Unity State, South Sudan | *Mathieu Rouquette for Mercy Corps*

THE WEAPONIZATION OF SOCIAL MEDIA

How social media can spark violence and what can be done about it

NOVEMBER 2019

Social media has emerged as a powerful tool for communication, connection, community and, unfortunately, conflict. It's created new, highly accessible channels for spreading disinformation, sowing divisiveness and contributing to real-world harm in the form of violence, persecution and exploitation. The impact social media has on real-world communities is complex and rapidly evolving. It stretches across international borders and challenges traditional humanitarian aid, development and peacebuilding models.

This new paradigm requires a new approach.

Mercy Corps has partnered with Do No Digital Harm and Adapt Peacebuilding on a landscape assessment to examine how social media has been used to drive or incite violence and to lay the foundation for effective, collaborative programming and initiatives to respond quickly and help protect already fragile communities.

MERCY CORPS

What makes social media different

Disinformation, hate speech and recruitment to violent groups through social manipulation are not new phenomena. These activities have long been identified as drivers or triggers of conflict and a focus of violence prevention efforts. In the past, these activities happened through traditional media and in-person communication. Social media changes the game. Here's how:

- **Social media platforms increase communication power.** The international reach and ease of access to social media mean that a higher volume of weaponized information can reach more people faster, and via multiple channels.
- **The personalization of social media targets individuals and amplifies impact.** Social media platforms tailor information to individual users' preferences. Machine learning takes that personalization further, serving up more targeted content, and narrowing the scope of information an individual receives to topics and viewpoints that confirm and reinforce one another.
- **Personalization increases polarization and exacerbates conflict risk.** Social media platforms organize users into groups that share preferences and demographic characteristics, creating "bubbles" or "echo chambers" that align with ethnic, ideological, linguistic or other societal divisions. Rumors can seem like credible facts, and collective online outrage can quickly trigger real-world violence.
- **Weaponized information is difficult to police.** Online conversations and the offline actions that flow from them can evolve quickly, and it can be nearly impossible to identify individual wrongdoers among the billions of social media users. Because the same qualities that make social media prone to weaponization also make it a powerful driver of positive engagement, regulations and technology companies' own policies have struggled to isolate and keep pace with threats.

Helmand, Afghanistan | *Toni Greaves for Mercy Corps*

UNDERSTANDING TYPES OF DIGITAL HARM

***Misinformation**: incorrect information spread by people without the intent to deceive*[1]

***Disinformation**: incorrect information spread by people intentionally in order to deceive or manipulate others*[2]

***Hate Speech**: any form of expression (speech, text, images) which "demeans or attacks a person or people as members of a group with shared characteristics such as race, gender, religion, sexual orientation or disability"*[3]

***Dangerous Speech**: speech that has a special capacity to catalyze or amplify violence by one group against another*[4]

[1] Bertolin, Giorgio. "Digital Hydra: Security Implications of False Information Online," (NATO StratCom COE: May 2016). https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online.

[2] *Ibid.*

[3] Robert Faris et al., "Understanding Harmful Speech Online" *Berkman Klein Center for Internet and Society* (2016), 5-6.

[4] Awori, 470; Theo Dolan, et al. "Youth and Radicalization in Mombasa, Kenya: A Lexicon of Violent Extremist Language on Social Media" (PeaceTech Lab: 2018), 6.

How weaponization works

This assessment explores how weaponized social media can contribute to offline conflict by examining real-world case studies. These examples are not exhaustive. Rather, they surface a range of concepts and implications that can help humanitarian, development and peacebuilding organizations — as well as technology companies and policymakers — understand what's happening and develop effective responses.

> *"No technology has been weaponized at such an unprecedented global scale as social media."*
>
> — Jonathan Ong & Jason Cabañes[5]

Case studies

Information operations (IO): Coordinated disinformation campaigns are designed to disrupt decision making, erode social cohesion and delegitimize adversaries in the midst of interstate conflict. IO tactics include intelligence collection on specific targets, development of inciteful and often intentionally false narratives and systematic dissemination across social and traditional channels. The Russian government used such tactics to portray the White Helmets humanitarian organization operating in Syria as a terrorist group, which contributed to violent attacks against the organization.

Political manipulation (PM): Disinformation campaigns can also be used to systematically manipulate political discourse within a state, influencing news reporting, silencing dissent, undermining the integrity of democratic governance and electoral systems, and strengthening the hand of authoritarian regimes. These campaigns play out in three phases: 1) the development of core narratives, 2) onboarding of influencers and fake account operators, and 3) dissemination and amplification on social media. As an example, the president of the Philippines, Rodrigo Duterte, used Facebook to reinforce positive narratives about his campaign, defame opponents and silence critics.

Digital hate speech (DHS): Social media platforms amplify and disseminate hate speech in fragile contexts, creating opportunities for individuals and organized groups to prey on existing fears and grievances. They can embolden violent actors and spark violence — intentionally or sometimes unwittingly. The rapid proliferation of mobile phones and Internet connectivity magnifies the risks of hate speech and accelerates its impacts. Myanmar serves as a tragic example, where incendiary digital hate speech targeting the majority Muslim Rohingya people has been linked to riots and communal violence.

Radicalization & recruitment (RR): The ability to communicate across distances and share user-generated, multimedia content inexpensively and in real time have made social media a channel of choice for some violent extremists and militant organizations, as a means of recruitment, manipulation and coordination. The Islamic State (ISIS) has been particularly successful in capitalizing on the reach and power of digital communication technologies.

---

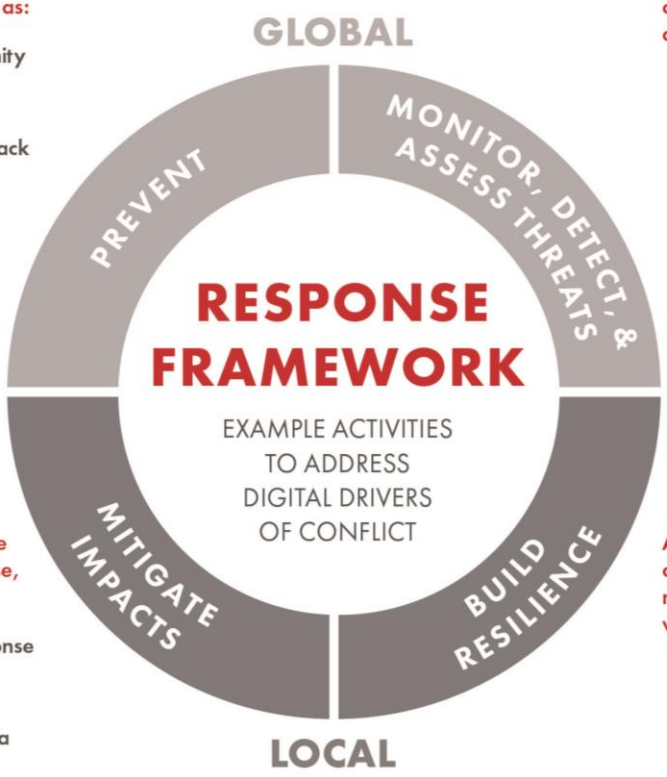
[5] Ong, Jonathan and Jason Cabañes, "Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines" (Newton Tech4Dev Network: 2018). https://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf.

A new framework for response

Based on insights from the case studies, we outline a framework for collective, comprehensive responses to digital drivers of conflict, identifying key entry points in the life cycle of weaponized social media where public, private and nonprofit organizations can make a difference. The framework is illustrated here and described in further detail below.

Front-end activities that reduce the incidence of weaponization, such as:

- Technology company community standards
- Technology product updates (e.g., credibility scores, feedback mechanisms) and inauthentic account removal
- Industry regulations (e.g., transparency, user data protection, accountability mechanisms)
- Civil society advocacy



Activities that detect or make sense of weaponization threats, associated drivers, and their impact, such as:

- Information and threat mapping
- Open-source rumor monitoring and management
- Identification and analysis of online hate speech
- Social network monitoring, analysis, and reporting

Activities that minimize or manage the worst impacts when crises arise, such as:

- Referral or warning and response mechanisms integrated into monitoring systems
- Weaponization of social media crisis response plans
- Addressing and countering polarization and radical narratives

Activities that improve the ability of vulnerable populations to resist adverse impacts from weaponization, such as:

- Digital media literacy training
- General awareness-raising campaigns on weaponization of social media
- Offline social cohesion building activities

Prevention: ***Reducing the incidence of weaponization with activities that include influencing policies and regulations of governments, multinational bodies, industry associations and technology companies.*** For example, the European Union has developed a set of data protection rules that outlines regulations for businesses and organizations in how to process, collect and store individuals' data, establishing rights for citizens and means for redress.[6]

Monitoring, detection and assessment of threats: ***Bringing together a wide variety of stakeholders, from intelligence organizations to civil society organizations, to identify threats and their potential impact.*** In Kenya's Tana Delta, for example, the Sentinel Project's Una Hakika program counters rumors

[6] EU Data Protection Rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

that have contributed to inter-ethnic violence by creating a platform for community members to report, verify and develop strategies to address misinformation.[7]

Building resilience: ***Helping fragile populations resist the worst impacts of the weaponization of social media, with digital media literacy training, online and offline awareness-building and education, and strategies to build social cohesion.*** For example, the Digital Storytelling initiative in Sri Lanka seeks to build skills in citizen storytelling as a way to balance polarizing online rhetoric, while also helping individuals become more responsible consumers of online information.[8] In another example, Mercy Corps' peacebuilding work in Nigeria's Middle Belt has increased trust and perceptions of security across farmer and pastoralist groups while including specific initiatives to support religious and traditional leaders in analyzing and leading discussions aimed at reducing the impacts of hate speech in social media.[9]

Mitigation: ***Minimizing harm once weaponized information has already spread, particularly in times of crisis.*** These activities might take place offline or online and include integrating referral or warning and response components into monitoring systems, establishing crisis and response plans, and addressing and countering online hate speech and radical or violent extremist narratives. An example is the Dangerous Speech Project's Nipe Uwell in Kenya project, which provided public information on dangerous speech as well as mechanisms to report and remove such speech online during the height of electoral tensions.[10]

Collaboration to counter weaponization

Social media has created fertile ground for online misinformation and manipulation that can lead to offline violence. For organizations working in international humanitarian aid, development and peacebuilding, weaponized social media adds complexity to the already difficult work of preventing and responding to violent conflict. Responding effectively to weaponized social media requires building new knowledge, capabilities and partnerships to better understand what's possible, what works and what doesn't.

By working together, aid and development organizations, governments and private sector companies can help make the world safer, responding to the threat of weaponized information on social media with actions and programs that meet the scale and sophistication of the challenge.

Next steps

The response framework outlined here includes a range of possible actions to address weaponized information on social media, drawing from cybersecurity, communications studies, cognitive science, conflict resolution and media studies. Our next step is to pilot and test this response framework in a variety of relevant contexts and, from this, build a working model and playbook for how to combat weaponized information and advance peace.

---

[7] "How It Works: Una Hakika." Sentinel Project. https://thesentinelproject.org/2014/02/17/how-it-works-una-hakika/.

[8] Digital Literacy Project. https://www.linkedin.com/school/digitalstorytelling/about/.

[9] Mercy Corps. https://www.mercycorps.org/research/does-peacebuilding-work-midst-conflict.

[10] Dangerous Speech Project. Nipe Ukweli. https://dangerousspeech.org/nipeukweli/.

CONTACT

MEGHANN RHYNARD-GEIL
Senior Adviser | Technology for Development
mrhynardgeil@mercycorps.org

LISA INKS
Acting Director | Peace and Conflict
links@mercycorps.org

Mercy Corps is a leading global organization powered by the belief that a better world is possible. In disaster, in hardship, in more than 40 countries around the world, we partner to put bold solutions into action — helping people triumph over adversity and build stronger communities from within. Now, and for the future.

Do No Digital Harm is the world's first on-call support mechanism for humanitarian organizations, NGOs, and at risk civil society groups looking to mitigate against the harms resulting from digital surveillance, electronic exploitation, and weaponized information. It provides field-research support, digital risk audits, strategic design workshops, and workflow integration support for a variety of clients.

Adapt Peacebuilding produces knowledge, provides advice, and implements programs that better the practice and policies of peacebuilding, and improve outcomes for people affected by violent conflict. Adapt Peacebuilding advises organizations globally, and implement direct programs in Myanmar and Colombia.

45 SW Ankeny Street
Portland, Oregon 97204
888.842.0842
mercycorps.org