to civilian populations, such as by helping them avoid or withstand attacks.[25]

A. Interventions Focused on Civilian Privacy and Visibility

Interventions Focused on Civilian Privacy and Visibility	
Description	Tools or interventions aimed at restricting the visibility of digital content that may put civilians at risk in atrocity risk settings
Theory of Change	If digital content could be used to target civilians, restricting the visibility of that content can contribute to civilian protection.
Examples	● Using features such as Facebook's "locked profile," which limits the ability to view various elements of a person's social media account, or similar interventions to limit the ability to view a user's affiliations or friends lists ● Obscuring users' previously shared location information ● Reviewing features to which users may be added without their consent that could make them more readily visible to perpetrators ● Creating channels for users' social media accounts to be secured or locked down in case of detention or arrest ● Proactively sharing instructions on the deletion or deactivation of social media accounts

During the Holocaust, thousands of Jews survived by hiding in plain sight—by changing their names, refusing to wear the compulsory Star of David, and taking great lengths to obscure their identity.[26] Today, persecuted minorities around the world continue to hide their identities in periods of surging hate, in an attempt to evade persecution and ensure their safety.[27] The ability to conceal elements of identity that may put individuals at risk must similarly make the leap into the digital environment.

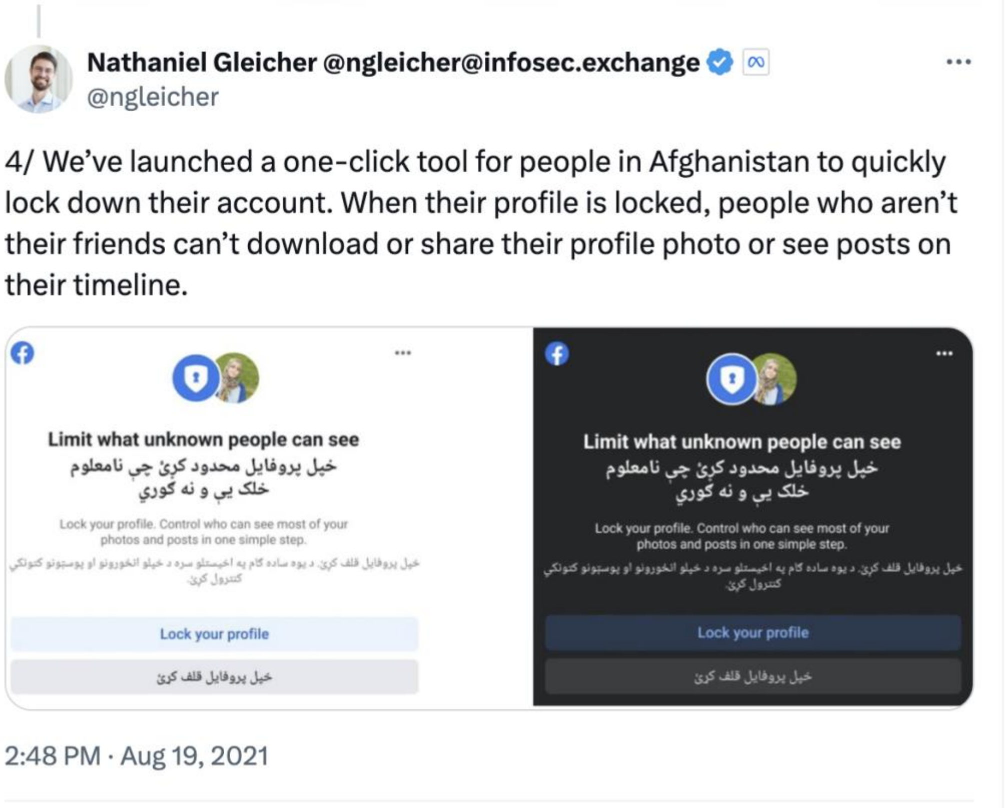
In moments of heightened atrocity risk, ensuring the privacy of information about civilians, particularly pertaining to their identities and affiliations, can be a matter of life and death. This is not only limited to information that is classically understood to be sensitive, such as one's home address or immigration status, but also can include information that may be commonly shared in safer contexts, but which may suddenly place someone at risk as atrocity risk factors emerge—such as someone's place of employment, political affiliation, or friend network. Perhaps the most precarious situations arise when information that civilians once felt comfortable openly sharing about themselves in the digital environment is unexpectedly accessed and abused by potential perpetrators, opening civilians up to targeting and potential harm.[28]

> ***"When they started to look into me, they were able to find my family members very quickly."***
>
> - Afghan activist, on the risks associated with using social media

In the digital environment, personal profiles and digital histories can be weaponized to target civilians but are often not the first thing people remember to secure in a moment of risk. Throughout consultations for this report, interviewees referenced the importance of interventions focused on ensuring the privacy of civilians in atrocity risk settings. They urged platforms to develop and deploy social media tools and interventions aimed at restricting the visibility of digital content that may put civilians at risk in atrocity risk settings. The core theory of change for this set of interventions is that where digital content is at risk of being weaponized to target civilians, restricting the visibility or accessibility of individuals' digital information can contribute to civilian protection.

Perhaps the leading example of this kind of intervention is 'locked profile,' a feature Facebook has made available in several contexts amid heightened risks of violence. According to Meta, when someone chooses to lock their profile, only their friends—and not other users—have the ability to view several aspects of their profile, such as their photos and posts, their full-size profile picture, or their Stories.[29] Interviewees explained that the product, which was informed by feedback from activists, journalists, and civil society groups, was intended to provide a streamlined option for people to lock down their profile in a single click, rather than trying to navigate a range of privacy settings in the midst of an active crisis. Its objective was to make it more difficult for someone to use open-source tools to gather information about individuals, particularly during dangerous periods, when it may be unrealistic to expect social media users to understand and adjust a number of individual privacy settings.

Facebook has offered its locked profile feature to users in Bangladesh and India, particularly to support women facing harassment on the platform, and in Burma in 2021, as anti-coup protests railed against military rule in the country.[30] The company also deployed the feature during the Taliban takeover of Afghanistan, amid fears that the Taliban would weaponize information available on social media to target individuals and in response to emerging reports that people were deleting content from their social media profiles that could link them to Western organizations or affiliations with the former Afghan government.[31] The feature was also offered in the context of the Russian invasion of Ukraine.[32]



Nathaniel Gleicher @ngleicher@infosec.exchange
@ngleicher

4/ We've launched a one-click tool for people in Afghanistan to quickly lock down their account. When their profile is locked, people who aren't their friends can't download or share their profile photo or see posts on their timeline.

Limit what unknown people can see
خپل پروفایل محدود کړئ چې نامعلوم
خلک يې و نه ګوري

Lock your profile. Control who can see most of your photos and posts in one simple step.

خپل پروفایل قلف کړئ. د یوه ساده ګام په اخیستلو سره د خپلو انځورونو او پوسټونو کتونکي
کنترول کړئ.

Lock your profile

خپل پروفایل قلف کړئ

Limit what unknown people can see
خپل پروفایل محدود کړئ چې نامعلوم
خلک يې و نه ګوري

Lock your profile. Control who can see most of your photos and posts in one simple step.

خپل پروفایل قلف کړئ. د یوه ساده ګام په اخیستلو سره د خپلو انځورونو او پوسټونو کتونکي
کنترول کړئ.

Lock your profile

خپل پروفایل قلف کړئ

2:48 PM · Aug 19, 2021

In August 2021, Meta launched 'locked profile' in Afghanistan, enabling users to quickly lock down their social media accounts. Nathaniel Gleicher, Facebook's Head of Security Policy, posted about the feature on Twitter (now X). *Twitter account of Nathaniel Gleicher*

Similar interventions may assist in obscuring an individual's relationships or connections on social media platforms. For example, in Afghanistan and Ukraine, Meta paired its locked profile feature with an intervention that proactively removes users' ability to view and search an individual's list of friends on Facebook, to protect them from those who would target individuals based on their networks.[33] It also proactively hid information about individuals' followers (and who those individuals are following) on Instagram in Ukraine and Russia.[34] Interviewees referenced similar efforts at other platforms, where features that might give away individuals' allegiances or affiliations were reviewed or temporarily disabled. X (formerly Twitter), for example, has a feature called "Lists," which allows users to compile curated lists of other users' accounts, organizing those one may want to follow according to topic or issue.[35] While this feature can help users identify lists of accounts that may offer information on a certain issue (such as journalists, subject matter experts, or those sharing resources in a crisis), one may also be added to a List without consent—opening up potential for abuse. Being involuntarily added to a List was described by one interviewee as akin to placing a bull's-eye on an individual's account, opening floodgates for that individual to be harassed or targeted.[36] Interventions that change the way involuntary features like Lists operate may support efforts to protect civilians from being readily identified by perpetrators in digital spaces.[37]

In other instances, privacy interventions can help obscure location information that an individual may have previously shared in a safer context, but which may now put them in danger. Those deploying these interventions, however, must tread carefully, as location data may also be intentionally shared in atrocity risk settings to identify the location of loved ones, humanitarian aid, or evacuation routes.[38] Where platform features are used to support civilian protection and can be abused simultaneously to target civilians, platforms must carefully consider the effects and unintended consequences of disabling or limiting the use of those features.[39]

At times, it may also be necessary for individuals—particularly members of vulnerable groups—to temporarily deactivate or delete their social media accounts entirely. In the context of the invasion of Ukraine, for example, Twitter shared information about how to deactivate individual accounts, noting that when using the platform "in conflict zones or other high-risk areas, it's important to be aware of how to control your account and digital information."[40] While this information is generally available, making sure that it is readily accessible and understandable (including in local languages) in periods of heightened risk may support civilian protection by preventing people from being targeted as a result of their social media profiles.

Unique challenges, however, arise when individuals are detained, arrested, or otherwise unable to access their accounts, yet information on their social media profiles puts them at risk. In Iran, for example, reports describe authorities accessing the social media accounts of detained individuals to identify the networks of activists with whom detainees are in touch.[41] To address these risks, interviewees referenced the importance of features or policies that enable delegated account access, allowing family members or trusted organizations to lock down the accounts of arrested or detained individuals when they cannot do so themselves, thereby making information on their social media profiles no longer visible to authorities or potential perpetrators. Consent for another user or organization to have delegated access to an individual's social media account, however, typically needs to be given in advance, a limitation for those who may not anticipate their arrest or detention. In addition, according to interviewees, requests to protect the accounts of detained users are typically handled on an individual basis by platform staff and can run into capacity constraints from teams unable to respond to requests around the clock in emergency settings. Interviewees emphasized the need for private channels allowing human rights defenders to flag that an individual has been detained and that the relevant account should be locked down until it is established that the individual is no longer at risk.[42]

While interviewees broadly agreed on the importance of privacy interventions, they also emphasized that "context matters enormously." Interventions aimed at restricting the visibility of civilian information may be most valuable in contexts where individuals are at risk of being targeted on the basis of their affiliations or speech. The archetypal example would be a setting such as Afghanistan during the 2021 Taliban takeover, where an individual's allegiances or affiliations may not be immediately obvious but where information could be readily gleaned through the individual's social media presence. Khalida Popal, Afghanistan's former women's soccer captain, for example, urged Afghan women at the time to "take down their names, remove their identities, take down their photos, for their safety," noting that obscuring identity was at once essential but also "painful for me, as an activist and someone who stood up to achieve and earn that identity."[43]

By contrast, interviewees underscored that these interventions may be less useful in settings where the identifying characteristic that makes an individual a target cannot be readily concealed. In situations where individuals are targeted on the basis of their ethnicity or religion, for example, they may be at risk in ways that privacy interventions cannot readily obscure, such as on the basis of surnames in a context where names may indicate an individual's ethnic group. In these settings, the ability to have a fully pseudonymous account may be more useful in protecting individual privacy. In addition, in contexts where perpetrators possess sophisticated technological and surveillance resources, interviewees warned that it may be difficult to persuade civilians that privacy tools are sufficient to thwart perpetrators' ability to obtain their private information.

Other interviewees felt that existing privacy interventions did not go far enough, noting that the bios of social media users often remain public even for those who choose to obscure other elements of their online profiles. Visible information may continue to show an individual's area of work, title, or organizational affiliation, each of which can potentially put them at risk in certain contexts, such as in Taliban-controlled Afghanistan. In other instances, interviewees called for greater control over what elements of their profiles would be obscured, noting that human rights activists may want to remain public while obscuring their family or friend connections from potential perpetrators. They emphasized the importance of tools that enable users to view their profile from the perspective of a stranger, so that they can make adjustments as necessary.

Interviewees also described the importance of communicating these features to vulnerable populations, noting that some companies have developed useful tools that affected communities do not know about. At the same time, interviewees emphasized the need to ensure that the communications about these tools do not overpromise to people who are at risk, and that they clearly articulate the limitations of what is being offered.

Decisions relating to these features may present tensions between proactively protecting individual privacy and offering individual agency concerning what people may want to share on social media. Although personal information available on social media platforms may put people at risk, their digital presence may simultaneously be enabling essential coordination and information sharing. When elements of an individual's profile are locked down or made less visible, for example, making connections may become more difficult for them, which may be in tension with other needs in a moment of crisis, such as trying to access resources or locate other members of their community. Interviewees expressed particular concern about platforms removing features for vulnerable individuals who may not have had an opportunity to provide input.

Interviewees at social media companies expressed the need for guidance as to when they should deploy privacy interventions to support atrocity prevention, while others called for greater clarity from social media platforms on the criteria they use for activating those features.[44] Given the trade-offs inherent in privacy features, some felt that all users might benefit from having greater agency to hide or obscure their digital information, rather than making these features available only in select settings. At the same time, platforms must often make difficult decisions about whether certain information should be made public or private as a default, taking into consideration those who cannot access their accounts to make a needed change, who do not know the tools available to them, or who are preoccupied with more pressing physical security concerns.

It was also suggested that interventions aimed at restricting visibility be complemented with digital literacy initiatives so that people could make more informed choices about what to share, as well as available privacy settings and tools. Others, however, noted that digital literacy and security tools largely place the burden back on civilians. Interviewees emphasized the need for these interventions to be easy for people to use, noting the challenges in communicating with crisis-affected populations.

Other possible privacy features referenced during consultations included "app cloaking," which disguises the icons for social media apps in case an individual's phone is confiscated, as with Grindr's Discreet App Icon feature.[45] Interviewees also suggested giving users the option to set up dummy profiles, so that they would have the option to show a less dangerous version of their true social media profile to potential perpetrators if detained or arrested. Another idea referenced was the creation of a "panic button," a feature that has been long requested by civil society groups in various contexts.[46] This type of feature could rapidly delete a user's digital history or direct users to external resources if activated.[47]

More generally, interviewees urged social media companies to think about privacy broadly in atrocity risk settings and to avoid overly narrow definitions of what constitutes private information, referencing the "mosaic effect" of information.[48] Interviewees noted that, while protecting vulnerable civilian populations is an atrocity prevention strategy in its own right, protecting individual privacy can also be an effective means of degrading perpetrator capacity, given how perpetrators can target individuals through information on social media. They also recommended that personally identifiable information be tightly protected, particularly against requests by governments that may use that data to target civilians.

Core guidance and recommendations:

- Platforms should explore interventions to proactively restrict the visibility of digital information that could be used to target civilians in atrocity risk contexts, such as their affiliations or location history.
- Privacy interventions aimed at protecting civilians should be carefully balanced against civilians' potential interests in sharing information in atrocity risk settings. Wherever feasible, civilians should be afforded agency over their digital presence.
- Platforms should carefully review features through which civilians' digital information may be visible without their consent, or where they may not realize they gave prior consent.
- Platforms should ensure that vulnerable civilian populations can readily understand how to temporarily deactivate or delete their social media accounts should they deem it necessary for their protection.
- Platforms should communicate the available privacy tools to vulnerable populations in advance of crises and should clearly articulate relevant limitations, to avoid overpromising to people who are at risk.

B. Interventions Focused on Account Security

Interventions Focused on Securing Online Accounts	
Description	Interventions aimed at protecting social media users against hacking, impersonation, and account takeover efforts
Theory of Change	Civilian protection includes ensuring that civilians' digital information cannot be obtained and used to target them through hacking and impersonation campaigns. This can in turn protect others who may be misled by hacked and impersonated accounts.
Examples	• Account security push notifications, deployed in Ukraine • End-to-end encryption channels

In the weeks that followed Hamas' devastating attack on Israel of October 7, 2023, the International Committee of the Red Cross (ICRC) released the following statement: "*We are aware that some individuals are impersonating the ICRC and asking for personal information from families of the hostages and missing people. Please note that we are not currently contacting families over the phone to ask for personal information or photos. Also, we do not call families using messaging apps.*"[49]

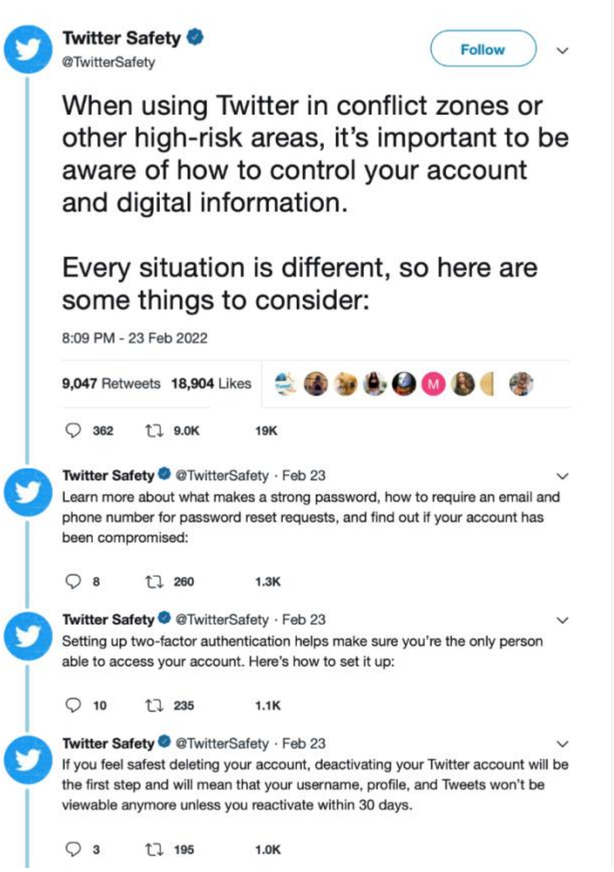
Social media platforms offer new vectors for sensitive individual information to be compromised and for people to be misled, with grave consequences. Information that can be accessed through hacking or impersonation campaigns can be used to target people or to further the commission of violence. In the context of Afghanistan, for example, a consultation interviewee recalled the Taliban's impersonation of aid workers to request sensitive information from civilians, under the auspices of helping them evacuate from the country. This information was then reportedly used to kidnap and retaliate against targets.

> ***"When you lose your bank card, you can suspend your account. Social media is a lot more valuable because you have a lot more information in there—your friends, family, and colleagues—you should be able to take control of it."***
>
> - Afghan activist

According to interviewees, hacking efforts may be particularly prevalent in crisis settings, where people may be displaced from their belongings or devices, leaving them exposed if perpetrators acquire their possessions. Perpetrators who hack civilians' social media accounts may be able to obtain sensitive information from messaging history, friend lists, and other digital information that can expose civilians to further targeting and retaliation.

Relevant interventions referenced by interviewees included the use of push notifications that prompt users in high-risk settings to secure their social media accounts. In the midst of the Russian invasion of Ukraine, for example, some social media platforms pushed out account security information to users in the region, outlining steps they could take to protect their accounts. This intervention may be particularly important for highly vulnerable and particularly authoritative accounts in a crisis or atrocity risk setting. Other interviewees, however, felt investing in account security should be "business as usual," across contexts.

Interviewees also referenced longstanding debates related to the end-to-end encryption (or E2EE) of messaging features on social media, as a way to guard against hacking campaigns. E2EE scrambles messages such that they can only be deciphered by the sender and intended recipient, preventing third parties—including social media companies themselves—from viewing messages.[50] Some interviewees emphasized the importance of E2EE of messaging features as a matter of default for safety reasons, and as a safer way for affected communities to



Twitter Safety
@TwitterSafety

When using Twitter in conflict zones or other high-risk areas, it's important to be aware of how to control your account and digital information.

Every situation is different, so here are some things to consider:

8:09 PM - 23 Feb 2022

9,047 Retweets 18,904 Likes

Twitter Safety @TwitterSafety · Feb 23
Learn more about what makes a strong password, how to require an email and phone number for password reset requests, and find out if your account has been compromised:

Twitter Safety @TwitterSafety · Feb 23
Setting up two-factor authentication helps make sure you're the only person able to access your account. Here's how to set it up:

Twitter Safety @TwitterSafety · Feb 23
If you feel safest deleting your account, deactivating your Twitter account will be the first step and will mean that your username, profile, and Tweets won't be viewable anymore unless you reactivate within 30 days.

In February 2022, around the invasion of Ukraine, Twitter shared guidance on account security when using the platform in conflict zones, including on how to delete or deactivate accounts if necessary. *Twitter, accessed via Wayback Machine.*

document atrocities. Others noted that encryption limits investigative teams' ability to study and address dangerous content being shared on messaging surfaces.

Core guidance and recommendations:

- Platforms should ensure they put in place and stringently enforce policies prohibiting account impersonation in atrocity risk settings.
- Platforms should explore opportunities to proactively communicate information to civilians about how to best secure their online accounts, such as through push notifications or prompts.
- Platforms should clearly communicate their policies around the use of encrypted or unencrypted features, so that users readily understand the security of the tools they use in atrocity risk settings.

C. Interventions Focused on Surfacing Crisis Resources and Credible Information

Interventions Focused on Surfacing Crisis Resources and Credible Information	
Description	Tools or interventions aimed at connecting social media users to crisis resources and/or credible information
Theory of Change	Ensuring that civilians are able to access information about crisis resources can enable vulnerable civilian populations to obtain lifesaving information and coordinate actions to protect themselves in moments of crisis. **OR** Ensuring that civilians are able to access reliable information about evolving developments can prevent misinformation or disinformation from inciting violence.
Examples	• Creating centralized landing pages or information hubs that compile authoritative information in atrocity risk settings • Modifying approaches to ranking and amplifying information to align with needs in atrocity risk settings • Amplifying content from credible accounts, such as reliable media or civil society organizations • Providing ad credits to credible local organizations • Making use of push or pop-up notifications or "nudges" to direct people to important resources or news items

Among the most powerful opportunities afforded by social media in atrocity risk settings is the ability to connect vulnerable civilian populations with crisis resources, including credible information. Interviewees expressed significant interest in this category of interventions, which can include both platforms' general approaches to ranking content and specific efforts to connect individuals with information to support their protection.[51] Enabling access to crisis resources, such as information about evacuation routes or humanitarian aid, can contribute to the protection of vulnerable civilian populations by helping them withstand or avoid attacks. In addition, surfacing reliable information about evolving developments in atrocity risk contexts may support self-protection efforts or help prevent misinformation or disinformation from inciting violence. Interviewees in this report's consultations emphasized the importance of social media for the real-time dissemination of information in crisis settings, noting that, although traditional avenues exist for sharing this information, social media offers opportunities to direct people more quickly toward “what they need to know, where they need to go, and who can help people get what they need.”[52] These opportunities may be particularly important in locations with limited independent media, such as authoritarian contexts where journalists lack the freedom to openly report on atrocities or emerging risks.

> ***“Throughout the history of social media, it has been deployed as a way to share information rapidly in moments of crisis to allow people to take action where they are.”***
>
> - Trust and safety practitioner

In atrocity risk settings, interviewees stressed that social media platforms’ approaches to the ranking and integrity of information take on heightened importance. These settings may warrant review of the way social media platforms rank and surface information, to ensure that they align with local needs and concerns. For example, interviewees called on platforms to identify indicators of more reliable information and to use them to ensure that more authoritative information is surfaced to users over less credible information. While modifications may be helpful generally, they may offer particular benefit in atrocity risk contexts, where, for instance, civilians need to readily access statements put out by humanitarian agencies, peacekeeping missions, and local leaders and filter more authoritative communications from rumors and hearsay. In discussing how this might be operationalized, some interviewees called for platforms to prioritize content based on external indicators of quality or credibility rather than on “user engagement,” with the aim of privileging more authoritative information on timelines or user feeds over information that may be less trustworthy. This may include, for example, bringing content from credible journalists and news organizations to the top of a user’s feed in an atrocity risk setting.

Interviewee views differed on the most valuable approaches to ranking information in atrocity risk settings, including whether a chronological or nonchronological content feed would better surface important information for civilian protection. Some interviewees noted that under schemes that de-prioritize chronology (in other words, schemes that rank information in a nonchronological fashion, such as by how much users are engaging with it), more recent posts can be buried, which could adversely affect users needing time-sensitive information about quickly changing events on the ground. Others observed that chronological feeds benefit people churning out high volumes of posts (by putting their content at the top of news feeds solely due to recency), which can inadvertently privilege recent but inaccurate information, or misinformation campaigns that generate high volumes of low-quality information. Other interviewees suggested enabling social media users to toggle back and forth between chronological and nonchronological content feeds, to ensure they could access information important to their needs and protection.[53] This issue may particularly warrant further research, with attention to the unique needs of people in atrocity risk settings.

Other interventions may be aimed at intentionally directing users to specific credible information or resources that could support their protection, though how well-placed social media companies are to identify the most useful content remains an open question. Interviewees suggested that, with relevant partnerships in place, this could be done through push or pop-up notifications or by developing “nudges” asking people if they need help or support. Platforms also could redirect users who search for certain terms to relevant resources, a tactic which has previously been used in the context of natural disasters and to counteract Holocaust denial.[54] Interviewees noted, however, that in some settings, redirects “can be stressful,” in that they take users away from primary sources at a moment when they may be trying to access information. Interviewees also expressed concern about how quickly a redirect would be modified to respond to changing events in the context of evolving developments.

This category of interventions could also take the form of amplifying credible accounts, such as those of reliable media, civil society organizations, or in some settings, state accounts.[55] Interviewees also noted that amplifying select accounts can be problematic, either because of the content those accounts ultimately post or because the way they may be perceived on the ground may differ from how they are viewed by decision makers at social media companies. Even content from reputable humanitarian organizations, one interviewee noted, can be received poorly by local populations that may feel such organizations should be doing more in a given setting.

Humanitarian or civil society organizations themselves can make use of social media opportunities to surface important information for civilian protection. The Signpost project, for example, leverages social media channels to provide critical, accurate information to vulnerable civilian groups, adapted to suit local contexts and needs.[56] Relatedly, interviewees referenced that platforms have occasionally granted ad credits to credible organizations in crisis settings, to enable them to push information themselves to users free of charge. Interviewees emphasized, however, that ad credits should ideally be paired with support and resources to help organizations leverage the social media space effectively, noting that civil society groups may lack not only money to promote critical information, but also the resources to create useful content in the first place.

Some platforms have also attempted to develop or fund the creation of centralized landing pages or resource centers that compile authoritative information in moments of risk. Amid rising tensions between Israelis and Palestinians in 2022, for example, Twitter used its Moments feature to develop and aggregate credible news articles related to ongoing air strikes.[57] In other instances, it collaborated with the Associated Press and Reuters to create user prompts that linked to digital public service announcements aimed at elevating credible information in crisis settings.[58] These information hubs or landing pages, however, are difficult and resource-intensive to keep current, interviewees warned. Credible sources are difficult to identify and vet in real time and can change quickly, rendering interventions “fraught with risks”—both for the company and for vulnerable individuals being served information. To mitigate these concerns, interviewees recommended that these pages or hubs be “owned” by external groups, such as humanitarian or protection organizations, and that platforms enable those organizations to update information directly.

Several considerations are relevant when adapting these interventions to atrocity risk contexts. As one participant articulated, “There’s going to be a lot of information that’s unverified but is important to keeping people safe, and social media plays an essential part in that information circulating.” As a result, platforms need to develop principled approaches for how they will identify credible and useful information for civilian protection and under what circumstances it will be amplified. Interviewees also emphasized the importance of considering the timing of these interventions, noting that information and credibility change over time. In addition, choosing to point users to particular information can be seen as taking a perspective on it. As one participant put it, getting these interventions right is “difficult, and the risks of getting it wrong are serious.”

Interventions that amplify the visibility of essential information for civilian protection, such as the location of bomb shelters or evacuation corridors, may also render it simultaneously more visible to potential perpetrators. Although social media can assist in surfacing resources, interviewees cautioned that unintended recipients of that information may be able to use it as well, rendering already vulnerable communities even more vulnerable. Drawing on experiences in non–social media contexts, interviewees recalled that Google elected to temporarily remove the ability for users to submit locations on Google Maps in Russia, Ukraine, and Belarus "out of an abundance of caution," after false rumors began circulating that the product was being used to coordinate Russian air strikes.[59] Similarly, those using social media features to highlight humanitarian or refugee resources, for example, should be mindful of the potential for this information to be misused by perpetrators, or by those looking to profit by, for example, diverting aid or "taxing" humanitarian workers at checkpoints. Given these tensions, interviewees emphasized the need to afford vulnerable communities control and agency in opportunities to surface crisis resources on social media, to address risks associated with information sharing.

Core guidance and recommendations:

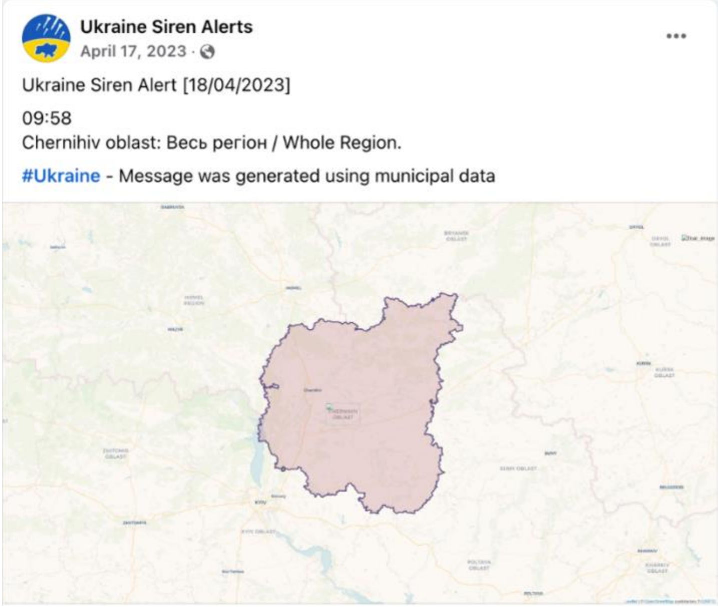
- In atrocity risk settings, the way information is ranked and prioritized takes on heightened importance. Platforms should review approaches to the ranking and amplification of information to align with needs in atrocity risk settings.
- Platforms should develop principled approaches for how they will identify credible and useful information for civilian protection, and under what circumstances specific information will be amplified.
- Platforms should consider affording users choice in how content is prioritized in user feeds so they can quickly identify resources and information important to their protection.
- Platforms should, in partnership with relevant organizations and humanitarian agencies, explore opportunities to direct users to credible information or resources that could support their protection.
- Platforms should explore interventions that would enable users themselves to designate information as critical and to have that information amplified in atrocity risk settings.
- Platforms should explore opportunities to afford vulnerable communities greater control and agency in opportunities to surface crisis resources on social media, to mitigate risks associated with information sharing.

D. Interventions Focused on Early Warning and Awareness

Interventions Focused on Early Warning and Awareness	
Description	Interventions that make use of social media to communicate warnings about the risk of mass atrocities
Theory of Change	Social media may be used to communicate warnings (either to or between civilians at risk, or to policy makers), with a view to influencing outcomes on civilian protection.
Examples	● Publishing emergency air raid alerts on social media ● Using social media posts to warn people about safe/unsafe locations in Libya, or potential targets for air strikes in Syria

Digital tools may also be used to support civilian protection by communicating warnings about the risk of mass atrocities, either to those in vulnerable communities, to high-level decision makers, or both, with a view to prompting action and changing outcomes.[60] Importantly, interventions in this category may be available to those outside of social media platforms, such as actors in international organizations and civil society, and members of affected communities.

Social media may be used, for example, to communicate warnings among and between local communities.[61] Interviewees referenced efforts in Tigray, Ethiopia, to blast alerts on social media if bombs were reported, or if soldiers were reported in certain areas. Similarly, the Ukraine Siren Alerts or UASA system collects data on air raid alerts from across the country and publishes it on social media. Particularly in instances where physical siren alerts have failed, these social media warnings can be critical.[62] In Libya, citizens have used social media to share imminent warnings, using the phrase "red light" to alert one another about the locations of militia fighting or other dangers.[63] Relatedly, the Sentry system in Syria leveraged social media, along with data from remote sensors and civilian volunteers, as part of an alert system to warn civilians of potential targets for Syrian warplanes.[64]



Ukraine Siren Alerts
April 17, 2023 ·

Ukraine Siren Alert [18/04/2023]

09:58
Chernihiv oblast: Весь регіон / Whole Region.

#Ukraine - Message was generated using municipal data

The Ukraine Siren Alerts system compiles data on imminent air attacks from across the country, and shares warnings on social media. *Ukraine Siren Alerts on Facebook*

Social media may also be used to communicate warnings from local communities to the international community, including to high-level decision makers who may be able to affect outcomes. Existing literature highlights the "essential communicative role that civil society can play in the midst of violence," noting that civilians and civil society actors "generally are the most knowledgeable about local conditions, especially in more remote areas that are often the site of mass atrocity episodes."[65] Social media may serve as a mechanism to enhance the dissemination of information about emerging atrocity risks and to bring the experiences of those in remote atrocity risk settings to the attention of the international community.

During the consultations for this report, interviewees reflected on the need for social media platforms to ensure that content moderation policies permit the sharing of information that could support early-warning efforts. They recalled instances in which social media content depicting violence that was aimed at warning individuals about safe and unsafe locations was removed for violating content moderation policies prohibiting graphic content. Given that these policies may prevent or discourage people from sharing essential information, interviewees suggested that atrocity risk settings may warrant modifying the way policies are applied. For instance, platforms could permit certain forms of content depicting violent events, but put them behind warning labels or "interstitials," or they could leverage technical interventions to help mitigate psychosocial harm to users, such as depicting media in grayscale, with low audio, and/or with images blurred.[66] Interviewees suggested that platforms might also ensure that content remains available to researchers and civil society organizations.

Platforms might also amplify content communicating early warnings, perhaps by prioritizing that content on user feeds, or by partnering with third-party entities engaged in early-warning initiatives. Interviewees referenced the opportunity to learn from past initiatives aimed at early warning in non–social media contexts, such as Google's development of a rapid air raid alert system for smartphones in partnership with the Ukrainian government, or the use of radio to broadcast information about armed groups in Uganda and the Central African Republic, enabling communities to organize self-protection efforts.[67]

When adapting these interventions to the social media space, interviewees cautioned that warnings issued by a social media company may lack credibility, leading civilians not to trust the information. Interviewees also noted that while partnerships with government officials (for example, to alert civilians about air raids) may be helpful in some contexts to support early warning, it would not be a viable option in atrocity risk contexts where the government is itself a perpetrator.

Core guidance and recommendations:

- Platforms operating in atrocity risk settings should review the way that content moderation policies on graphic media or violent content are applied and enforced. They should also explore the use of technical interventions to mitigate psychosocial harm associated with the viewing of graphic content, such as the use of interstitials, grayscale, or image blurring.
- Platforms should explore opportunities to support early-warning initiatives by trusted third-party entities, but they should implement safeguards to carefully assess information credibility and timeliness.

E. Interventions Focused on Enhancing Communication and Coordination Capabilities

Interventions to Support Communication and Coordination	
Description	Interventions that expand or enhance civilians' ability to communicate and coordinate
Theory of Change	Supporting open communication and coordination between civilians will enable them to better avoid or withstand atrocities.
Examples	● "Groups," "Communities," or group messaging features ● Features that help users connect to social media platforms via proxy servers, bypassing restrictions on internet access

A restricted information space is a risk factor for mass atrocities. The United Nations (UN) *Framework of Analysis for Atrocity Crimes* flags "imposition of strict control on the use of communication channels, or banning action to them," as an enabling circumstance for the commission of atrocity crimes, contributing to an environment conducive to their commission.[68] Widespread restrictions on access to information "[enable] the authorities to reinforce prejudicial policies, incite further xenophobia and identity-based divisions and perpetrate widespread human rights violations and crimes against humanity with impunity."[69]

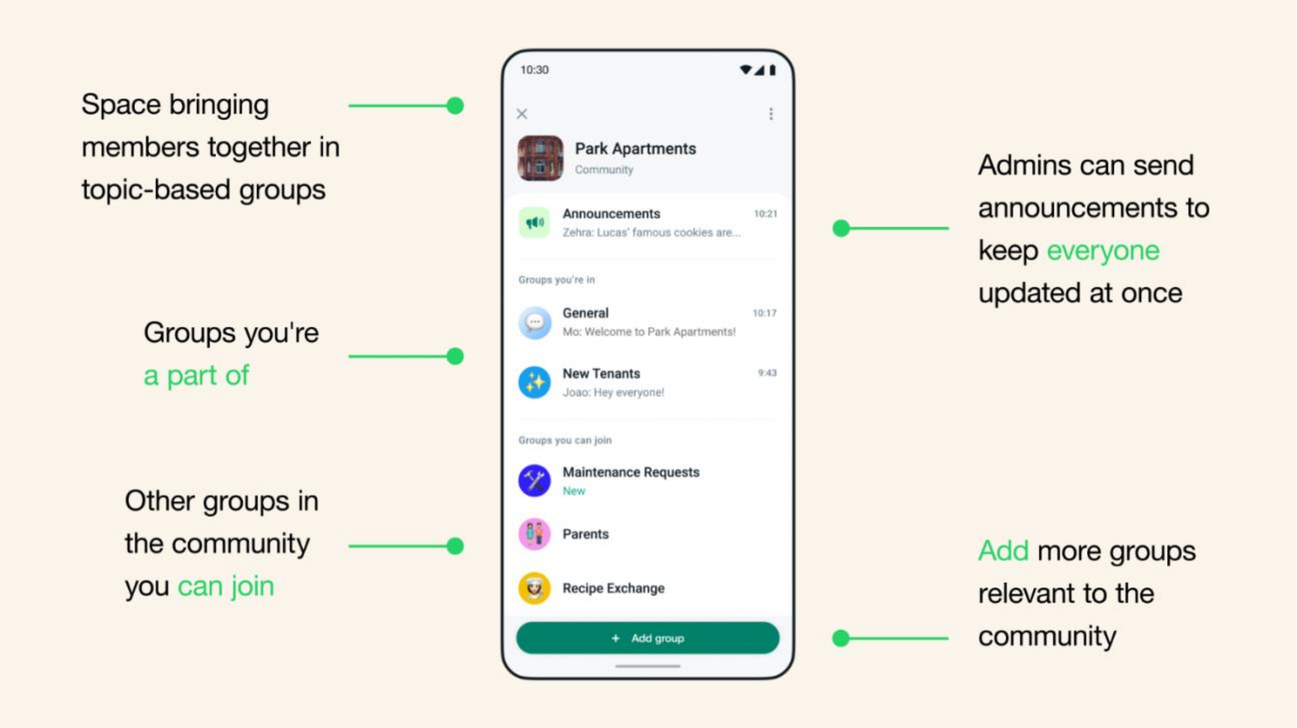
This category of interventions is aimed at supporting civilian communication and coordination in periods of heightened atrocity risk. Although the presence of social media in itself may support information sharing, this category covers interventions that go beyond social media's mere presence and looks instead at specific social media tools and features that can enable communication and coordination between civilians.

Supporting open communication and coordination among civilians in moments of risk may enable them to better withstand or avoid atrocities. Recent scholarship has challenged the portrayal of civilians as lacking agency in atrocity settings, noting that, "far from hapless victims, civilians are able to protect themselves and their communities despite the circumstantial, social, and institutional constraints of mass violence."[70] Civilians often, however, face serious obstacles to communicating and coordinating during crisis events, including internet blackouts and governments throttling or blocking access to social media. These obstacles are often compounded by restrictions on media freedom and demands for social media platforms to remove certain forms of digital content. This category explores the specific tools or interventions that can counter these obstacles and support expanded communication and coordination among civilians.

"The ability of civilians to communicate and connect with one another in private settings is fundamentally one of the most important civilian protection efforts."

- Expert on social media and human rights

This category can take the form of features that specifically enable group communication. Facebook's "Groups" feature, for example, enables subsets of users to communicate in a digital forum that can be either public or private, while X's "Communities" feature offers a space for public-facing communication, but with only members of a particular community permitted to engage in the discussion.[71] TikTok offers Group Chat, enabling up to 32 participants to engage in group discussion and shared videos.[72] WhatsApp, a direct messaging app, now also offers channels that, like social media, offer users the ability to communicate in private groups, as well as to organize constellations of groups in "communities."[73] Though these features have broad utility, the spaces they create may offer particular benefit in supporting civilian communication and coordination in atrocity risk settings.


Space bringing members together in topic-based groups
Groups you're a part of
Other groups in the community you can join
Admins can send announcements to keep everyone updated at once
Add more groups relevant to the community
10:30
Park Apartments
Community
Groups you're in
Groups you can join
+ Add group

Announcements 10:21
Zehra: Lucas' famous cookies are...

General 10:17
Mo: Welcome to Park Apartments!

New Tenants 9:43
Joao: Hey everyone!

Maintenance Requests
New

Parents

Recipe Exchange

WhatsApp shares guidance on the use of its 'Communities' feature, which offers users the ability to form topic-based groups. The feature may offer value to support communication and coordination in crisis or conflict settings. *WhatsApp*

Consultation interviewees highlighted the importance of social media features that create “spaces for self-organization,” as well as opportunities to exchange warnings about emerging risks. Even in settings where activists believe their messages are being monitored, interviewees emphasized that they may still choose to share information on social media, calling on their communities to be ready to move to a particular location, for example, or sharing information on where and how to receive further instructions. The use of group or community features was referenced as particularly valuable in atrocity risk settings where the cost of speaking out about emerging risks in offline public spaces becomes too high.

In addition to self-organization and the exchange of warnings, interviewees highlighted the power of social media to “build connection between groups that aren’t always speaking to each other,” such as interfaith groups that may be able to engage in dialogue in moments of risk. They emphasized the need for platforms to give the administrators of these groups the tools they need to moderate healthy discussions.

Platforms have also experimented with opportunities to support connectivity in settings where there are heightened restrictions on communication, such as internet blackouts or platform shutdowns. Amid a series of internet shutdowns during the Iranian women’s movement, for example, WhatsApp launched a feature to help users connect to the platform via proxy servers, bypassing government controls on internet access.[74] Similarly, as Russia blocked access to social media platforms during the war in Ukraine, Twitter launched a “Tor onion service” to allow users to access the platform even when it is blocked in a certain country, through an anonymized service that also helps guard against surveillance.[75] These interventions were referenced by interviewees as important opportunities to ensure civilians could remain connected and exchange information in atrocity risk contexts.

Finally, interviewees also proposed interventions that would allow users themselves to communicate and exchange information about their own safety and that of loved ones, and to have that information pushed up using algorithms. For example, Facebook's "mark as safe" feature, which has been deployed in several contexts, allows users to let their network know they are safe if they are located in a given geographic area during a crisis.[76]

As with other features, social media features that enable coordination and communication for self-protection are subject to risk of abuse, with serious consequences. In addition to moderating these forums, platforms should give further consideration to empowering users to make their own decisions about the visibility and accessibility of these spaces. Administrators should be empowered to approve or remove members of spaces, and limitations on the privacy of relevant forums should be clearly communicated.

Core guidance and recommendations:

- Platforms should be mindful of the value of features that enable group discussion for coordination and communication between civilians in atrocity risk settings. When considering modifying or updating these features, platforms should take particular care in assessing the needs of those in atrocity risk settings that may be using them for this purpose.
- Platforms should pay particular attention to the moderation of group discussion forums in settings where there is a heightened risk of mass atrocities. This may include ensuring that the administrators of discussion forums have the tools and resources they need to support moderation, such as the ability to approve or remove members from digital spaces.
- Platforms should ensure that the visibility or privacy of group discussion forums is clearly communicated to all participants.
- In atrocity risk settings where there are heightened restrictions on communication, such as internet blackouts or platform shutdowns, platforms should consider exploring opportunities to expand access to social media, particularly for vulnerable or isolated communities.

VI. THE DIGITAL TOOLBOX AND PERPETRATOR CAPACITY

On January 11, 1994, Major General Romeo Dallaire, the UN force commander in Rwanda, sent a cable to United Nations headquarters in New York. Therein, he detailed reports from an informant describing suspected plans to exterminate the country's Tutsi population and identifying the location of a "major weapons cache." General Dallaire noted that the informant was prepared to raid the arms cache within 36 hours if the United Nations would guarantee the protection of himself and his family.[77]

In a now-infamous response cable, General Dallaire was told by then-chief of UN Peacekeeping Kofi Annan to stand down.[78] For years to come, this moment would be regarded as a critical missed opportunity to disrupt the capacity and ability of perpetrators to carry out a genocide in Rwanda—in which those weapons were used at breathtaking scale and pace during the spring of 1994.

Degrading the capacity of perpetrators to commit mass violence is a core atrocity prevention strategy. The commission of mass atrocities depends on perpetrators having certain material and operational capacities, such as weapons, finances, and communication channels.[79] Traditionally, tools to support this strategy have included policy actions like financial sanctions, arms embargoes, and the disruption of communication networks, but perpetrators' ability to leverage digital spaces today introduces the need for novel interventions. In the digital era, weapons caches need not be physical; they can also include the repository of social media features at risk of being weaponized by perpetrators active in the digital environment. Because perpetrators have appeared to effectively use social media to disseminate exclusionary ideologies, deceive communities, and incite violence, comprehensive efforts to degrade perpetrator capacity should assess opportunities to disrupt the use of those tools.

In the digital era, weapons caches need not be physical; they can also include the repository of social media features at risk of being weaponized by perpetrators

The theory of change and the framing of this section draw, in part, on work led by experts in disrupting online influence operations who have articulated what they refer to as an "online operations kill chain."[80] A "kill chain" helps "identify the sequence of activities that attackers go through in their operations and looks for ways to disrupt them."[81] Applied to the mass atrocity space, the analytical framework of a kill chain can help identify opportunities to disrupt atrocity perpetrators that are weaponizing social media to commit mass violence. As set out in this section, the framework may include opportunities to prevent perpetrators from gaining a foothold on social media platforms, organizing and coordinating on social media, engaging with other users at scale, and mobilizing bystanders to mass violence. Finally, this section also describes a set of "break glass" interventions that might be deployed as a last resort to prevent perpetrators from using social media to further violence.[82]

A. Interventions Aimed at Preventing Perpetrators from Acquiring a Foothold on Social Media Platforms at Scale

Interventions to Prevent Perpetrators Gaining Foothold on Platform at Scale	
Description	Interventions aimed at preventing perpetrators from setting up a large presence on social media platforms
Theory of Change	Preventing perpetrators from establishing or maintaining extensive networks of accounts will make them less able to weaponize social media in furtherance of atrocities (such as to incite or coordinate violence).
Examples	• Prevent perpetrators from registering social media accounts. • Expand detection of coordinated networks of accounts of potential perpetrators. • Designate and ban perpetrators under violent individuals and organizations policies. • Deplatform perpetrators, or subject them to heightened monitoring against content moderation policies.

In the early stages of an online operation, threat actors—including atrocity perpetrators—will need to "set up" their operation on a social media platform. Just as they obtain weapons, physical locations, and bank accounts in the physical realm, perpetrators will often need to acquire or create social media accounts.[83] This category describes interventions aimed at preventing perpetrators from, as one interviewee put it, "acquiring a foothold" on social media platforms at scale as a first line of defense in degrading their digital capacity. The underlying theory of change is that preventing perpetrators from establishing or maintaining extensive networks of accounts on social media platforms will render them less able to weaponize digital spaces to commit or further mass atrocities.

> ***"At the most fundamental level—at risk of sounding simplistic—any online operation has to be able to get online."***
>
> - Ben Nimmo and Eric Hutchins, "Phase-based Tactical Analysis of Online Operations"

A logical place to begin in preventing atrocity perpetrators from gaining access to social media platforms at scale is preventing account registration in the first place. As one intervention in this category, interviewees recommended preventing automated account registration—that is, preventing users from registering a new account through automated means.[84] Others, however, pointed to contexts in which perpetrators had repurposed or acquired old accounts, and these interviewees urged interventions aimed at disrupting perpetrator networks to avoid focusing solely on new account registration. In settings where a proliferation of fake or inauthentic accounts impersonate others and/or spread misinformation, taking enhanced measures to ensure that accounts belong to real individuals can help prevent perpetrators from setting up a large and inauthentic platform presence.

Once perpetrators have registered social media accounts, further interventions may be used to disrupt coordinated networks of accounts, which are often used to influence and manipulate social media users. In Burma, for example, networks of coordinated inauthentic accounts that purported to be independent news and opinion pages were used to "covertly push the messages of the Burma military."[85] During the consultations for this paper, interviewees urged platforms to enhance their in-house investigative capacities to detect coordinated networks and to remove those networks when they violate platforms' terms of service, as when they are made up of inauthentic accounts. As part of this effort, interviewees emphasized the importance of platforms taking signals from local civil society groups, who often have insights about perceived bot networks but lack the technical capacity to conclusively detect them.[86]

Interviewees identified several other interventions aimed at preventing potential atrocity perpetrators from acquiring access to platforms at scale or at disrupting their access once obtained. One such intervention is platforms' ability to designate (and then ban) certain accounts as belonging to "dangerous individuals and organizations." Meta, for example, has a policy prohibiting "organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence" on the platform—and during the 2023 conflict in Sudan, suspended the account for the Rapid Support Forces (RSF) under this policy, as well as the account of RSF leader Hemedti.[87] TikTok, similarly, has a policy prohibiting the use of the platform by "individual perpetrators of mass violence."[88] According to interviewees, these policies can be effective in preventing perpetrators from using social media platforms to spread propaganda or gain legitimacy in the eyes of the international or local community.

Another approach is to deplatform individual perpetrator accounts, an intervention which generally refers to denying someone the ability to post on a platform, typically by suspending their account.[89] Debate is ongoing about the advantages and disadvantages of deplatforming individuals from social media (as well as when and whether they should ever be "rehabilitated" by restoring their account access), but interviewees broadly supported using this intervention for perpetrators who repeatedly violate content moderation policies.[90] Interviewees noted that the policy is particularly effective when used to deplatform influential public figures (who may be able to use social media to broadcast to large groups of followers). Interviewees also, however, recognized the limitations, as the same individual or their representative might set up alternative accounts on a different platform if banned from one platform. Nevertheless, deplatforming atrocity perpetrators (i.e., preventing them from having an account at all) rather than merely deamplifying their content may have advantages, because dangerous content that is deamplified can be re-shared by other actors. As with account suspensions under a dangerous organization policy, deplatforming individual perpetrators on the basis of their online conduct terminates their ability to post content and to weaponize social media to support atrocity crimes.

Opinions among interviewees differed, however, about whether individuals should be deplatformed on the basis of online behavior (i.e., for repeatedly violating content moderation policies), or whether individuals should ever be deplatformed solely on the basis of offline behavior, such as if they are suspected or convicted of committing war crimes or have a documented history of human rights violations. Some interviewees suggested subjecting the accounts of potential atrocity perpetrators to heightened monitoring by content moderators, to ensure that platform policies are stringently enforced, given these individuals' offline behavior. Others suggested that platforms might increase penalties for violations of content moderation policies for all users in settings with heightened atrocity risks. This may mean, for example, that policies that typically require multiple "strikes" or offenses before a user would be suspended from the platform might have a lower threshold before users are suspended in periods of heightened atrocity risk, given the risk of harm. At the same time, interviewees noted the tension between suspending individuals from social media platforms and the need to preserve a record of their content for accountability purposes, an issue addressed later in this report.

Disrupting networks and accounts belonging to potential perpetrators of atrocity crimes can carry unintended consequences. For example, to the extent that platforms make these interventions publicly known, this can put social media companies at risk, particularly if they have employees based in the country. In addition, interviewees warned that platforms may be perceived as taking sides in a conflict, by virtue of removing networks of accounts or deplatforming individual leaders and organizations. This can lead to their services being blocked or throttled in the affected country, which may inhibit access to information for at-risk communities, hampering early warning, information sharing, and coordination for protection.

Core guidance and recommendations:

- Platforms should explore interventions that can prevent atrocity perpetrators from setting up networks of inauthentic accounts. This should include both efforts to prevent the registration of new accounts and review of older accounts that may exhibit suspicious behavior.
- Platforms should enhance their in-house investigative capacities to detect and remove coordinated networks of inauthentic accounts that may be used by atrocity perpetrators.
- Platforms should, in atrocity risk settings, proactively review potential perpetrators against criteria for designation under policies that target violent organizations.

- Platforms should explore heightened monitoring of accounts of atrocity perpetrators and more stringent enforcement of content moderation policies given these individuals' offline behavior.

B. Interventions Aimed at Disrupting Perpetrators from Coordinating and Organizing on Social Media

Interventions to Prevent Perpetrators from Coordinating and Organizing on Social Media	
Description	Interventions aimed at disrupting perpetrators from using social media to coordinate and organize the commission of violence
Theory of Change	To the extent that digital spaces are being used to coordinate and organize violence, disrupting perpetrators' ability to use social media to advance the planning and organization of violence will degrade their overall capacity to commit atrocities.
Examples	• Enforcing content moderation policies that prohibit weapons sales or promotion of criminal activities • Heightened monitoring of online spaces where perpetrators may be organizing violent tactics, such as groups or pages • Implementing policies prohibiting the use of social media for surveillance

Mass atrocities do not occur spontaneously but are "processes that take time to plan, coordinate, and implement."[91] Across countries and contexts, perpetrators must engage in planning, recruitment, and organization before they are able to carry out violence at scale. To the extent that digital spaces are being used to organize the commission of mass atrocities, logic dictates that degrading perpetrator capacity should include disrupting the organization of atrocities online as well as offline.

This category of interventions includes tools aimed at preventing perpetrators from using social media to coordinate and organize the commission of mass atrocities. The theory of change at play is that these restrictions can reduce perpetrators' ability to use social media to advance the planning and organization of mass violence. This category also draws on the concept of the kill chain, which identifies opportunities to disrupt threat actors from coordinating their activities on social media as a key link in the chain.[92] For example, experts on threat actors in other contexts point to the training of recruits using private online groups, the publication of lists of targets on social media, the use of hashtags, and the use of bot networks to automate posting across accounts.[93] They also reference threat actors' gathering of information on social media to support their planning, such as searching for targets, surveilling journalists and dissidents, and monitoring trending topics.[94]

In many instances, social media companies have policies relevant to prohibiting the abuse of their platforms for the coordination and organization of mass violence. This includes policies prohibiting the purchase or sale of weapons, ammunition, and explosives; policies prohibiting the use of the platform to promote criminal activities; and policies prohibiting the use of the platform by violent extremist organizations.[95] During the consultations for this report, interviewees emphasized the importance of enforcing these policies in countries at risk of atrocities or where atrocities are already under way. They also pointed to the need to better moderate public and semi-public spaces on social media where individuals might be gathering to organize acts of violence, such as in private groups or Pages.[96]

Beyond preventing perpetrators from organizing violence, opportunities may exist for platforms to prevent perpetrators from readily collecting information on social media about potential targets. One intervention proposed was effectively the inverse of Facebook's locked profile feature (discussed previously), in which social media users in a certain country are temporarily limited in the extent of information they can view about other accounts. This intervention would prevent all accounts in a given country or region from being able to view information such as which accounts someone follows or their membership in online groups. Interviewees suggested this could help prevent perpetrators from using a single account they follow to collect information on a chain of other accounts.

Interviewees also emphasized the importance of policies prohibiting the use of social media data for surveillance.[97] X, for example, prohibits monitoring sensitive events such as protests and rallies, investigating or tracking sensitive groups and organizations, and using the platform for facial recognition.[98] Interviewees considered policies of this nature a "meaningful contribution" in instances where protesters or activists in at-risk settings are using social media platforms.

In adapting these interventions to atrocity risk contexts, platforms should be mindful of the need to ensure adequate resourcing and enforcement of policies prohibiting the coordination and organization of violence in settings with a heightened risk of atrocities. They may also need to consider the specific forms of online organization and coordination that perpetrators engage in before the commission of atrocities. To the extent this differs from other forms of violence (such as terrorism or violent extremism), opportunities may exist to better align policies with the needs of affected communities.

Core guidance and recommendations:

- Platforms should ensure that they have policies in place that prohibit the abuse of their platforms for the coordination and organization of mass violence, including but not limited to the purchase and sale of weapons and recruitment to violent organizations.
- Platforms should also ensure that the enforcement of these policies is sufficiently resourced in atrocity risk settings, particularly in online spaces where perpetrators may be gathering.
- Platforms should explore interventions to prevent perpetrators from readily collecting information on social media about potential targets and ensure policies are in place prohibiting the use of social media data for surveillance.

C. Interventions Aimed at Limiting the Presence or Visibility of Dangerous Content in Atrocity Risk Settings

Interventions to Limit the Presence or Visibility of Dangerous Content in Atrocity Risk Settings	
Description	Interventions aimed at reducing the presence or visibility of potentially inflammatory digital content during periods of heightened atrocity risk
Theory of Change	Reducing the presence, audience reach, or visibility of potentially inflammatory digital content limits potential perpetrators' ability to use social media to incite atrocities or further societal divisions.
Examples	• Implementing policies governing how platforms manage dangerous misinformation in crisis settings, such as limiting it from appearing on users' home feed or timeline, or limiting its ability to be re-shared • Deamplifying content that could create a serious risk of harm, such as potentially dehumanizing language or exclusionary ideologies • Using "rate limits" or "forwarding limits," which reduce the number of people to whom a user can forward content at scale

Once perpetrators are active and operating on social media platforms, other interventions could limit the visibility and presence of dangerous content that contributes to violence. The theory of change underlying this category is that by reducing the visibility or presence of potentially inflammatory digital content, platforms can reduce the likelihood that this content could incite atrocities or further societal divisions that contribute to atrocity risk. This category can take several forms, including (but not limited to) policies on mis/disinformation, deamplification interventions, and rate limits.

Policies on crisis misinformation and disinformation and incitement to violence

Because misinformation/disinformation is one of the primary vehicles through which perpetrators can create conditions conducive to mass atrocities, interventions that seek to limit its visibility or presence might play a meaningful role in degrading perpetrator capacity. Throughout this report's consultations, interviewees discussed the importance of platform policies to manage dangerous misinformation in crisis settings, perhaps by limiting users' ability to share, recommend, or amplify unverified and potentially harmful information.[99] These policies can create a basis for platforms to remove misinformation where it can be linked to a risk of harm on the ground.[100] These policies may also provide for engagement between platforms and third-party fact-checkers who can help validate misinformation in real time.

In August 2022, for example, Twitter released a crisis misinformation policy, focused on its handling of false or misleading information that could serve as a pretext for aggression, trigger the displacement of vulnerable populations, affect the ability of humanitarian actors to support members of affected communities, incite the targeting of vulnerable groups, or disrupt peacekeeping operations or ceasefire agreements.[101] Similarly, Facebook's misinformation policy prohibits "misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people," and Snapchat's Community Guidelines prohibit spreading false information that could cause harm.[102]

A range of further interventions can also be deployed to address misinformation, beyond having a policy in place on what types of misinformation are prohibited on the platform. As described by interviewees, these may include, for example, placing warning labels over that content, preventing it from appearing on users' home feed or timeline, or preventing certain terms from being suggested or "typed ahead" when using the search tool on platforms. In instances where removal of the content is not warranted, these "soft interventions" may help degrade perpetrators' ability to spread dangerous rumors or misinformation that can contribute to the incitement of violence. These interventions may also be aimed at limiting the virality of dangerous misinformation to create more time for credible information on emerging events to surface.

Interviewees also highlighted the importance of interventions that combat dangerous misinformation in the form of audiovisual media, rather than text-based digital content. Interviewees reflected, for example, on inflammatory images of destroyed health centers that were circulating on social media during the conflict in Ethiopia, yet had actually depicted buildings in Libya. Misinformation interventions may require the use of an interstitial providing context on the source of certain media, may prevent it from being re-shared if it is going viral in a moment of atrocity risk, or may require its removal from the platform entirely.

Interviewees described partnerships between platform representatives and civil society organizations as "the key" to operationalizing crisis misinformation policies, explaining that external partners, particularly those with close knowledge of the local context, can help detect and debunk emerging narratives that pose a risk of harm in real time. Platforms can then choose to remove dangerous misinformation, deamplify it, or add warning labels or interstitials to better contextualize content for users. Interventions may also be deployed to prevent misinformation from being actively re-shared or recommended on the platform.

Interviewees emphasized that understanding and deamplifying crisis misinformation is a "very heavy lift," and doing this work at scale in conflict zones is "incredibly difficult." Understanding and addressing harmful misinformation is ideally done by experts, and is not easily simplified into work that can be done by frontline content moderators or through automated enforcement. This situation is further complicated by the fact that dangerous claims "age and metastasize," making them difficult to keep up with. As one interviewee noted: "the most dangerous narrative is the next narrative."

Platforms should ensure that they have policies in place prohibiting incitement to violence and that these policies are rigorously enforced. These policies should be developed and enforced with an understanding of the patterns surrounding the commission of mass violence, such as the dehumanization of individuals and groups, the use of hate speech, and the use of coded language or "dog-whistling" to incite violence.[103] Interviewees also recommended that these policies apply to all users across contexts, including political leaders, military leaders, and elected officials.

Deamplification

The ability to deamplify content on social media presents another potential opportunity to (indirectly) degrade perpetrator capacity. Interviewees suggested that inflammatory digital content that poses a serious risk of violence—yet which does not warrant removal—can sometimes be demoted on user feeds in instances where platforms have a principled framework to support those decisions. For example, platforms may proactively deamplify certain terms or phrases if they could create a risk of harm, such as dehumanizing language, or derogatory terms and slurs. This work can be supported by either sophisticated processes to recognize and classify content, or by simple “key word demotion.” In Burma for instance, interviewees recalled that platforms partnered with civil society organizations to co-design “slur lists” that could be used by social media platforms in various interventions.[104] Interviewees described demotion as a powerful tool because it affects the prioritization of all content, rather than only affecting individual pieces of content flagged for review by moderators.

Interviewees expressed serious concerns, however, with deamplification rather than removal of dangerous content. Interviewees highlighted the difficulty in designating categories of dangerous content for deamplification and noted the risks of platforms being perceived as adopting a particular political viewpoint when deploying this intervention. Deamplifying on the basis of particular terms, as opposed to highly contextualized analysis, may inadvertently suppress broad swathes of information or discussion relating to a conflict, inhibiting expression and access to information. One participant referenced how groups in Ethiopia perceived that the word “genocide” was being demoted, so they avoided using it in their advocacy, to what they saw as the detriment of the strength of their messaging. Interviewees also observed that content-based deamplification can be perceived as “shadow banning,” thereby undermining trust in social media companies.[105] Deamplification may be particularly damaging for groups seeking to use certain terms as counter-speech to raise awareness about human rights violations, and interviewees called on platforms to be transparent when considering its implementation.

Rate limits

Interviewees suggested that product design interventions may further reduce the presence or visibility of dangerous content on social media. For example, emerging scholarship has emphasized the importance of interventions that place reasonable limits on the number of accounts one user can contact at once, to “mirror real-life processes whereby individuals have to gain some level of trust to be able to reach broad groups of others.”[106] In the wake of mob violence in India seemingly fueled by rumors on WhatsApp, for example, the platform imposed a 20-person limit on the number of people to whom users could forward messages—after its prior threshold permitted users to share messages with up to 250 people at once.[107] Interviewees emphasized that these interventions continue to permit information-sharing during a crisis, but they draw on research suggesting that after messages have been re-shared repeatedly, the value of the information being shared drops.[108]

Notably, rate limits on messaging services such as WhatsApp are effectively “content neutral”—that is, imposed to prevent the mass forwarding of messages regardless of the content of those messages—and are also agnostic to the identity of the sender. Rate limits on traditional social media platforms could take the form of limits prohibiting users from sending direct messages *en masse*, or from sharing high volumes of invitations to join groups. Interviewees highlighted that rate limits could help restrain mass engagement beyond a certain threshold, or require users to achieve some level of trustworthiness on the platform before permitting them broad reach and engagement.

Limitations on ad use

Interventions that link “trustworthiness” with reach may also be used to limit perpetrators’ use of social media advertisements in atrocity risk settings. In Ukraine, for example, consultation interviewees referenced with concern how newly created social media accounts were able to use targeted advertisements to portray Ukrainian armed forces disparagingly, depicting them as criminal actors. Ensuring that social media users exhibit some indicia of trustworthiness before they can make use of social media may help reduce the presence of dangerous content in high-risk settings.

In other instances, platforms have suspended advertising services in conflict settings altogether, or have barred certain categories of users from running advertisements. Facebook, for example, prohibited advertisers within Russia from creating or running ads anywhere in the world during the conflict in Ukraine.[109] To the extent advertisements are permitted in atrocity risk settings, they should be rigorously scrutinized against policies prohibiting hate speech and incitement to violence.[110]

Core guidance and recommendations:

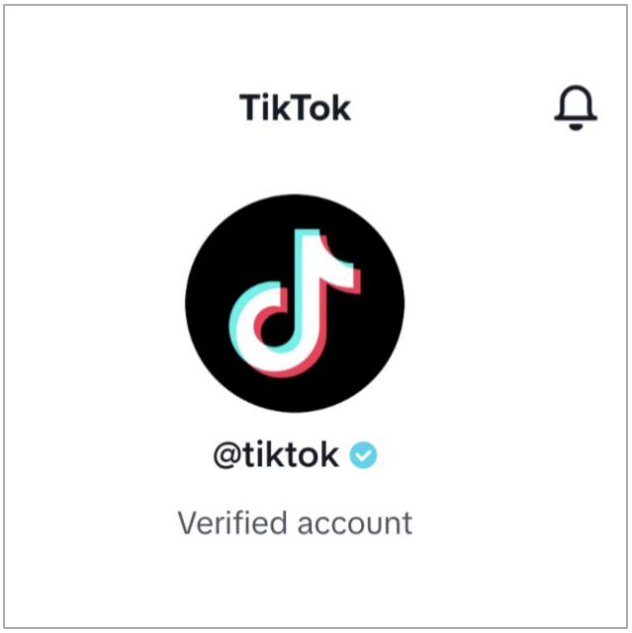
- Platforms should ensure they have policies in place to manage dangerous misinformation in atrocity risk settings, perhaps by limiting users' ability to share, recommend, or amplify unverified and potentially harmful information.
- Platforms should also ensure they have policies in place prohibiting the incitement of violence and that these policies are rigorously enforced in atrocity risk settings. These policies should also be developed and enforced with an understanding of behaviors and patterns concerning the commission of mass violence, such as the use of dehumanization, hate speech, and coded language or "dog-whistling" to incite violence.
- Where dangerous misinformation remains online in atrocity risk settings, platforms should explore the use of "soft interventions" to reduce the risk of misinformation contributing to violence, such as placing warning labels over the content.
- Platforms should engage in further research on the benefits, risks, and unintended consequences of deamplifying dangerous content (such as dehumanizing language or derogatory terms) in atrocity risk settings, but they should be transparent about their approach.
- In atrocity risk settings, platforms should explore reasonable rate limits or requirements that users accumulate some indicia of trustworthiness before they are permitted broad reach and engagement on the platform, to prevent perpetrators from reaching other users *en masse*.
- Platforms should explore opportunities to link indicia of "trustworthiness" to the ability to use features like ads in atrocity risk settings, or to prohibit the use of ads outright in certain contexts. To the extent ads are permitted, they should be rigorously scrutinized against policies prohibiting hate speech and incitement to violence.

D. Interventions Aimed at Contextualizing Perpetrator Content

Interventions Aimed at Contextualizing Perpetrator Content	
Description	Interventions aimed at providing additional information or context about inflammatory digital content, where it is not removed outright
Theory of Change	Situating inflammatory digital content in the context of credible, factual information can reduce the ability of potential perpetrators to spread and persuade people of dangerous rumors or to incite violence.
Examples	● Placing warning labels or interstitials over potentially inflammatory digital content, sharing further context about what is depicted or asserted ● Verifying and labeling accounts belonging to certain types of users, such as government officials, electoral candidates, or state-affiliated media ● Providing further context on or labeling the provenance of misleading media ● Prebunking or inoculating users against dangerous misinformation

This next category of interventions aims to provide additional information or context around inflammatory content posted on social media. The theory of change underlying this category is that situating inflammatory narratives, images, or videos on social media in the context of credible information can help dispel dangerous rumors and prevent them from contributing to the incitement of mass atrocities. These interventions may degrade perpetrator capacity by short-circuiting cycles of misinformation and disinformation, through techniques that counter or debunk dangerous narratives and rumors, and by enhancing people's understanding of the digital content they encounter.

Account labeling and verification



TikTok offers a verified badge for accounts where the platform has confirmed the user is who they purport to be, helping users make informed choices about the accounts they engage with. *TikTok*

One intervention to contextualize digital content is the use of labeling, which typically refers to visual tags affixed to accounts or posts on social media, designating or categorizing them so users can better understand the information they encounter. For example, many platforms have labeled accounts belonging to government officials, state-affiliated media, and electoral candidates to help ensure that people understand who is advancing a particular issue.[111] When the authentic accounts of leaders and influential figures are clearly denoted, this intervention can also help prevent users from being misled by impersonation efforts that may seek to misrepresent such persons or their views.[112] Relatedly, this has also been a motivation behind user verification, which consists of "checkmarks" or other indicators that allow users to more readily identify the authentic accounts of public figures, brands, and institutions. Interviewees observed that labeling and verification are important not only to prevent people from being misled, but also to support users' understanding of which accounts to trust.[113] Interviewees emphasized the importance, in an atrocity risk setting, of being able to identify government accounts that may post evacuation routes or locations for aid distribution—as well as being able to distinguish them from those of impersonators.

Labeling may be paired with secondary interventions to demote or deamplify content being shared by certain types of accounts—often referred to as "visibility filtering." Interviewees noted that, taken together, these interventions may help simultaneously limit both the impact and the dissemination of perpetrator content. In the context of the war on Ukraine, for example, Twitter prohibited accounts belonging to Russian state-affiliated media from being either recommended or amplified across the platform, and according to the company's internal research, this action decreased the reach of these outlets by approximately 30 percent.[114] Interviewees contrasted the labeling and deamplification of state-affiliated media with bans on state-affiliated media, as the European Union required for Russian state-affiliated media during the war in Ukraine. By labeling and deamplifying (but not removing) content from these accounts, interviewees noted that it remains available on social media platforms to support war crimes investigations and accountability efforts.

Interviewees referenced difficulties in scaling labeling efforts, particularly when labeling individual categories of users. Verifying and labeling every single electoral candidate (in local, regional, and national elections around the world), for example, was described as an "incredibly fragile and labor-intensive process," which would also need to be regularly updated.[115]

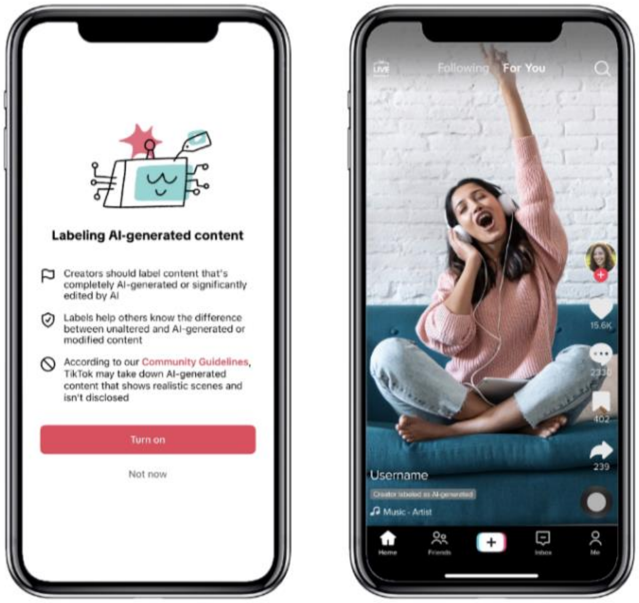
Content labeling or interstitials

One intervention to contextualize digital content is the use of labeling, which typically refers to visual tags affixed to accounts or posts on social media, designating or categorizing them so users can better understand the information they encounter. Information on social media may also be usefully contextualized by adding interstitials or warning labels to misleading content. Rather than removing dangerous narratives or rumors outright, platforms have deployed interstitials in certain circumstances that encourage users to beware or think twice when engaging with that content. Interviewees noted that, in some cases, these interstitials are paired with interventions to reduce visibility of the underlying content, such as removing the ability to re-share the post. Interviewees recalled platforms using interstitials successfully in the context of election misinformation, noting that the labels can offer users greater context and information about the content they encounter. In addition, interstitials can (but do not always) refer users directly to third-party resources or information on external pages, an additional element that interviewees noted required more resources but that may offer value in an atrocity risk setting. Interstitials have also been applied to images that are authentic but that are circulating without context in ways that are misleading, such as in relation to a different conflict or period of time; these interstitials inform users that "this looks like it circulated five years ago" or similar language.

> ***"The manipulation of people into becoming perpetrators is, generally speaking, built upon a foundation of lies. The more we can address these misunderstandings . . . the more we can reduce the likelihood or impact of atrocities."***
>
> - Expert on technology and mass atrocities

Interviewees noted that it is difficult to deploy interstitials at scale but said it may be feasible for a small subset of content, such as when misleading content is being shared by a public official or figure. Some interviewees cautioned, however, that in some past instances, people rarely seemed to actually engage with or click-through the interstitials to view information provided. Interviewees also questioned whether adding context to dangerous content may inadvertently draw more attention to it, raising the visibility and risk associated with those posts. Indeed, while some studies indicate that content labeling offers significant potential, reducing users' likelihood to believe and re-share misinformation, other studies suggest it can have unintended consequences, such as increasing the likelihood of users re-sharing unlabeled false content, assuming that it is true.[116]

Interviewees also suggested labeling the provenance of misleading media as another tool to degrade perpetrator capacity. These labels would aim to explicate powerful images or media that are not what they purport or appear to be and to help users understand the post's true origin or context, degrading perpetrators' ability to weaponize those images to build support for violence. Interviewees who supported this intervention spoke to the power of images on social media, especially where images evoke strong emotions and may exacerbate animosity against individuals and groups, thereby contributing to an environment where atrocities are believed to be "permissible." This intervention may also be useful to blunt the impact of deepfakes or other types of synthetic and manipulated media. TikTok, for example, has developed labels for AI-generated content to help users more readily identify manipulated or manufactured content.[117]



Labeling AI-generated content

Creators should label content that's completely AI-generated or significantly edited by AI

Labels help others know the difference between unaltered and AI-generated or modified content

According to our Community Guidelines, TikTok may take down AI-generated content that shows realistic scenes and isn't disclosed

TikTok announces efforts to label AI-generated content, aimed at helping users contextualize content they encounter on the platform. *TikTok*

Prebunking

Other interventions in this category aim to degrade perpetrators' ability to persuade users to believe misinformation and disinformation. These may include "prebunking" or "inoculating" against dangerous narratives, an intervention which draws on evidence that providing people with "microdoses" of misinformation, while refuting them with authoritative sources, can prevent users from being swayed when they encounter similar misinformation and disinformation in the future.[118]

Prebunking can take the form of either reminding people of the basic tenets of engaging in critical analysis of information, or supporting people in critically assessing specific claims. Some interviewees preferred the former approach, believing it to be more transferable across contexts, given that narratives unfolding on social media are “moving targets.” As one interviewee put it, people want to be “taught how to think, not forced to think in a certain way.” Interviewees felt that prebunking strategies can support users in engaging in the process of critical thought, as well as teach people how to spot weaknesses in arguments but invite them to draw their own conclusions.

Interviewees described successful prior efforts to inoculate against misinformation in instances in which platforms had the ability to anticipate crisis events, such as elections. In the context of elections, platforms can prepare messaging to inform people about key tactics or narratives that will likely be used to manipulate them. As one example, in the lead-up to the 2024 European Union election, Google and partner organizations developed social media ads teaching users how to spot common misinformation tactics, such as scapegoating of marginalized groups.[119] By contrast, interviewees described the challenges of inoculating users from misinformation when unfolding events occur suddenly or are difficult to predict. When asked how inoculation campaigns might be optimized for use in atrocity risk settings, interviewees discussed the need to identify types of speech that carry the most potential to drive harm, as well as types of speech that may persuasively draw upon preexisting grievances in a particular context.

KRYZYS W POLSCE
31/08/2022
Wojna trwa, uchodźcy zostaną w Polsce na wiele miesięcy
Wanessa Franczak
Wczoraj, 10:05
Miliony uchodźców przekroczyły polską granicę od początku marca...
To win this war, some will try to manipulate you

A still from a prebunking video that aimed to counter emerging false narratives about Ukrainian migrants on social media. Google's Jigsaw partnered with local organizations in Central and Eastern Europe to produce the prebunking videos, which were viewed over 38 million times, reaching social media users in Poland, Czechia and Slovakia. *Jigsaw*

Others noted that prebunking initiatives are resource-intensive and difficult to scale, referencing the amount of research and tailoring involved. It also remains unclear how often people need to be “inoculated” against misinformation and disinformation, or how often these efforts should be deployed over a protracted period.[120] Interviewees also expressed concern that prebunking could foster distrust on social media platforms generally, which may itself be a perpetrator strategy, such as Russian efforts to erode trust in mainstream information during the war in Ukraine.[121] In general, interviewees called for greater transparency concerning the effectiveness of prebunking initiatives, and ensuring that researchers have access to data so that resources can be directed effectively.

An essential precursor to addressing misinformation and disinformation is understanding local context, including by engaging with local partners. Where contextual knowledge is absent, users may be able to themselves flag or correct misleading information they encounter, such as through X's Community Notes feature, or by developing improved channels for people to identify misinformation from the bottom up. These community-to-company interventions would leverage the awareness of people on the ground in atrocity risk settings rather than relying on platforms' understanding of local context. Interviewees also noted that offering users an opportunity to provide input on the accuracy of posts subtly encourages them to think through content and their reactions to it. Others, however, were skeptical of these interventions, noting that they may be more effective in settings where there are broad distributions of political beliefs and less helpful in authoritarian settings where the state has successfully manipulated perceptions on a given issue. Interviewees also pointed to issues that have plagued past interventions, such as X's Community Notes feature, including vulnerability to manipulation and lack of oversight.[122]

Core guidance and recommendations:

- In atrocity risk settings, platforms should explore labeling and verifying certain categories of accounts, such as those belonging to government officials or electoral candidates, to help ensure users are not persuaded by impersonation attempts.
- Platforms should explore the use of interstitials, paired with deamplification, for a small subset of high-risk, high-visibility content in atrocity risk contexts. They should also explore options to communicate the provenance of misleading media, so users better understand the source of content they encounter.
- In partnership with local organizations, platforms should explore the use of prebunking to reduce the potency of misinformation and disinformation in atrocity risk contexts and support independent research on the efficacy of these efforts.
- Where local partnerships are absent, platforms should explore the possibility of user-led or community-to-community interventions that would enable users to flag misinformation themselves.

E. Interventions Aimed at Preventing Perpetrators from Mobilizing Bystanders

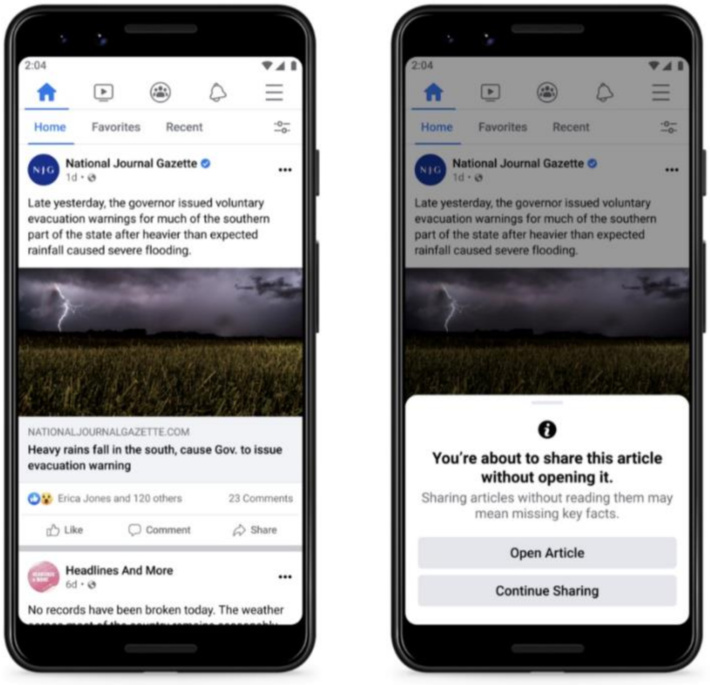
Interventions to Prevent Perpetrators from Mobilizing Bystanders	
Description	Interventions aimed at reducing the incentives for bystanders or third-party enablers to inadvertently contribute to narratives and ideologies being advanced by perpetrators
Theory of Change	Reducing the likelihood that third-party enablers contribute to the dissemination of dangerous narratives and ideologies advanced by perpetrators can reduce perpetrators' ability to weaponize social media to incite or fuel atrocities.
Examples	● "Nudges" suggesting that users "think twice" before re-sharing certain content on social media ● Prompts warning users if they are about to share a potentially harmful or hurtful reply or comment ● Interventions to interrupt the user interface to make it more difficult to rapidly re-share content that may contribute to violence

The concept of "bystanders" to atrocities is often used to describe those who are passive or indifferent to episodes of mass violence.[123] In the wake of the Holocaust, it was used to reference a range of domestic and international actors who were neither perpetrators nor victims, but who failed to speak out or take action to protect those being persecuted.[124] In some instances, those who claimed to be bystanders were persuaded to become active participants in the Holocaust, out of economic or personal interest, prejudice, or pressure within their communities.

Today, the idea of bystanders or third-party enablers encapsulates the array of actors that contribute to mass atrocities—by supplying resources, services, or political or moral support for perpetrators. The US Agency for International Development (USAID) has described these enablers as local, national, regional, and international actors that, even if motivated by their own power or economic interests rather than the same interests as perpetrators, nevertheless "facilitate a supply chain that fuels violence against civilians."[125]

In the digital environment, actors who are otherwise neutral may be incentivized to contribute to or re-share dangerous narratives being advanced by perpetrators. This category of interventions looks to reduce the incentives for bystanders—who may not otherwise be interested in contributing to mass violence—to support perpetrators' use of social media in an effort to garner attention, money, or power. The theory of change is that by reducing the likelihood that bystanders contribute to the dissemination of dangerous narratives being advanced by perpetrators, perpetrators may have less ability to weaponize social media to incite or fuel violence.

One of the interventions proposed in this category was the use of "nudges" to encourage users in atrocity risk settings to pause in some way before responding to or re-sharing a post with potentially dangerous or misleading information. This intervention is premised in behavioral science theory, which aims to influence the likelihood that individuals choose to take a certain action over others.[126] Participants discussed the potential of nudges to encourage critical thinking and reflection when engaging online, noting that they could take the form of encouraging users to read an article before re-sharing it, or to think twice. In 2021, for example, Twitter launched a prompt that asked users whether they would like to open an article before re-tweeting it; users had the option to click through the prompt and continue to share the article or to go to the article instead.[127] Both Facebook and Twitter have also used prompts to warn users if they are about to share a potentially harmful or hurtful reply.[128] In an alternative version of this intervention, interviewees proposed not prompting users to pause, but rather interrupting the user interface in moments of risk to make it more difficult for someone to rapidly re-share information.



National Journal Gazette
1d ·

Late yesterday, the governor issued voluntary evacuation warnings for much of the southern part of the state after heavier than expected rainfall caused severe flooding.

NATIONALJOURNALGAZETTE.COM
Heavy rains fall in the south, cause Gov. to issue evacuation warning

Erica Jones and 120 others
23 Comments

Headlines And More
6d ·
No records have been broken today. The weather

National Journal Gazette
1d ·

Late yesterday, the governor issued voluntary evacuation warnings for much of the southern part of the state after heavier than expected rainfall caused severe flooding.

You're about to share this article without opening it.
Sharing articles without reading them may mean missing key facts.

Facebook explains its launch of a 'nudge' feature that prompts users to open and read news articles before sharing them with others. *Meta Newsroom*

Interviewees emphasized, however, the different considerations at play in settings where atrocities have not yet begun as opposed to places where people are already at risk of direct violence. These "speed bumps" to sharing information on social media can be valuable in some settings, but people under direct threat may need to share and receive information quickly. Interviewees were also concerned about how social media users would view these types of nudges from social media companies, who are not always perceived to be acting in users' best interests.

In other instances, bystanders may be financially motivated to spread disinformation or to impersonate members of affected communities. According to one interviewee, the Taliban used monetary incentives to persuade people to create pages impersonating Hazara and Uzbek communities. The content shared on those pages purported to subscribe to Taliban ideology, creating a perception of normalization and acceptance that can be highly damaging. Initiatives to verify the accounts of public figures and help prevent impersonation (both discussed previously) may offer potential to address bystander mobilization in these contexts.

Interviewees also pointed to emerging research exploring the use of social media data in targeted messaging campaigns that aim to disrupt potential perpetrators' motivation to commit atrocities. Dr. Rhiannon Neilsen has proposed the use of targeted educational campaigns for individuals in atrocity risk contexts, advocating restraint and peace "in the same way that individuals receive personalized advertisements for products on social media.”[129] While this intervention remains as yet untested, interviewees referenced its potential to dismantle the incentives and the broader logic structure in which potential perpetrators are operating.

Core guidance and recommendations:

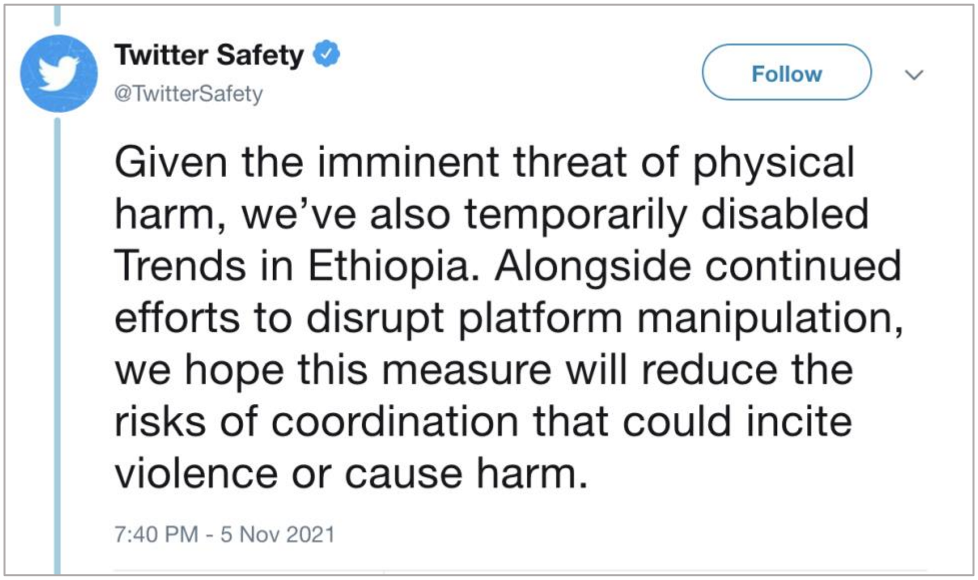
- Platforms should, in atrocity risk settings, explore the use of nudges to encourage critical thinking and to make it more difficult for bystanders to rapidly re-share information that could contribute to violence.
- In settings where atrocities have already begun, platforms may want to consider suspending interventions that add friction to users’ ability to rapidly share content that may be necessary for their protection.

F. Last Resort or “Break Glass” Measures

Last Resort or “Break Glass” Measures	
Description	Interventions that temporarily and intentionally disable or degrade social media features in moments of heightened atrocity risk
Theory of Change	Where social media features are at risk of being abused to contribute to atrocities, disabling features reduces tools available to perpetrators.
Examples	● Intentionally disabling features that allow users to share hashtags, to avoid inciting violence in Ethiopia ● Intentionally slowing down or degrading the functionality of certain features (i.e., adding “friction”) to prevent content from rapidly circulating on social media

In some instances, certain social media features or products may present such a grave risk of abuse or weaponization by perpetrators that they may be worth disabling until the situation stabilizes. This final category of interventions covers such interventions, those that would temporarily and intentionally disable or degrade social media features in moments of heightened atrocity risk. The theory of change is that, where social media features are at grave risk of being abused to contribute to atrocities, “turning off” those features (or deliberately making them less effective) meaningfully reduces the tools available to perpetrators.

Amid surging ethnic violence in Ethiopia in 2021, for example, Twitter temporarily disabled its “trends” feature in the country. In a public statement on the decision, Twitter said that its intervention was aimed at reducing the “risks of coordination that could incite violence or cause harm.”[130] Interviewees also referenced past instances when friction has been added in various ways to slow down content on platforms, which was described as a blunt but occasionally necessary tool. Others referenced prior efforts at social media companies to develop contingency plans to disable or degrade certain features in the event of large-scale violence, but noted the difficulties inherent in persuading a company to intentionally “degrade” the user experience. Interviewees emphasized the importance of platforms having written protocols in place that systematically address how and when to determine the implementation of these interventions, ideally well in advance of mass violence.


Twitter Safety
@TwitterSafety

Given the imminent threat of physical harm, we've also temporarily disabled Trends in Ethiopia. Alongside continued efforts to disrupt platform manipulation, we hope this measure will reduce the risks of coordination that could incite violence or cause harm.

7:40 PM - 5 Nov 2021

In November 2021, Twitter decided to temporarily disable its 'Trends' feature in Ethiopia. It announced the feature in a tweet, stating that the decision was aimed at reducing the risk that the feature could be used to incite violence. *Twitter, https://twitter.com/TwitterSafety/status/1456813764184055808, accessed via Wayback Machine*

Although instances of platforms intentionally degrading their own features are rare, interviewees felt these interventions should be kept on the table, describing them as among the "biggest emergency tools in the toolkit." While interviewees largely preferred a more "nuanced approach" to mitigating risk associated with platform features, many felt that disabling entire features in a given setting may be a necessary last resort, particularly if platforms become aware that they lack the resources or tools in place to mitigate grave risks of abuse. In the words of one interviewee, if a feature is being deployed in a way that drives a dimension of the conflict, platforms turning that feature off may be "relevant and justified."

At the same time, interviewees emphasized that these drastic interventions come with difficult trade-offs. Given the importance of social media to information sharing, interviewees were hesitant to disable features or products entirely, particularly in crisis settings. They expressed concern that these interventions might end up hurting victims more than perpetrators, but noted that platforms are used in different ways, making this a unique calculation for each company. Some interviewees felt products or features might be more usefully disabled for specific accounts rather than for entire countries, yet recognized that action at the account level may come with additional liability or exposure for social media companies. These interviewees also noted that even disabling product features in a certain country may have its limitations, considering the continued ability of diaspora communities or other users outside that country to use features limited in-country. Both risk assessments and interventions, they noted, must carefully account for where dangerous content is actually originating before deploying these "break glass" measures.

Core guidance and recommendations:

- Considering the gravity and irremediability of mass atrocities, platforms should keep on the table interventions that would temporarily degrade or disable platform features at risk of severe abuse by atrocity perpetrators.

- At the same time, because of the dual-use nature of most social media features, these measures should typically be used as a “last resort” or “break glass” measure, deployed only after assessing relevant limitations and trade-offs.

VII. CONCLUDING THOUGHTS AND RECOMMENDATIONS

Digital spaces today are at the front lines of where atrocity risks materialize and unfold. They are among the arenas where the classic phases of mass violence play out anew—dehumanization, the rise of exclusionary ideologies, the exposure of individuals as enemies or traitors, and the organization and coordination of genocide. But they are also places where atrocity prevention strategies—such as protecting vulnerable civilian populations and degrading perpetrators—have opportunities to play out, by expanding the atrocity prevention toolbox to leverage digital tools and interventions. In some cases, this entails adapting traditional tools to the realm of social media, while in other cases, it requires developing an expanded awareness of new tools with the potential to support prevention strategies.

At the same time, a note of caution is warranted for anyone assessing the use of social media to support atrocity prevention efforts. Interviewees consulted for this report referenced the hard-learned lessons of many in the humanitarian aid and development community who have come to understand that, just by being in the space, they have effectively become part of the conflict, and their interventions may be instrumentalized in complicated ways. While the social media environment offers unique tools to support atrocity prevention efforts, interviewees urged caution, thorough assessment, and data-driven research before deploying new interventions that may carry unintended consequences.

Interviewees also urged social media companies to review lessons learned from their use of prior interventions, so that these tools can be iterated and improved for future use. This may include collecting data on the efficacy of various tools and reviewing both interventions and policies surrounding their use with external stakeholders such as civil society organizations, so that risks and concerns can be anticipated and addressed.

In surveying the tools available to use when confronted with moments of atrocity risk, the atrocity prevention community must not forget the digital realm —because its implications reverberate for communities that are both online and offline. By expanding the atrocity prevention toolbox to include digital tools and interventions, we have an opportunity to become more active participants in digital spaces, and to develop more modern atrocity prevention strategies to meet the challenges of the moment.

A. Recommendations for Platforms

As discussed, this report sets out the landscape of social media tools and interventions that may be able to support either (a) protecting vulnerable civilian populations or (b) degrading perpetrator capacity. Because many of these interventions are within the control of platforms, most of the resulting recommendations are directed at social media companies.

1. Preliminary Recommendations: Interventions to Support Civilian Protection

First, preliminary recommendations on specific tools and interventions that may be able to contribute to the protection of vulnerable civilian populations are as follows:

Atrocity Prevention Strategy: Protect Vulnerable Civilian Populations

Key Assumptions

- Social media can enable vulnerable civilian populations to access critical information and coordinate actions to protect themselves in moments of crisis.
- At the same time, information available on social media can place civilians at greater risk of physical attack.
- Social media can enable communication between members of affected communities about emerging atrocity risks, and from affected communities to policy makers.

Mechanisms

- Safeguarding sensitive information about vulnerable civilian populations (for example, by *locking profiles or increasing account security measures to prevent hacking*)
- Coordinating and facilitating self- or external-protection efforts (for example, by *users communicating warnings on unsafe locations or circulating information on humanitarian aid access points*)
- Supporting access to essential information that could be used for protection

TOOLS & INTERVENTIONS TO PROTECT VULNERABLE CIVILIAN POPULATIONS

Protect Online Privacy	**THEORY OF CHANGE:**	**EXAMPLES:**	**PRELIMINARY GUIDANCE FOR PLATFORMS:**
Tools or interventions aimed at restricting the visibility of digital content that may put civilians at risk in atrocity risk settings	If digital content could be used to target civilians, restricting the visibility of that content can contribute to civilian protection.	• Facebook's locked profile feature, which limits the ability to view various elements of a person's social media account, or similar interventions to limit the ability to view a user's affiliations or friends lists • Obscuring users' previously shared location information • Reviewing features to which users may be added without their consent that could make them more readily visible to perpetrators • Creating channels for users' social media accounts to be secured or locked down in case of detention or arrest • Proactively sharing instructions on the deletion or deactivation of social media accounts	• Platforms should explore interventions to proactively restrict the visibility of digital information that could be used to target civilians in atrocity risk settings, such as their affiliations or location history. • Privacy interventions aimed at protecting civilians should be carefully balanced against their potential interests in sharing information in atrocity risk settings. Wherever feasible, civilians should be afforded agency over their digital presence. • Platforms should carefully review features through which civilians' digital information may be visible without their consent, or where they may not realize they gave prior consent. • Platforms should ensure that vulnerable civilian populations can readily understand how to temporarily deactivate or delete their social media accounts should they deem it necessary for their protection. • Platforms should communicate available privacy tools to vulnerable populations in advance of crises, and should clearly articulate relevant limitations to avoid overpromising to people who are at risk.

Secure Social Media Accounts *Interventions aimed at protecting social media users against hacking, impersonation, and account takeover efforts*	**THEORY OF CHANGE:** Civilian protection includes ensuring that civilians' digital information cannot be obtained and used against them through hacking and impersonation campaigns. This can in turn protect others who may be misled by hacked and impersonated accounts.	**EXAMPLES:** • Account security push notifications, deployed in Ukraine • End-to-end encryption channels	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • Platforms should ensure they put in place and stringently enforce policies prohibiting impersonation in atrocity risk settings. • Platforms should explore opportunities, such as through push notifications or prompts, to proactively communicate information to civilians about how to best secure their online accounts. • Platforms should clearly communicate their choices around the use of encrypted or unencrypted features, so that users readily understand the security of the tools they use in atrocity risk settings.
Surface Crisis Resources and Credible Information *Tools or interventions aimed at connecting social media users to crisis resources, amplifying credible information, or both*	**THEORY OF CHANGE:** Ensuring that civilians can access information about crisis resources can contribute to protection by helping them avoid or withstand attacks. OR Ensuring that civilians can access reliable information about evolving developments can prevent misinformation and disinformation from inciting violence.	**EXAMPLES:** • Creating centralized landing pages or information hubs that compile authoritative information in atrocity risk settings • Modifying approaches to ranking and amplification of information to align with needs in atrocity risk settings • Amplifying content from credible accounts, such as reliable media or civil society organizations • Providing ad credits to credible local organizations	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • In atrocity risk settings, the way information is ranked and prioritized takes on heightened importance. Platforms should review approaches to the ranking and amplification of information to align with needs in atrocity risk settings. • Platforms should develop principled approaches for how they will identify credible and useful information for civilian protection, and under what circumstances specific information will be amplified. • Platforms should consider affording users choice in how content is prioritized in user feeds so they can quickly identify resources and information important to their protection.

		• Using push or pop-up notifications, or “nudges,” to direct people to important resources or news items	• Platforms should, in partnership with relevant organizations and humanitarian agencies, explore opportunities to direct users to credible information or resources that could support their protection. • Platforms should explore opportunities to afford vulnerable communities greater control and agency in efforts to surface crisis resources on social media and mitigate risks associated with information sharing.
Disseminate Early-Warning Information *Interventions that make use of social media to communicate warnings about atrocity risks*	**THEORY OF CHANGE:** Social media may be used to communicate warnings (either to civilians at risk or to policy makers), with a view to influencing outcomes on civilian protection.	**EXAMPLES:** • Publishing emergency air raid alerts on social media • Using social media posts to warn people about safe/unsafe locations in Libya, or potential targets for air strikes in Syria	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • In light of the use of social media for early warning, platforms operating in atrocity risk settings should review the way that content moderation policies on graphic media or violent content are applied and enforced, with consideration to the needs of affected populations to understand emerging events and risks of violence. Platforms should also explore the use of technical interventions to mitigate psychosocial harm associated with the viewing of graphic content, such as the use of interstitials, grayscale, or image blurring. • Platforms should explore opportunities to support early-warning initiatives by trusted third-party entities, but they also should implement safeguards to carefully assess information credibility and timeliness.

Enhance Communication and Coordination Capabilities *Interventions that expand or enhance civilians' ability to communicate and coordinate*	**THEORY OF CHANGE:** Supporting open communication and coordination between civilians will enable them to better avoid or withstand atrocities.	**EXAMPLES:** • "Groups" or "Communities" features on social media • Social media features that enable group messaging • Features that help users connect to social media platforms via proxy servers, bypassing restrictions on internet access	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • Platforms should be mindful of the value of features that enable group discussion for coordination and communication between civilians in atrocity risk settings. When considering modifying or updating these features, platforms should take particular care in assessing the needs of those in atrocity risk settings who may be using them for this purpose. • Platforms should pay particular attention to the moderation of group discussion forums in settings where there is a heightened risk of mass atrocities. This may include ensuring that the administrators of discussion forums have the tools and resources they need to support moderation, such as the ability to approve or remove members from digital spaces. • Platforms should ensure that the visibility or privacy of group discussion forums is clearly communicated to all participants. • In atrocity risk settings where restrictions on communication are heightened, such as through internet blackouts or platform shutdowns, platforms should consider exploring opportunities to expand access to social media, particularly for vulnerable or isolated communities.

2. Preliminary Recommendations: Interventions to Degrade Perpetrator Capacity

Preliminary recommendations on specific tools and interventions that may be able to contribute to degrading the capacity of atrocity perpetrators are as follows:

Atrocity Prevention Strategy: Degrade Potential Perpetrators’ Capacity to Commit Atrocities

Key Assumptions

- Mass atrocities depend on perpetrators having certain material and operational capacities. In many countries at risk of mass atrocities today, perpetrators may use social media as a resource for facilitating systematic attacks.
- Social media can enable potential perpetrators to communicate rapidly and persuasively with large audiences in ways that may contribute to atrocity risk, by inciting violence, spreading exclusionary ideologies, or disseminating disinformation or misinformation about a particular group.
- Social media can also play a role in the planning and organization of attacks, such as by providing forums for recruitment or weapons sales.
- Tools that make social media platforms less effective or efficient means of advancing perpetrators’ goals can therefore contribute to degrading their overall capacity to commit atrocities.

Mechanisms

- Decreasing the speed and audience-reach efficiency of social media features for potential perpetrators (for example, via *content moderation policies on crisis misinformation, rate limits, or nudges suggesting users think twice before re-sharing certain content*)
- Decreasing the persuasiveness of inciting, misleading, or otherwise dangerous content (for example, via contextualizing content or labeling the source of posts, such as state-affiliated media)
- Disrupting digital spaces in which perpetrators are organizing or planning the commission of atrocities (*for example, weapons sales and recruitment*)
- Denying potential perpetrators access to social media platforms entirely or to specific social media features or platforms (for example, via *detection and removal of coordinated networks of accounts of potential perpetrators, deplatforming violent organizations, or disabling social media features in moments of heightened atrocity risk*)

TOOLS & INTERVENTIONS TO DEGRADE POTENTIAL PERPETRATORS' CAPACITY TO COMMIT ATROCITIES

	THEORY OF CHANGE:	EXAMPLES:	PRELIMINARY GUIDANCE FOR PLATFORMS:
Prevent Perpetrators Gaining Foothold on Platforms at Scale *Interventions aimed at preventing perpetrators from setting up a large presence on social media platforms*	Preventing perpetrators from establishing or maintaining extensive networks of accounts will make them less able to weaponize social media in furtherance of atrocities (such as to incite or coordinate violence).	• Preventing perpetrators from registering social media accounts • Expanding detection of coordinated networks of accounts of potential perpetrators • Designating and banning perpetrators under policies governing violent individuals and organizations • Deplatforming perpetrators, or subjecting them to heightened monitoring against content moderation policies	• Platforms should explore interventions that can prevent atrocity perpetrators from setting up networks of inauthentic accounts. This should include both efforts to prevent the registration of new accounts, and review of older accounts that may exhibit suspicious behavior. • Platforms should enhance their in-house investigative capacities to detect and remove coordinated networks of inauthentic accounts that may be used by atrocity perpetrators. • Platforms should, in atrocity risk settings, proactively review potential perpetrators against criteria for designation under violent organizations policies. • Platforms should explore heightened monitoring of accounts of atrocity perpetrators and more stringent enforcement of content moderation policies given these individuals' offline behavior.

Disrupt Perpetrators from Coordinating and Organizing on Social Media	**THEORY OF CHANGE:**	**EXAMPLES:**	**PRELIMINARY GUIDANCE FOR PLATFORMS:**
Interventions aimed at disrupting perpetrators from using social media to coordinate and organize the commission of violence	To the extent that digital spaces are being used to coordinate and organize violence, disrupting perpetrators' ability to use social media to advance the planning and organization of violence will degrade their overall capacity to commit atrocities.	• Enforcement of content moderation policies prohibiting weapons sales or to promote criminal activities • Heightened monitoring of online spaces where perpetrators may be organizing violent activities, such as groups or pages • Policies prohibiting the use of social media for surveillance	• Platforms should ensure that they have policies in place that prohibit the abuse of their platforms for the coordination and organization of mass violence, including but not limited to the purchase and sale of weapons and recruitment to violent organizations. • Platforms should also ensure that the enforcement of these policies is sufficiently resourced in atrocity risk settings, particularly in online spaces where perpetrators may be gathering. • Platforms should explore interventions to prevent perpetrators from readily collecting information on social media about potential targets, and ensure policies are in place prohibiting the use of social media data for surveillance.

	THEORY OF CHANGE:	EXAMPLES:	PRELIMINARY GUIDANCE FOR PLATFORMS:
Limit the Presence or Visibility of Dangerous Content in Atrocity Risk Settings *Interventions aimed at reducing the presence or visibility of potentially inflammatory digital content during periods of heightened atrocity risk*	Reducing the presence, audience reach, or visibility of potentially inflammatory digital content, limits potential perpetrators' ability to use social media to incite atrocities or further societal divisions.	• Having policies governing how platforms manage dangerous misinformation in crisis settings, such as limiting it from appearing on users' home feed or timeline or limiting its ability to be re-shared • Deamplifying content that could create a serious risk of harm, such as potentially dehumanizing language or exclusionary ideologies • Implementing rate limits or forwarding limits that reduce the number of people a user can forward content to at scale	• Platforms should ensure they have policies in place to manage dangerous misinformation in atrocity risk settings, perhaps by limiting users' ability to share, recommend, or amplify unverified and potentially harmful information. • Platforms should also ensure that they have policies in place prohibiting the incitement of violence and that these policies are rigorously enforced in atrocity risk settings. These policies should also be developed and enforced with an understanding of behaviors and patterns around the commission of mass violence, such as the use of dehumanization, hate speech, and coded language or "dog-whistling" to incite violence. • Where dangerous misinformation remains online in atrocity risk settings, platforms should explore the use of "soft interventions" to reduce the risk of misinformation contributing to violence, such as placing warning labels over the content. • Platforms should engage in further research on the benefits, risks, and unintended consequences of deamplifying dangerous content (such as dehumanizing language or derogatory terms) in atrocity risk settings, but they should be transparent about their approach. • In atrocity risk settings, platforms should explore reasonable rate limits or

			requirements that users accumulate some indicia of trustworthiness before they are permitted broad reach and engagement on the platform, to prevent perpetrators from reaching other users en masse. • Platforms should explore opportunities to link indicia of trustworthiness to the ability to use features like ads in atrocity risk settings, or to prohibit the use of ads outright in certain contexts. To the extent ads are permitted, they should be rigorously scrutinized against policies prohibiting hate speech and incitement to violence.
Contextualize Perpetrator Content *Interventions aimed at providing additional information or context around inflammatory digital content posted on social media by potential perpetrators, where it is not removed outright*	**THEORY OF CHANGE:** Situating inflammatory digital content in the context of credible, factual information can reduce perpetrators' ability to spread and persuade people of dangerous rumors or incite violence.	**EXAMPLES:** • Placing warning labels or interstitials over potentially inflammatory digital content, sharing further context about what is depicted or asserted • Verifying and labeling accounts belonging to certain types of users, such as government officials, electoral candidates, or state-affiliated media • Providing further context on or labeling the provenance of misleading media • "Prebunking" or inoculating users against dangerous misinformation	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • In atrocity risk settings, platforms should explore labeling and verifying certain categories of accounts, such as those belonging to government officials or electoral candidates, to prevent users from being persuaded by impersonation attempts. • Platforms should explore the use of interstitials, paired with deamplification, for a small subset of high-risk, high-visibility content in atrocity risk contexts. They should also explore options to communicate the provenance of misleading media, so users better understand the source of content they encounter. • In partnership with local organizations, platforms should explore the use of prebunking to reduce the potency of mis/disinformation in atrocity risk contexts,

			and support independent research on the efficacy of these efforts. • Where local partnerships are absent, platforms should explore the possibility of user-led or community-to-community interventions that would enable users to flag misinformation themselves.
Prevent Perpetrators from Mobilizing Bystanders *Interventions aimed at reducing the incentives for bystanders or third-party enablers to inadvertently contribute to narratives and ideologies being advanced by perpetrators*	**THEORY OF CHANGE:** By reducing the likelihood that third-party enablers contribute to the dissemination of dangerous narratives and ideologies advanced by perpetrators, reduce perpetrators' ability to weaponize social media to incite or fuel atrocities.	**EXAMPLES:** • "Nudges" suggesting users think twice before re-sharing certain content on social media • Prompts warning users if they are about to share a potentially harmful or hurtful reply or comment • Interventions to interrupt the user interface to make it more difficult to rapidly re-share content that may contribute to violence	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • Platforms should, in atrocity risk settings, explore the use of "nudges" to encourage critical thinking, and they should make it more difficult for bystanders to rapidly re-share information that could contribute to violence. • In settings where atrocities have already begun, platforms may want to consider suspending interventions that add friction to users' ability to rapidly share content that may be necessary for their protection.

Last Resort or “Break Glass” Measures Interventions that temporarily and intentionally disable or degrade social media features in moments of heightened atrocity risk	**THEORY OF CHANGE:** Where social media features are at risk of being abused to contribute to atrocities, disabling features reduces the tools available to perpetrators.	**EXAMPLES:** • Intentionally disabling features that allow users to share hashtags, to avoid inciting violence in Ethiopia • Intentionally slowing down or degrading the functionality of certain features (i.e., adding friction) to prevent content from rapidly circulating on social media	**PRELIMINARY GUIDANCE FOR PLATFORMS:** • In light of the gravity and irremediability of mass atrocities, platforms should keep on the table interventions that would temporarily degrade or disable platform features at risk of severe abuse by atrocity perpetrators. • At the same time, because of the dual-use nature of most social media features, these measures should typically be used as a last resort or “break glass” measure, deployed only after assessing relevant limitations and trade-offs.

3. General Recommendations to Platforms

Finally, this section summarizes overarching recommendations that apply broadly when developing or deploying digital tools and interventions and are not specific to individual categories of tools. They arose in consultations focused on interventions for both civilian protection and degrading the capacity of atrocity perpetrators.

Platforms should invest in research and development concerning social media tools that hold potential to help prevent mass atrocities. The inventory of tools in this report offers a starting point for both deepening understanding of when and how different tools can address mass atrocity risks and expanding the range of available tools.

a. Invest in Atrocity Prevention Capacity and Expertise

Platforms should invest in building internal atrocity prevention capacity and expertise. They should ensure they have a dedicated crisis response function that can define and categorize potential atrocity risk situations according to a principled risk assessment process, coordinate between teams to collaborate on potential issues, and develop clear protocols on when various interventions and policies will be deployed. Platforms should also ensure that, when specific interventions must be deployed by relevant teams, those teams can obtain needed resources. As articulated in recent US-EU guidance for social media companies on protecting human rights defenders, platforms should "mobilize additional capacity when they identify a foreseeable risk of harm."[131] In addition, platforms should hold tabletop or scenario-based simulations to prepare for atrocity risk settings.[132]

More broadly, platforms should build their awareness on how their products are being used in atrocity risk settings to create a baseline for further assessment of risks and opportunities. According to interviewees, identifying what content is most viewed and engaged with, as well as which accounts have greatest visibility and reach in an atrocity risk context, for example, could help platforms better assess the need for further interventions. This can help companies avoid investing time and energy into features that will have only minimal impact on civilian protection or degrading perpetrators, and avoid instances where efforts are made to modify product features that ultimately do not pose significant risks. Throughout the consultations for this report, social media companies were urged to invest in understanding how products are being used in contexts where there is a risk of mass atrocities, rather than making rash decisions to restrict product features that may also have important and positive applications. Interviewees also called for platforms to invest in robust trust and safety teams that would have the capacity to monitor for abuse of the platform and be proactive in removing dangerous content.

> ***"This is not a story of technical tools—it would be really nice if it was. This requires understanding atrocity dynamics, and it takes lots of human resources to do that well."***
>
> \- Expert on digital technology and international law

In some cases, interventions referenced in this report may be in tension or conflict with one another. For example, exploring opportunities to permit graphic content that provides early warning of atrocities may make efforts to limit the presence of dangerous or inflammatory content on social media more difficult. Expanded internal capacity to understand local risk dynamics, and experienced crisis response teams, can help navigate these difficult decisions, in partnership with affected communities wherever possible.

Rather than solely focusing on specific interventions, interviewees urged all stakeholders to invest in obtaining greater understanding of threats in the information environment. As one participant put it, “More than focusing on any one product or policy intervention, we need to foster greater openness on what platforms have done in the past, share access to that data, and really understand what’s working so resources can be directed effectively.” This may also involve supporting independent research on areas like disinformation, as well as how it plays out on social media platforms.

b. Preserve Digital Evidence of Mass Atrocities

Without precautions, tools and interventions that aim to address dangerous, graphic, or inflammatory social media content may inadvertently contribute to the loss of potential evidence of atrocity crimes that holds important value for justice and accountability efforts.[133] Platforms should preserve digital evidence of mass atrocities, and, where appropriate, share information to assist in the investigation and prosecution of atrocity crimes. They should also clarify their policies on data preservation in atrocity risk and conflict settings, and consult with civil society organizations (and, where feasible, affected communities) to identify content relevant to international justice and accountability efforts.[134] Preservation of digital evidence may fall within the confines of “civilian protection” by raising the cost of violence, supporting a long-term theory of deterrence, or may constitute an opportunity to degrade perpetrator capacity by contributing to justice and accountability efforts. Even in the event that it does not fit neatly into these strategies, it warrants mention as a core consideration when implementing interventions aimed at atrocity prevention.

Interviewees noted that the preservation of digital content is only effective if it is ultimately shared with investigative and prosecutorial authorities pursuing accountability initiatives. Facebook’s Oversight Board, for example, has found that Facebook “has a responsibility to collect, preserve and, where appropriate, share information to assist in the investigation and potential prosecution of grave violations of international criminal, human rights and humanitarian law by competent authorities and accountability mechanisms.”[135] Preservation initiatives may also be undertaken to respond to targeted requests from legal entities charged with gathering evidence, open calls for potential evidence issued by entities with the mandate to investigate and prosecute core international crimes, or in response to users themselves flagging digital content that could be useful as future evidence, should platforms make that option available.[136]

c. Localize Resources

Platforms should localize all resources to ensure accessibility and ease of use for affected communities. Any tools or interventions developed for use by individuals in at-risk communities must be made available in the relevant local languages of affected populations. Effectively serving communities affected by atrocity risks requires ensuring that not only the tools, but also communications around their roll-out, launch, and risks are accessible and understandable to those that may need to use them. Interviewees suggested, for example, localizing help center articles or blog posts by social media companies around their actions in atrocity risk contexts. Because localization may take time and resources, it should ideally be planned and coordinated well in advance of the launch of relevant features and policies.

d. Invest in Local Partnerships

Platforms should invest heavily in local partnerships that can support awareness of atrocity risk dynamics. Across the board, interviewees emphasized that local partnerships are the cornerstone of atrocity prevention interventions, and local partners are essential to support platform awareness on how atrocities risks intersect with online dynamics. Interviewees urged platforms to build local relationships well in advance of crisis events, and to proactively establish clear processes and channels through which civil society organizations can report dangerous content.

> ***“The question that underlies these conversations is, what is the investment in local connections? I don’t think any of this can be done well without that.”***
>
> - Expert on digital technology and international law

The implementation of interventions should be informed by local experts and organizations, with tools implemented in relevant local languages. Indeed, core principles aimed at interventions to counteract dangerous speech find broad applicability for social media interventions to support atrocity prevention: interventions should (a) be developed in partnership with local partners, (b) be goal oriented and strategic, and (c) do no harm, while managing relevant risks.[137]

Platforms should provide training on relevant product and policy interventions so they can be rolled out more effectively in at-risk communities.[138] At the same time, interviewees referenced the difficulties of identifying fact-checkers and partners in contexts where there are acute physical safety and security risks, as local partners may need to leave the country, creating gaps in available credible information. Although many platforms already have formal partnership programs, experts urged companies to make a greater investment in ensuring these are mutually supportive rather than extractive, and in ensuring that external partners understand the impact of their contributions.

e. Cross-platform Communications

In addition to partnerships with civil society, platforms should expand communication with peer companies to support collective efforts to protect civilians and degrade perpetrator capacity. This may include sharing information about risk dynamics, vulnerable groups in need of protection, or particularly inflammatory online dynamics that pose a heightened risk of mass atrocities. Given the frequent migration of dangerous content from one platform to another, interviewees argued that perpetrators may be less able to weaponize social media overall if platforms expand coordination of efforts and strategically share information relevant to atrocity risk contexts. This may include, for example, sharing potentially dangerous narratives that may be surging, or strategies through which perpetrators are seeking to evade platform policies. Interviewees also pointed to an intervention called “hashing” dangerous content, whereby platforms can quickly identify visually similar content that has been removed by another platform.[139] This, however, requires close coordination between platforms, which, according to some interviewees, has been *ad hoc* to date.

B. Recommendations for Policy Makers

Most of the recommendations set out in this report are aimed at platforms as the primary actors in conceptualizing, developing, and deploying the types of features and interventions described herein. This report, however, is not solely aimed at platforms, but also at atrocity prevention policy makers responsible for developing strategies that could better integrate digital tools and interventions. First and foremost, policy makers should ensure that atrocity prevention strategies include an assessment of both risks and opportunities in the digital environment, taking into account how both at-risk communities and perpetrators are using social media. Further, policy makers should consider taking the following actions:

- Partner with social media platforms to research the benefits and risks of specific interventions in atrocity risk settings;
- Establish dedicated channels for communication between the atrocity prevention community and social media companies;
- Engage in greater information sharing with social media companies on settings where there is a heightened risk of mass atrocities, with the aim of raising awareness of the need for digital interventions;
- Explore opportunities to share atrocity prevention expertise with platforms, to support the development and deployment of interventions focused on prevention; and
- Explore opportunities to incorporate social media tools and interventions into atrocity prevention strategies.

ACKNOWLEDGMENTS

This report is dedicated to the thoughtful community of trust & safety practitioners, who work (tirelessly, invisibly) behind the scenes to address online dynamics that could fuel offline violence. This community constitutes, in my view, a new form of human rights defenders, and I hope this report helps to spotlight and empower your work.

This report was made possible by the Leonard and Sophie Davis Genocide Prevention Fellowship, and the generous support of the United States Holocaust Memorial Museum's Simon-Skjodt Center for the Prevention of Genocide. I am particularly grateful to Naomi Kikoler and Lawrence Woocher for being willing to support a novel (and niche!) area of atrocity prevention research, and for working with me at length to bring this project to life. This report builds on the body of past work that the Simon-Skjodt Center has published, and I am grateful to the entire team for sharing their expertise and ideas in support of this report. Finally, special thanks goes to Tallan Donine, who not only assisted with research, consultations, and editing of this report, but who was a trusted thought partner across the entire lifespan of this project. Thank you so much for the care and attention you put into everything you do.

In developing this report, I benefited from the time and insights of a great number of people with diverse areas of expertise, who were generous enough to speak with me and share their perspectives on the potential of social media for atrocity prevention. This includes current and former representatives of social media companies, atrocity prevention experts and practitioners, and members of at-risk communities. While some of them preferred to not be named, this report draws greatly on their expertise and lived experiences at the intersection of digital dynamics and atrocity risks. My deep gratitude goes to the following individuals, whose thoughtful insights and experience in this field were foundational to the development of this report: Brian Clarke, Federica D'Alessandra, Jenny Domino, Maksym Dvorovyi, Jennifer Easterday, Matt Graydon, Joseph Guay, Rebecca Hamilton, Brittan Heller, Kristina Hook, Ravi Iyer, Alexa Koenig, Iain Levine, Raquel Vazquez Llorente, Rhiannon Neilsen, Mia Newman, Sarah Oh, Betsy Popken, Homira May Rezai, Yoel Roth, Miranda Sissons, Theodora Skeadas, Sam Toizer, Christopher Tuckwood, and Ernesto Verdeja.

Shannon Raj Singh

ABOUT THE AUTHOR

Shannon Raj Singh served as the Leonard & Sophie Davis Genocide Prevention Fellow at the Simon-Skjodt Center for the Prevention of Genocide, US Holocaust Memorial Museum (2023-2024). She is an expert in the intersection of digital technology and mass violence, and the former Human Rights Counsel for Twitter, where she advised on risks related to the platform's use in conflicts such as Ethiopia, Afghanistan, and Ukraine. Today, she is the Principal and Founder of Athena Tech & Atrocities Advisory, where she advises governments, international organizations, and civil society groups on the digital dynamics of armed conflict and opportunities to leverage technology for civilian protection. She is Co-Chair of the War Crimes Committee of the International Bar Association, and a Special Advisor on Social Media & Conflict to the Centre for Humanitarian Dialogue, a Geneva-based organization supporting peace negotiations around the world. Shannon has served as a legal advisor at the Special Tribunal for Lebanon, a legal fellow at the International Criminal Tribunal for Rwanda, and a Visiting Fellow of Practice with the Oxford Programme on International Peace and Security. She holds a JD from the University of Southern California and a dual BA from UCLA.

ENDNOTES

[1] Samantha Power, "Bystanders to Genocide," *The Atlantic*, September 2021, https://www.theatlantic.com/magazine/archive/2001/09/bystanders-to-genocide/304571/; see also Samantha Power, *"A Problem from Hell": America and the Age of Genocide* (Basic Books, 2002), 384.

[2] Remarks by Ambassador Samantha Power at POLITICO's Women Rule Event, *Politico*, November 21, 2013 (as provided by the United States Mission to the United Nations), https://www.politico.com/story/2013/11/samantha-power-remarks-women-rule-event-100205.

[3] Tallan Donine, "Introducing a Strategic Framework for Helping Prevent Mass Atrocities," *Genocide Prevention Blog*, United States Holocaust Memorial Museum, September 5, 2023, https://www.ushmm.org/genocide-prevention/blog/introducing-a-strategic-framework-for-helping-prevent-mass-atrocities.

[4] "Simon-Skjodt Center Deputy Director Testifies at Hearing on Mass Atrocities," *Genocide Prevention Blog*, United States Holocaust Memorial Museum, February 6, 2018, https://www.ushmm.org/genocide-prevention/blog/simon-skjodt-center-deputy-director-testifies-at-hearing-on-mass-atrocities.

[5] For purposes of this report, the concepts of social media tools and interventions are used relatively interchangeably to denote products, policies, and tactics that may be used on social media to support atrocity prevention strategies. Generally, the concept of social media "tools" references individual features or "products," while "interventions" reference the broader initiatives that make use of those tools.

[6] See, e.g., David J. Simon; Samhitha Josyula; Joshua Lam; and Julian D. Melendi (2024) "Atrocity Prevention in the Digital Era: Adapting Norms, Laws, and Code to Changes in the Ways Atrocities Are Committed," Genocide Studies and Prevention: An International Journal: Vol. 17: Iss. 3: 1–23, pp. 2-4; Rebecca Hamilton, Platform-Enabled Crimes: Pluralizing Accountability When Social Media Companies Enable Perpetrators to Commit Atrocities, 63 Boston College Law Review 1349 (2022) https://digitalcommons.wcl.american.edu/facsch_lawrev/2220; Federica D'Alessandra, Ross James Gildea, 'Technology, R2P, and the UN Framework of Analysis for Atrocity Crimes," Global Responsibility to Protect (2024), pp. 7-12.

[7] United States Holocaust Memorial Museum (USHMM), "Social Media, Mass Atrocities, and Atrocity Prevention: 2023 Sudikoff Interdisciplinary Seminar on Genocide Prevention" (background paper, April 2023), https://www.ushmm.org/m/pdfs/2023_Sudikoff_Interdisciplinary_Seminar_on_Genocide_Prevention_-_Background_Paper.pdf.

[8] Caroline Crystal, "Facebook, Telegram, and the Ongoing Struggle against Online Hate Speech" (Carnegie Endowment for International Peace, Sept. 7, 2023), https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en.

[9] Kristina Hook and Ernesto Verdeja, "Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities" (Stimson Issue Brief, July 7, 2022), https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/.

[10] Global Centre for the Responsibility to Protect, "The Relationship between Digital Technologies and Atrocity Prevention" (Policy Brief, March 2024), 5, https://www.globalr2p.org/publications/the-relationship-between-digital-technologies-and-atrocity-prevention/.

[11] See, e.g., Caroline Crystal, "Facebook, Telegram, and the Ongoing Struggle against Online Hate Speech" (Carnegie Endowment for International Peace, Sept. 7, 2023), https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en.; Emma Irving, "Suppressing Atrocity Speech on Social Media," *AJIL Unbound* 113 (2019) doi:10.1017/aju.2019.46; Kristina Hook and Ernesto Verdeja, "Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities" (Stimson Issue Brief, July 7, 2022), https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/; USAID, "Atrocity Prevention: A Development Practitioner's Guide," 2024, Annex A, https://www.usaid.gov/sites/default/files/2024-05/Atrocity-Prevention-Guide-2024.pdf; "Defining a Brave New Field: Technology and the Protection of Civilians in Conflict," Columbia SIPA, https://www.sipa.columbia.edu/sites/default/files/migrated/downloads/International%2520Peace%2520Institute_Defining%2520a%2520Brave%2520New%2520Field.pdf.

[12] See, e.g., Ushahidi, https://www.ushahidi.com/; Peter van der Windt and Marcartan Humphreys, "Crowdseeding in Eastern Congo: Using Cell Phones to Collect Conflict Events Data in Real Time," *Journal of Conflict Resolution* 60, no. 4 (June 2016): 748–81; Charles Martin-Shields, "Information Communication Technologies in Atrocity Response and Prevention: Deepening Our Understanding of the Legal, Ethical and Practical Challenges," *Genocide Studies and Prevention: An International Journal* 11, no. 1 (2017): 100–3, doi: http://doi.org/10.5038/1911-9933.11.1.1484; Christopher Tuckwood, "The State of the Field: Technology for Atrocity Response," *Genocide Studies and Prevention: An International Journal* 8, no. 3 (2014): 81–86, doi: http://dx.doi.org/10.5038/1911-9933.8.3.7.

[13] See, e.g., The Sentinel Project, https://thesentinelproject.org/what-we-do/the-role-of-technology/; Agathe Sarfati, "New Technologies and the Protection of Civilians in UN Peace Operations" (International Peace Institute, September 2023), https://www.ipinst.org/wp-content/uploads/2023/09/IPI-E-RPT-New-Technologies.pdf; Kroc Institute for International Peace Studies, "Artificial Intelligence, Social Media, and Political Violence Prevention, n.d., https://kroc.nd.edu/research/artificial-intelligence-social-media-and-political-violence-prevention/.

[14] See, e.g., Simone Bunse, "Social Media: A Tool for Peace or Conflict?" (Stockholm International Peace Research Institute backgrounder, August 20, 2021), https://www.sipri.org/commentary/topical-backgrounder/2021/social-media-tool-peace-or-conflict; DPPA Mediation Support Unit and swisspeace, "Social Media in Peace Mediation: A Practical Framework," June 2021), https://www.swisspeace.ch/assets/publications/downloads/PeaceMediationSocialMedia_SwissPeace_UNO_Web_v1.pdf.

[15] See, e.g., Kristin Bergtora Sandvik and Nathaniel A. Raymond, "Beyond the Protective Effect: Towards a Theory of Harm for Information Communication Technologies in Mass Atrocity Response," *Genocide Studies and Prevention: An International Journal* 11, no. 1 (2017): 9–24; Peter Mandaville and Julia Schiwal, "A New Approach for Digital Media, Peace and Conflict" (United States Institute of Peace, February 15, 2023), https://www.usip.org/publications/2023/02/new-approach-digital-media-peace-and-conflict.

[16] Rebecca Hamilton, "Atrocity Prevention in the New Media Landscape," Symposium on Non-State Actors and New Technologies in Atrocity Prevention, 2019, https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=2290&context=facsch_lawrev.

[17] See, e.g., Sandvik and Raymond, "Beyond the Protective Effect."

[18] "Tools for Atrocity Prevention," United States Holocaust Memorial Museum website, https://preventiontools.ushmm.org/.

[19] Some of the digital tools and interventions discussed in this report could also be considered information tools, defined as measures "intended to heighten awareness of the situation, gain support for [national] policy, and convince perpetrators and their supporters that they are being watched, which may dissuade them from conducting criminal behavior for which they could be held accountable" as well "measures that support strategic communication to advance [government] themes and messages as well as measures, including intelligence, to improve situational understanding." *Mass Atrocity Prevention and Response Options (MAPRO): A Policy Planning Handbook* (Carlisle, PA: U.S. Army Peacekeeping and Stability Operations Institute, March 2012), https://pksoi.armywarcollege.edu/wp-content/uploads/2020/07/MAPRO_handbook_final.pdf. In other cases, however, digital tools discussed in this report do not fall squarely within that category. Such tools include those that seek to conceal information in the name of civilian protection, such as privacy measures; efforts to protect civilians against hacking or impersonation attempts; or initiatives to restrict the presence of dangerous digital content to avoid it from contributing to atrocity risk dynamics. For this reason, this report has posited a distinctive category of digital tools.

[20] USHMM, "Social Media, Mass Atrocities, and Atrocity Prevention: 2023 Sudikoff Interdisciplinary Seminar on Genocide Prevention."

[21] "A Strategic Framework for Helping Prevent Mass Atrocities" (Simon-Skjodt Center for the Prevention of Genocide Special Report, United States Holocaust Memorial Museum, September 2023), https://www.ushmm.org/m/pdfs/A_Strategic_Framework_for_Helping_Prevent_Mass_Atrocities_.pdf.

[22] "A Strategic Framework for Helping Prevent Mass Atrocities."

[23] "A Strategic Framework for Helping Prevent Mass Atrocities."

[24] Executive Summary, "Lessons Learned in Preventing and Responding to Mass Atrocities," United States Holocaust Memorial Museum website, last updated April 2024, https://www.ushmm.org/genocide-prevention/simon-skjodt-center/work/lessons-learned/summary.

[25] "A Strategic Framework for Helping Prevent Mass Atrocities."

[26] Anderson Cooper, "How a WWII-Era Forger Saved Lives One Fake Document at a Time," CBS News *60 Minutes*, October 29, 2017, https://www.cbsnews.com/news/how-a-wwii-era-forger-saved-lives-one-fake-document-at-a-time/; "Thousands of False Identities," Twelve Years That Shook the World Podcast, United States Holocaust Memorial Museum, December 18, 2018, https://www.ushmm.org/learn/podcasts-and-audio/12-years-that-shook-the-world/thousands-of-false-identities.

[27] Evan Gerstmann, "Are Jewish Students Feeling Forced to Hide Their Identity on Campus?," *Forbes.com*, updated September 24, 2021, https://www.forbes.com/sites/evangerstmann/2021/09/24/are-jewish-students-feeling-forced-to-hide-their-identity-on-campus/?sh=4bd61b9178e8.

[28] As an example, Ukrainians who openly shared information about their sexual orientation on social media in the first post-Soviet country to decriminalize homosexuality could find that their old posts presented new and acute risks in areas that suddenly came under Russian control. See, e.g., National LGBTI Consortium, "The Situation of LGBTI People in the Temporarily Occupied Territories of Ukraine," https://lgbti-consortium.org.ua/en/projects/stanovyshhe-lgbti-lyudej-na-tymchasovo-okupovanyh-terytoriyah-ukrayiny/, and Johanna Chisholm, "Analysis: Putin's Homophobia Is Advancing LGBTQ Rights in Ukraine," FP, April 16, 2023, https://foreignpolicy.com/2023/04/16/ukraine-russia-war-putin-homophobia-lgbtq-rights-military-civil-unions/.

[29] Facebook Help Center, "Lock Your Facebook Profile," https://www.facebook.com/help/196419427651178; Nathanial Gleicher (@ngleicher@infosec.exchange), post on Twitter, August 19, 2021, 5:48 p.m., https://twitter.com/ngleicher/status/1428474000611573762.

[30] Rafael Frankel, "An Update on the Situation in Myanmar," post on Meta, Feb. 11, 2021, https://about.fb.com/news/2021/02/an-update-on-myanmar/; KrASIA, "Facebook Rolls Out New Safety Feature to Protect Myanmar Protestors," Apr. 6, 2021, https://kr-asia.com/facebook-rolls-out-new-safety-feature-to-protect-myanmar-protestors; Manish Singh, "Facebook Rolls Out Feature to Help Women in India Easily Lock Their Accounts," *TechCrunch*, May 21, 2020, https://techcrunch.com/2020/05/21/facebooks-new-safety-feature-for-women-in-india-easily-lock-the-account-from-strangers/; as well as interviews conducted for this report.

[31] Kim Lyons, "Facebook Hides Friends Lists on Accounts in Afghanistan as a Safety Measure," *The Verge*, August 20, 2021, https://www.theverge.com/2021/8/20/22634209/facebook-hides-friends-lists-instagram-safety-afghanistan-taliban-security; as well as interviews conducted for this report.

[32] Nathaniel Gleicher (@ngleicher@infosec.exchange), post on Twitter, February 24, 2022, 1:07 p.m., https://twitter.com/ngleicher/status/1496909655754219526; Karissa Bell, "Facebook Turns on 'Lock Profile' Tool for People in Ukraine," *Engadget*, February 24, 2022, https://www.engadget.com/facebook-turns-on-lock-profile-tool-people-in-ukraine-184946515.html.

[33] Lyons, "Facebook Hides Friends Lists"; Nathaniel Gleicher, "Updates on Our Security Work in Ukraine," Meta, February 27, 2022, https://about.fb.com/news/2022/02/security-updates-ukraine/.

[34] "Meta's Ongoing Efforts Regarding Russia's Invasion of Ukraine," Meta, February 26, 2022, https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/.

[35] X Help Center, "How to Use X Lists," https://help.x.com/en/using-x/x-lists.

[36] For a description of the targeted harassment campaigns that can follow from adding users to Lists, see Samantha Cole, "Here's Why You're Being Added to Twitter Lists All of a Sudden," *Vice*, July 28, 2020, https://www.vice.com/en/article/twitter-list-notifications-vicariously-app/.

[37] In the context of Ukraine, for example, Twitter issued a public statement that it was "working across features like Topics, Lists, and Spaces to reinforce safety measures and guard against misuse." Sinead McSweeney, "Our Ongoing Approach to the War in Ukraine," *X Blog*, March 16, 2022, https://blog.twitter.com/en_us/topics/company/2022/our-ongoing-approach-to-the-war-in-ukraine. Relevant interventions may include, for example, an abuse-reporting feature for Lists or the ability for users to remove themselves from Lists to which they were added without their consent.

[38] *See*, e.g., ICRC, "Digital Dilemmas: Immersive experience exposes technology threats," Inspired, August 25, 2023, https://blogs.icrc.org/inspired/2023/08/25/digital-dilemmas-immersive-experience-exposes-technology-threats/, and corresponding exhibit.

[39] To draw from an example outside social media, in the context of Ukraine, Google temporarily disabled foreign access to its features that could depict live traffic and "area busyness" on Google Maps after consulting with Ukrainian authorities. Despite the potential for abuse by Russian perpetrators, countervailing considerations may have included civilians' ability to access information about traffic conditions along evacuation routes.

[40] Twitter Safety (@TwitterSafety), February 24, 2022, 4:09 a.m., https://web.archive.org/web/20220225234015/https://twitter.com/TwitterSafety/status/1496698664747687942.

[41] Katie Polglase and Gianluca Mezzofiore, "Iran's Government Accesses the Social Media Accounts of Those It Detains," CNN Business, December 19, 2022, https://www.cnn.com/2022/12/19/business/iran-social-media-accounts-intl-cmd/index.html.

[42] *See also* the digital security work of human rights organizations in this space, such as Access Now's Digital Security Helpline, https://www.accessnow.org/help/, and Frontline Defenders' Digital Security Resources, https://www.frontlinedefenders.org/en/digital-security-resources.

[43] Elizabeth Culliford, "Facebook, Twitter and LinkedIn Secure Afghan Users' Accounts amid Taliban Takeover," Reuters, August 19, 2021, https://www.reuters.com/technology/facebook-says-it-is-securing-afghan-user-accounts-amid-taliban-takeover-2021-08-19/.

[44] TripZilla Philippines, "The 'Facebook Profile Lock' and How Filipinos Can Do It Locally," July 19, 2021, https://www.tripzilla.ph/facebook-profile-lock/29296/.

[45] Grindr Help Center, Settings, https://help.grindr.com/hc/en-us/articles/1500011544001-Settings.

[46] Viktorya Vilk, Elodie Vialle, and Matt Bailey, *No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users*(New York: PEN America, 2021), https://pen.org/report/no-excuse-for-abuse/.

[47] Vilk, Vialle, and Bailey, *No Excuse for Abuse*.

[48] Jill Capotosto, "The Mosaic Effect: The Revelation Risks of Combining Humanitarian and Social Protection Data," *ICRC Humanitarian Law & Policy Blog*, February 9, 2021, https://blogs.icrc.org/law-and-policy/2021/02/09/mosaic-effect-revelation-risks/.

[49] International Committee of the Red Cross (ICRC), "Information for People Affected by the Conflict in Israel," October 20, 2023, https://www.icrc.org/en/document/information-people-affected-conflict-israel.

[50] Nicole Perfroth,"What Is End-to-End Encryption? Another Bull's-Eye on Big Tech," *New York Times*, November 19, 2019, https://www.nytimes.com/2019/11/19/technology/end-to-end-encryption.html.

[51] Content "ranking" refers to the ways in which social media platforms determine the order of content visible across various surfaces of a social media platform, such as their newsfeed or timeline. For example, Facebook describes its approach to ranking content in user feeds here: Meta Transparency Center, "Our Approach to Facebook Feed Ranking," November 28, 2023, https://transparency.meta.com/features/ranking-and-content/.

[52] At the same time, other participants referenced the difficulty of relying on these interventions in settings with limited connectivity, such as during internet blackouts.

[53] Emerging regulations, such as the European Union's Digital Services Act, also require certain social media companies to provide greater information and optionality around the ranking and prioritization of digital content. See, e.g., European Commission website, The Digital Services Act, "Europe Fit for the Digital Age: New Online Rules for Platforms," https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act/europe-fit-digital-age-new-online-rules-platforms_en; Steven Lee Meyers, "E.U. Law Sets the Stage for a Clash over Disinformation," *New York Times*, Sept. 27, 2023, https://www.nytimes.com/2023/09/27/technology/disinformation-law-european-union.html.

[54] See, e.g., Matt O'Brien, "Facebook Bans Holocaust Denial, Distortion Posts," PBS News, October 12, 2020, https://www.pbs.org/newshour/nation/facebook-bans-holocaust-denial-distortion-posts.

[55] Participants pointed to past instances, for example, when platforms have consulted with civil society to develop a white list of reliable independent media for amplification.

[56] See https://www.signpost.ngo/.

[57] X, "World News: Israeli Fighter Jets Strike Gaza after Rockets Are Launched into Israel," April 20, 2022, https://x.com/i/events/1516902141046132736.

[58] Joanna Geary, "Bringing More Reliable Context to Conversations on Twitter," *X Blog*, August 2, 2021, https://blog.x.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter.

[59] Sarah Emerson and Richard Nieva, "Google Maps Blocked Edits after People Falsely Claimed It Was Used to Coordinate Russian Air Strikes," *BuzzFeed.News*, March 3, 2022, https://www.buzzfeednews.com/article/sarahemerson/russia-google-maps-tags-ukraine.

[60] In addition, social media data may be leveraged to better understand the risk of mass atrocities, such as through the use of large language models to understand potential atrocity risks. Although outside the scope of this report, various initiatives have analyzed social media to identify civilians in need of protection or to monitor trends around hate speech and expand early-warning capability.

[61] "Defining a Brave New Field: Technology and the Protection of Civilians in Conflict."

[62] Pariesa Brody, "Ukraine Siren Alerts: How a New Online System Updates Ukrainians about Air Raids," France 24, *The Observers*, April 19, 2022, https://observers.france24.com/en/tv-shows/the-observers/20220419-ukraine-siren-alerts-an-online-notification-system-for-air-raid-sirens.

[63] Chris Stephen, "Libya's Downward Spiral to Shortages, Militia Power, and Migrant Abuse," *IRIN News*, December 7, 2017, https://webarchive.archive.unhcr.org/20230519163516/https://www.refworld.org/docid/5a2a573f4.html.

[64] Louisa Loveluck, "The Secret App That Gives Syrian Civilians Minutes to Escape Airstrikes," *Washington Post*, August 18, 2018, https://www.washingtonpost.com/world/the-secret-app-that-gives-syrian-civilians-minutes-to-escape-airstrikes/2018/08/17/e91e66be-9cbf-11e8-b55e-5002300ef004_story.html; as well as interviews conducted for this report.

[65] Zachariah Mampilly et al., "The Role of Civilians and Civil Society in Preventing Mass Atrocities" (special report, Simon-Skjodt Center for the Prevention of Genocide, United States Holocaust Memorial Museum, Washington, DC, July 2020), https://www.ushmm.org/m/pdfs/TheRoleofCivilians.pdf.

[66] See, e.g., Alexa Koenig and Andrea Lampros, *Graphic: Trauma and Meaning in Our Online Lives* (Cambridge: Cambridge University Press, 2023).

[67] Mia Sato, "Google Will Send Air Raid Alerts to Android Phones in Ukraine," *The Verge*, Mar. 10, 2022, https://www.theverge.com/2022/3/10/22971213/google-ukraine-android-air-raid-alerts; Zack Baddorf, "Central Africans Use Radio Network to Stay Safe from LRA," VOA, Feb. 1, 2017, https://www.voanews.com/a/central-africans-use-radio-network-to-stay-safe-from-lra/3701720.html.

[68] United Nations Office on Genocide Prevention and the Responsibility to Protect, *Framework of Analysis for Atrocity Crimes: A Tool for Prevention* (New York: United Nations, 2014), https://www.un.org/en/genocideprevention/documents/about-us/Doc.3_Framework%20of%20Analysis%20for%20Atrocity%20Crimes_EN.pdf.

[69] Global Centre for the Responsibility to Protect, "The Relationship between Digital Technologies and Atrocity Prevention."

[70] Mampilly et al., "The Role of Civilians."

[71] X Help Center, Communities on X, https://help.x.com/en/using-x/communities; Facebook Help Center, Groups, https://www.facebook.com/help/1629740080681586.

[72] TikTok, Share your favorite TikTok moments with Direct messaging, Aug 12, 2024, https://newsroom.tiktok.com/en-us/share-your-favorite-tiktok-moments-with-direct-messaging.

[73] WhatsApp Help Center, About Building Private, Safe, and Secure Communities on WhatsApp, https://faq.whatsapp.com/458610306367976.

[74] "Connecting to WhatsApp by Proxy," WhatsApp blog, January 5, 2023, https://blog.whatsapp.com/connecting-to-whatsapp-by-proxy; Stephanie Stacey, "WhatsApp's New Feature Could Help Iranians Bypass Censors and Coordinate Protests during Government-Imposed Internet Blackouts," *Business Insider*, January 7, 2023, https://www.businessinsider.com/whatsapps-new-feature-could-help-iranian-protesters-bypass-censors-2023-1.

[75] Jon Fingas, "Twitter Launches a Tor Service to Help Russians Evade Censorship," Engadget, March 8, 2022, https://www.engadget.com/twitter-tor-onion-service-evade-censorship-210549633.html; Guardian staff and agencies, "Twitter Launches Privacy-Protected Site on Dark Web to Bypass Russia's Block," *The Guardian*, March 9, 2022, https://www.theguardian.com/technology/2022/mar/09/twitter-tor-version-russia-block.

[76] See, e.g., "Defining a Brave New Field: Technology and the Protection of Civilians in Conflict," Columbia SIPA, https://www.sipa.columbia.edu/sites/default/files/migrated/downloads/International%2520Peace%2520Institute_Defining%2520a%2520Brave%2520New%2520Field.pdf.

[77] Outgoing code cable dated January 11, 1994, from Dallaire to General J. G. Maurice Baril, head of the military division of the UN Department of Peacekeeping Operations, PBS *Frontline* old website, https://www.pbs.org/wgbh/pages/frontline/shows/evil/warning/cable.html.

[78] Cable dated January 11, 1994, from Dallaire to Baril, PBS *Frontline* old website.

[79] "Strategic Framework for Helping Prevent Mass Atrocities," 7.

[80] See, e.g., works by Ben Nimmo and Eric Hutchins, including "Phase-based Analysis of Online Operations" (working paper, Carnegie Endowment for International Peace, March 2023) and "The Online Operations Kill Chain: A Model to Analyze, Describe, Compare, and Disrupt Threat Activity from Influence Operations to Cybercrime" (paper, Carnegie Endowment for International Peace, March 16, 2023).These experts have found that, despite the array of threat actors in the online environment, who "range from intelligence agencies and troll farms to child-abuse networks," as well as perpetrators of mass atrocities, these widely different actors "may follow the same chain of steps." The categories reflected here do not map directly to the kill chain posited by Nimmo and Hutchins, but they include observations relevant to several possible links in the chain relevant to atrocity perpetrators.

[81] Ben Nimmo and Eric Hutchins, "A New Kill Chain Approach to Disrupting Online Threats," *Lawfare*, June 23, 2023, https://www.lawfaremedia.org/article/a-new-kill-chain-approach-to-disrupting-online-threats.

[82] Nimmo and Hutchins, "Phase-based Analysis of Online Operations."

[83] Nimmo and Hutchins, "Phase-based Analysis of Online Operations."

[84] There were several technical suggestions as to how this intervention might be implemented that go beyond the scope of this paper.

[85] Meta, "Removing Myanmar Military Officials from Facebook," August 28, 2018, https://about.fb.com/news/2018/08/removing-myanmar-officials/; Eli Meixler, "Facebook Bans more Myanmar Military-Linked Accounts for Spreading Propaganda," *Time*, October 16, 2018, https://time.com/5425609/facebook-remove-myanmar-military-accounts/.

[86] Some participants also suggested that it may sometimes be useful to attribute coordinated inauthentic behavior to specific perpetrators, meaning that platforms that detect and remove these networks may choose to take the further step of sharing their findings. For example, if a social media platform were to expose the network of accounts behind a particular dangerous narrative, such as the concept of "denazification" in Ukraine (used as a justification for Russia's invasion), it could help demonstrate the hollowness of the claim.

[87] Meta Transparency Center, "Dangerous Organizations and Individuals," https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/; Sudan's Military Leader Accuses Rival of Committing War Crimes," *Al Jazeera*, August 14, 2023, https://www.aljazeera.com/news/2023/8/14/sudans-military-leader-accuses-rival-of-committing-war-crimes.

[88] TikTok, "Safety and Civility," April 17, 2024, https://www.tiktok.com/community-guidelines/en/safety-civility/.

[89] Merrill Perlman, "The Rise of 'Deplatform,'" *Columbia Journalism Review*, February 4, 2021, https://www.cjr.org/language_corner/deplatform.php.

[90] See, e.g., Danny Klinenberg, "Does Deplatforming Work?" *Journal of Conflict Resolution* 68(6), 2024, 1199-1225; Regine Cabato and Rebecca Tan, "Meta Oversight Board calls for Cambodian leader's accounts to be suspended," *Washington Post*, June 29, 2023, https://www.washingtonpost.com/world/2023/06/29/meta-hun-sen-cambodia-suspended-accounts/.

[91] United Nations, *Framework of Analysis for Atrocity Crimes*.

[92] Nimmo and Hutchins, "Phase-based Analysis of Online Operations" Phases 3–4.

[93] Nimmo and Hutchins, "Phase-based Analysis of Online Operations."

[94] Nimmo and Hutchins, "Phase-based Analysis of Online Operations."

[95] See, e.g., Meta Terms and Policies, Prohibited Content, Weapons, Ammunition, and Explosives, https://www.facebook.com/policies_center/commerce/weapons_ammunition_and_explosives; TikTok, Regulated Goods and Commercial Activities, https://www.tiktok.com/community-guidelines/en/regulated-commercial-activities; X, Illegal or certain regulated goods or services, https://help.x.com/en/rules-and-policies/regulated-goods-services; Snap, Community Guidelines: Hateful Content, Terrorism, and Violent Extremism, https://values.snap.com/policy/policy-community-guidelines?lang=en-US#:~:text=Terrorist%20organizations%2C%20violent%20extremists%2C%20and,advances%20terrorism%20or%20violent%20extremism.

[96] On Pages, see Facebook Help Center, Create and Manage a Page, https://www.facebook.com/help/135275340210354. These efforts, however, should nevertheless respect the right to privacy and focus on the enhanced moderation of public and semi-public spaces.

[97] See, e.g., X Developer Platform, "More about Restricted Uses of the X APIs," https://developer.x.com/en/developer-terms/more-on-restricted-use-cases.

[98] X Developer Platform, "More about Restricted Uses of the X APIs."

[99] See, e.g., Twitter Help Center (May 2022) (via Wayback Machine); https://web.archive.org/web/20220519204048/https://help.twitter.com/en/rules-and-policies/crisis-misinformation.

[100] One participant described a platform's policy to address dangerous misinformation in fragile settings as particularly significant in atrocity risk settings, calling it essentially a "genocide prevention policy."

[101] Twitter Help Center (May 2022) (via Wayback Machine); https://web.archive.org/web/20220519204048/https://help.twitter.com/en/rules-and-policies/crisis-misinformation.

[102] Meta Transparency Center, "Misinformation," https://transparency.fb.com/policies/community-standards/misinformation; Snapchat, How We Prevent the Spread of False Information on Snapchat, https://values.snap.com/news/how-we-prevent-the-spread-of-false-information-on-snapchat?lang=en-US.

[103] See, e.g., Abbie Richards, Robin O'Luanaigh and Lea Marchl, "How 'Gnome Hunting' Became TikTok's Latest Antisemitic Dog Whistle," Global Network on Extremism and Technology, June 9, 2023, https://gnet-research.org/2023/06/09/how-gnome-hunting-became-tiktoks-latest-antisemitic-dog-whistle/#.

[104] See, e.g., Meta, Oversight Board Selects Case About a Post Discussing the Situation in Myanmar While Using Profanity, June 12, 2023, https://transparency.meta.com/oversight/oversight-board-cases/post-discussing-the-situation-in-myanmar-while-using-profanity; NBC News, Hate speech in Myanmar continues to thrive on Facebook, November 18, 2021, https://www.nbcnews.com/tech/tech-news/hate-speech-myanmar-continues-thrive-facebook-rcna5982.

[105] The term "shadow banning" is typically used to refer to "stealth actions by social media platforms to limit a post's visibility." Ephrat Livni, "What Is 'Shadow Banning'?," *New York Times*, January 13, 2023, https://www.nytimes.com/interactive/2023/01/13/business/what-is-shadow-banning.html; Geoffrey A. Fowler, "Shadow Banning Is Real: Here's How You End Up Muted by Social Media," *Washington Post*, December 27, 2022, https://www.washingtonpost.com/technology/2022/12/27/shadowban/.

[106] Jonathan Stray, Ravi Iyer, and Helena Puig Larrauri, "The Algorithmic Management of Polarization and Violence on Social Media" (23-05 Knight First Amendment Institute, August 22, 2023), https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media.

[107] Alex Hern, "WhatsApp to Restrict Message Forwarding after India Mob Lynchings," *The Guardian*, July 20, 2018, https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings.

[108] Keach Hagey and Jeff Horwitz, "Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead." The Wall Street Journal, September 15, 2021, https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215 ; Ian Anderson, Girzem Ceylan and Wendy Wood, "People share misinformation because of social media's incentives — but those can be changed," NiemanLab, August 8, 2023, https://www.niemanlab.org/2023/08/people-share-misinformation-because-of-social-medias-incentives-but-those-can-be-changed/.

[109] "Meta's Ongoing Efforts Regarding Russia's Invasion of Ukraine."

[110] A test run by Global Witness submitted requests for advertisements on Facebook containing hate speech in the context of Ethiopia, all of which were approved for publication. See "'Now Is the Time to Kill': Facebook Continues to Approve Hate Speech Inciting Violence and Genocide during Civil War in Ethiopia," Global Witness, June 9, 2022, https://www.globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/.

[111] X Help Center, "About Election Labels on X," https://help.x.com/en/using-x/election-labels; Justin Erlich, "TikTok's state-affiliated media policy," January 18, 2023, https://newsroom.tiktok.com/en-us/tiktoks-state-affiliated-media-policy; Patricia L. Moravec, Avinash Collis, Nicholas Wolczynski, "Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook," Information Systems Research, August 9, 2023, https://doi.org/10.1287/isre.2022.0305.

[112] Barbara Ortutay, "Twitter brings back election labels for 2020 US candidates," AP, December 12, 2019, https://apnews.com/article/us-news-elections-social-platforms-ca-state-wire-election-2020-186af7a45caa51b36edc78d40dc4ee78; as well as interviews conducted for this report.

[113] See also Raquel Vasquez Llorente, "How Musk's Twitter is Jeopardizing War Crimes Investigations," Tech Policy Press, July 10, 2023, https://www.techpolicy.press/how-musks-twitter-is-jeopardizing-war-crimes-investigations/; Shannon Raj Singh, "What Happens When We Get What We Pay for: Generative AI and the Sale of Digital Authenticity," Just Security, June 20, 2024, https://www.justsecurity.org/96924/generative-ai-and-the-sale-of-digital-authenticity/.

[114] Dara Kerr, "Twitter Once Muzzled Russian and Chinese State Propaganda. That's over Now," NPR, *All Things Considered*, April 21, 2023, https://www.npr.org/2023/04/21/1171193551/twitter-once-muzzled-russian-and-chinese-state-propaganda-thats-over-now; McSweeney, "Our Ongoing Approach to the War in Ukraine."

[115] Some participants were uncertain whether the usefulness warranted the investment and called for platforms to collaborate with academic researchers to measure the impact of labeling initiatives.

[116] Peter Dizikes, "The Catch to Putting Warning Labels on Fake News," MIT News | Massachusetts Institute of Technology, March 3, 2020, https://news.mit.edu/2020/warning-labels-fake-news-trustworthy-0303; Jon Bateman and Dean Jackson, *Countering Disinformation Effectively: An Evidence-Based Policy Guide* (Washington, DC: Carnegie Endowment for International Peace, 2024), https://carnegie-production-assets.s3.amazonaws.com/static/files/Carnegie_Countering_Disinformation_Effectively.pdf; as well as interviews conducted for this report.

[117] Meta has also recently expanded its labelling of AI-generated imagery. Natasha Lomas, "Meta to expand labelling of AI-generated imagery in election packed year," TechCrunch, February 6, 2024, https://techcrunch.com/2024/02/06/meta-ai-generated-image-labelling/.

[118] See, e.g., Jon Roozenbeek, Sander van der Linden, and Thomas Nygren, "Prebunking interventions based on 'inoculation' theory can reduce susceptibility to misinformation across cultures," The Harvard Kennedy School Misinformation Review, January 2020, Volume 1:2, https://doi.org/10.37016/mr-2020-008; Stephan Lewandowsky and Sander van der Linden, "Countering Misinformation and Fake News Through Inoculation and Prebunking," European Review of Social Psychology, February 22, 2021, https://www.tandfonline.com/doi/full/10.1080/10463283.2021.1876983; Nico Grant and Tiffany Hsu, "Google Finds 'Inoculating' People Against Misinformation Helps Blunt its Power," The New York Times, August 24, 2022, https://www.nytimes.com/2022/08/24/technology/google-search-misinformation.html.

[119] Cat Zakrzewski et al., "Debunking Misinformation Failed. Welcome to 'Pre-bunking,'" *Washington Post*, May 26, 2024, https://www.washingtonpost.com/technology/2024/05/26/us-election-misinformation-prebunking/.

[120] See, e.g., Shannon Bond, "False information is everywhere. 'Pre-bunking' tries to head it off early," NPR, October 28, 2022, https://www.npr.org/2022/10/28/1132021770/false-information-is-everywhere-pre-bunking-tries-to-head-it-off-early.

[121] See Shannon Bond, "How Russia is losing — and winning — the information war in Ukraine," NPR, February 28, 2023, https://www.npr.org/2023/02/28/1159712623/how-russia-is-losing-and-winning-the-information-war-in-ukraine.

[122] Vittoria Elliott and David Gilbert, "Elon Musk's Main Tool for Fighting Disinformation on X Is Making the Problem Worse, Insiders Claim," *Wired*, October 17, 2023, https://www.wired.com/story/x-community-notes-disinformation/.

[123] United States Holocaust Memorial Museum, "Bystanders," in *Holocaust Encyclopedia*, https://encyclopedia.ushmm.org/content/en/article/bystanders.

[124] United States Holocaust Memorial Museum, "Bystanders."

[125] US Department of State's Bureau of Conflict and Stabilization Operations and USAID's Center of Excellence on Democracy, "Human Rights and Governance: Atrocity Assessment Framework Supplemental Guidance to State/USAID Conflict Assessment Frameworks" (working draft, n.d.), https://2009-2017.state.gov/documents/organization/241399.pdf.

[126] See, e.g., "What Is Nudge Theory?," Imperial College London, https://www.imperial.ac.uk/nudgeomics/about/what-is-nudge-theory/; Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird & Jevin D. West, "Combining interventions to reduce the spread of viral misinformation," Nature, June 23, 2022, https://www.nature.com/articles/s41562-022-01388-6.

[127] Taylor Hatmaker, "Twitter Tests a Feature That Calls You Out for RTing without Reading the Article," *TechCrunch*, June 10, 2020, https://techcrunch.com/2020/06/10/twitter-retweet-prompt-android/.

[128] Josh Constine, "Instagram Officially Tests Hiding Like Counts," *TechCrunch*, April 30, 2019, https://techcrunch.com/2019/04/30/instagram-hidden-like-counter/; Taylor Hatmaker, "Twitter Runs a Test Prompting Users to Revise 'Harmful' Replies," *TechCrunch*, May 5, 2020, https://techcrunch.com/2020/05/05/twitter-harmful-replies-prompt-harassment-test-feature/.

[129] Rhiannon Neilsen, "Coding protection: 'cyber humanitarian interventions' for preventing mass atrocities," International Affairs 99:1 (2023), p. 312.

[130] Megan A. Brown and Tessa Knight, "Trendless Fluctuation? How Twitter's Ethiopia Interventions May (Not) Have Worked," *Tech Policy.Press*, January 11, 2022, https://www.techpolicy.press/trendless-fluctuation-how-twitters-ethiopia-interventions-may-not-have-worked/. The action, however, was controversial. An investigation by New York University and DFRLab suggested that the intervention resulted in "no discernible change in the volume of tweets or the prevalence of toxic and threatening speech, meaning the Twitter intervention may not have worked as intended."

[131] "U.S.-EU Recommended Actions for Online Platforms on Protecting Human Rights Defenders Online" (March 2024), https://www.state.gov/wp-content/uploads/2024/03/HRD-Guidance_Joint_Updated-_-Accessible-3.12.24.pdf.

[132] Interviewees noted that these can help both improve substantive response efforts and build relationships across sectors in preparation for atrocity events.

[133] See, e.g., "'Video Unavailable': Social Media Platforms Remove Evidence of War Crimes," Human Rights Watch, September 10, 2020, https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes; Lindsay Freeman, "Digitally Disappeared: The Struggle to Preserve Social Media Evidence of Mass Atrocities," Georgetown Journal of International Affairs, Volume 23:1, Spring 2022, pp. 105-113; Alexa Koenig, "Big Tech Can Help Bring War Criminals to Justice," Foreign Affairs, November 11, 2020, https://www.foreignaffairs.com/articles/united-states/2020-11-11/big-tech-can-help-bring-war-criminals-justice.

[134] The EU Digital Services Act provides new obligations for social media companies on data preservation, which may implicate new concerns, including on the misuse of data and implications for affected communities.

[135] Former President Trump's Suspension [by Facebook] Upheld, Oversight Board, 2024, https://www.oversightboard.com/decision/FB-691QAMHJ/.

[136] One participant proposed an intervention under which social media companies could provide an option or "button" for users to click to notify the platform that the content may be useful to save to support accountability initiatives.

[137] Rachel Hilary Brown, *Defusing Hate: A Strategic Guide to Counteract Dangerous Speech* (Washington, DC: United States Holocaust Memorial Museum, 2016), https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf.

[138] See, e.g., U.S. Department of State, "U.S.-EU Recommended Actions for Online Platforms on Protecting Human Rights Defenders Online" (March 2024): https://www.state.gov/wp-content/uploads/2024/03/HRD-Guidance_Joint_Updated-_-Accessible-3.12.24.pdf.

[139] Global Internet Forum to Counter Terrorism, "Advances in Hashing for Counterterrorism," *Insight, News*, March 29, 2023, https://gifct.org/2023/03/29/advances-in-hashing-for-counterterrorism/.

A nonpartisan federal, educational institution, the **UNITED STATES HOLOCAUST MEMORIAL MUSEUM** is America's national memorial to the victims of the Holocaust, dedicated to ensuring the permanence of Holocaust memory, understanding, and relevance. Through the power of Holocaust history, the Museum challenges leaders and individuals worldwide to think critically about their role in society and to confront antisemitism and other forms of hate, prevent genocide, and promote human dignity.

ushmm.org/connect

UNITED STATES
HOLOCAUST
MEMORIAL
MUSEUM

SIMON-SKJODT CENTER
FOR THE PREVENTION OF GENOCIDE

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