



THE WEAPONIZATION OF SOCIAL MEDIA

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

November 2019



Executive Summary

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.

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.

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.

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 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 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 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.¹

¹ EU Data Protection Rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

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 that have contributed to inter-ethnic violence by creating a platform for community members to report, verify and develop strategies to address misinformation.²

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.³ 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.⁴

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.⁵

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.

² "How It Works: Una Hakika." Sentinel Project. <https://thesentinelproject.org/2014/02/17/how-it-works-una-hakika/>.

³ Digital Literacy Project. <https://www.linkedin.com/school/digitalstorytelling/about/>.

⁴ Mercy Corps. <https://www.mercycorps.org/research/does-peacebuilding-work-midst-conflict>.

⁵ Dangerous Speech Project. Nipe Ukweli. <https://dangerousspeech.org/nipeukweli/>.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This assessment was authored by:

Primary Authors:

Joseph Guay, Director, Research, Do No Digital Harm Initiative

Stephen Gray, Director, Adapt Peacebuilding

Secondary Authors:

Meghann Rhynard-Geil, Senior Advisor, Technology for Development, Mercy Corps

Lisa Inks, Acting Director, Peace and Conflict, Mercy Corps

We are indebted to Jenny Vaughan, Dina Esposito, Alan Donald, and Adrienne Karecki for their guidance throughout this initiative. Special thanks to Sarahna Khatri and Sona Pai. We are grateful to our other reviewers and supporters, including Rebecca Wolfe, Michael Bowers, Sanjay Gurung, Michael Young, Sasha Davis, Chelsea Wieber, Jennifer Schmidt, Matt Streng, Miranda Hurst, Selena Victor, Megan Doherty, Manasi Patwardhan, and Jennifer Seward. We are particularly grateful for the inputs of Faye Mooney, in memoriam.

Table of Contents

Introduction	7
Case Studies	8
Conflict Analysis	16
Responding to Weaponization	20
Conclusion	29
Bibliography	30
Appendix	40
Methodology	40
In-depth Case Studies	42

Introduction



“No technology has been weaponized at such an unprecedented global scale as social media”

— Jonathan Ong & Jason Cabañes (2018)

The so-called Digital Revolution has transformed the humanitarian, development and peacebuilding landscape, creating new pathways for data-driven interventions, along with a broader ecosystem of relevant actors, roles and relationships. Social media platforms, search engines and online news organizations, for example, play an increasingly significant role in elections integrity, civic discourse and group identity formation, with offline impacts on peace and social cohesion.

But while digital technologies can offer many opportunities to improve people’s lives, there is also growing concern around their possible negative implications as drivers of violence, persecution and exploitation. While the spread of malicious or inaccurate information has long been a driver of conflict through in-person communication and traditional media, this landscape assessment examines the ways in which digital platforms and behaviors — specifically on social media — uniquely contribute to conflict and may require peacebuilders to adapt their existing strategies or create new approaches.

Repressive authorities, armed groups and violent extremists are making innovative use of digitally enabled tools and methods to distort facts on the ground and spread incendiary rhetoric. Their goals: to obfuscate accountability, undermine community acceptance, erode social cohesion or incite panic and/or violence. Ordinary citizens become embroiled in these processes, whether intentionally or unconsciously, and contribute to networks of online and offline actions that undermine healthy societies or foment violence.

This assessment, based on a literature review and expert interviews, takes stock of these challenges and proposes ways for peacebuilders and other organizations to respond. It outlines four ways in which social media is weaponized, contributing to conflict dynamics and posing risks for humanitarian, development and peacebuilding processes: information operations, political manipulation, digital hate speech, and radicalization and recruitment. The assessment also analyzes how social media presents new drivers of conflict — and can exacerbate traditional ones — within the broader spectrum of root causes and triggers of violence, and proposes next steps for a comprehensive response to the weaponization of social media.

See Appendix for a description of the assessment methodology.

Case Studies

The following case studies provide snapshots of the contexts in which the weaponization of social media occurs and the specific tactics used. They are not intended to express opinions about who is “good” or “bad” in a specific context, but rather outline how social media has been weaponized and the impact of those activities on peace and stability.

The concepts of misinformation and disinformation appear through each case study, forming the foundation of how social media is often weaponized, and are defined as follows:

- *Misinformation*: incorrect information spread by people without the intent to deceive⁶
- *Disinformation*: incorrect information spread to intentionally deceive or manipulate others⁷, including deliberately false news stories, manufactured protests, doctored content (such as photos or videos), and tampering with private communications before release.



Warrap, Sudan | Miguel Samper for Mercy Corps

See Appendix for in-depth case study examinations and findings.

Case Study 1: Information operations — Russia’s targeting of the White Helmets in Syria

In the digital era, coordinated disinformation operations have re-emerged as a central component of Russia’s information warfare strategy in places like Syria, a country of significant geostrategic importance for Russia that has been plagued by one of the worst refugee crises in modern history. Specifically, the Russian government has made systematic use of information operations to amplify manufactured claims and false accusations against the Syrian Civil Defense, also known as the White Helmets, a Nobel-prize nominated humanitarian organization made up of more than 3,000 volunteers who are credited with saving thousands of lives in Syria.⁸ In the context of armed conflict in Syria, Russia’s government has labeled the White Helmets a terrorist organization with links to al-Qaeda and ISIS.

Weaponization via information operations

Information operations — defined as “the integrated employment ... of information-related capabilities in concert with other lines of operations to influence, disrupt, corrupt, or usurp the decision-making of adversaries” — is a central component of Russia’s Information Warfare strategy.⁹ In such situations, conflict is not declared overtly, and most activities are carried out below the threshold of conventional means.

6 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>.

7 Misinformation can have a powerful effect on divisiveness and chaos in a target society, as the truth becomes hard to discern. (CRS, 5) Civilians may knowingly or unknowingly be functioning as proxies on behalf of an adversary (CRS, 1). See also Bertolin, 5.

8 Di Giovanni, Janine. 16 October 2018. “Why Assad and Russia Target the White Helmets.” The New York Review of Books.

<https://www.nybooks.com/daily/2018/10/16/why-assad-and-russia-target-the-white-helmets/>

9 CRS, Information Warfare, 3 quoting Joint Chiefs of Staff: 3-13, “Information Operations” (November 27, 2012).

Clashes are “contactless,” using precision capabilities that target non-combatants (i.e., civilian populations, news media and/or the private sector).¹⁰

Technique and tactics

While there is some variation in the descriptions of the specific steps taken to implement information operations, a central sequence of practices across these sources makes up the “Digital Disinformation Playbook.”¹¹

- 1) **Targeting:** Propagators of disinformation operations carry out *intelligence collection* on their target audiences via open-source channels on the web and analysis gathered by digital advertising agencies.
- 2) **Content creation:** Operatives create and curate emotionally resonant or otherwise inciteful content (audio/visual, text-based information) for weaponization, including propaganda, misinformation and disinformation.
- 3) **Dissemination:** Narratives are systematically disseminated through multiple means, fusing together social and traditional media, as well as offline channels such as printed materials or public rallies.¹²
- 4) **Amplification:** Propagated narratives are then amplified via botnets, inauthentic accounts, influencers, hashtag hijacking, astroturfing¹³ (imitating grass-roots actions using coordinated inauthentic accounts) and trading up the chain¹⁴ (planting stories in small outlets where they can then be picked up by larger ones).
- 5) **Distraction:** All actors within the system work together to prevent objective sense-making within the target zone of operations by creating distractions, disrupting telecommunications infrastructure or banning social media platforms.

Impacts and implications

The Syria Campaign, with research from Graphika, estimates that “bots and trolls linked to other Russian disinformation campaigns have reached an estimated 56 million people on Twitter with posts related to the White Helmets during ten key news moments of 2016 and 2017.”¹⁵ These online defamation campaigns attempt to delegitimize the White Helmets’ status as a neutral and impartial humanitarian actor in an attempt to make them a legitimate target for kinetic attacks.¹⁶ Over 210 white helmet volunteers have been killed

10 Svetoka, 9-11; Lucas and Pomeranzev, 11.

11 See T.E. Nissen Framework – Svetoka, 11; Facebook “Information Operations,” 2016; Lucas and Pomeranzev, 6; Sarah Oh and Travis Adkins, “Disinformation Toolkit” (InterAction: June 2018), 11; Giorgio Bertolin, “Digital Hydra: Security Implications of False Information Online,” (NATO StratCom COE: May 2016), 8-9; Hossein Derakhshan and Claire Wardle, “Information Disorder: Definitions” in Understanding and Addressing the Disinformation Ecosystem (Annenberg School of Communication: 2017); Livingston, “Contentious Narratives” 2018; CRS 9-10; Lucas and Pomeranzev, “Winning the information War” 2018.

12 While **mixed media information campaigns** use multiple social media channels and website-based platforms to perpetuate and amplify the reach of a single narrative, **cross-media campaigns** leverage a central channel around which the campaign is built and hyperlinked to. Both are extremely effective at masking inauthenticity. See Bertolin, 40-42.

13 Russian IO fuse **astroturfing** (imitating grass-roots actions using coordinated inauthentic accounts) with **hybrid-trolling** (deliberately provocative behavior that aims to distort online discussions and cause conflict among participants in order to advance ideological, political, or military objectives). (Bertolin, 29).

14 According to Marwick and Lewis, “media manipulators are able to trade stories “up the chain” of media outlets...by planting a story with a small or local news outlet who may be too understaffed or financially strained to sufficiently fact-check it. If the story performs well enough...it gets amplified beyond its current scope.” (Marwick and Lewis, 38-39).

15 The Syria Campaign and Graphika, “Killing the Truth: How Russia is Fuelling a Disinformation Campaign to Cover up War Crimes in Syria, 2017, 13.

16 For example, “KARMA IS A BITCH -> #WhiteHelmets killed. That will teach you to kill innocent children to fake #syrichochemicalattack!! #SyriaHoax #MFAnews.” (@BinsakSB). Furthermore “Those at the heart of these conspiracy theories, such as Vanessa Beeley, call for the White Helmets to be killed as legitimate military targets.” See The Syria Campaign, 2017.

since 2013. Their centers “have been hit by missiles, barrel bombs and artillery bombardment 238 times between June 2016 and December 2017.”¹⁷

As a consequence, **the operational capacities of the White Helmets and their partners are eroded**, and the disinformation campaign has a net effect of distracting from or covering up activities by Syrian and Russian forces on the ground, including potential war crimes.

Case Study 2: Political Manipulation — Elections in the Philippines

President Rodrigo Duterte of the Philippines has proven adept at exploiting social media for political gain, leveraging social media to reinforce positive narratives about his campaign and to defame and silence opponents and critics.¹⁸ The Philippines-based online news website Rappler has been the target of coordinated and sustained disinformation campaigns after it exposed the systematic use of paid trolls¹⁹, bots²⁰, networks of fake accounts and contracted influencers propagating pro-Duterte narratives (including mis- and disinformation) during the 2016 presidential election.

Weaponization via political manipulation

Political manipulation is similar to information operations, but within the context of a single community or state. Political discourse is systematically manipulated by networked disinformation campaigns modeled after digital advertising approaches and operationalized through exploitative strategies and incentive structures. These practices have the power to set agendas, propagate ideas, debase political discourse and silence dissent, ultimately seeking to change the outcome of political events.

These disinformation campaigns play out in three phases:

- 1) **Design:** Establishing objectives, branding, core narratives, etc.
- 2) **Mobilization:** Onboarding influencers, fake account operators and grassroots intermediaries, and preparing media channels
- 3) **Implementation:** Disseminating and amplifying messages and implementing other tactics such as digital black ops, #trending and signal scrambling.²¹

Technique and tactics

Political misinformation and coordinated disinformation campaigns feature a range of operatives who design strategies and implement tactics, and are based on an architecture involving:

¹⁷ The Syria Campaign, 2017.

¹⁸ Paladino, 16-17.

¹⁹ In internet slang, trolls refers to people who post inflammatory content online in order to cause argument or harass an individual or organization, either for their own amusement, or for another form of gain. Trolls are also associated with the presentation of extraneous information that sows or normalizes tangential conversations or narratives.

²⁰ An internet bot, also known as a web robot, robot or simply bot, is a software application that runs automated tasks (scripts) over the internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone.

²¹ Much like in the Information Operations case study 1, these political messaging campaigns also make use of #hashtag hijacking, astroturfing, and trading up the chain tactics.

- 1) **Employing PR strategists and creatives:** Elite strategists use marketing techniques to align campaign objectives with consistent messaging through “branding” and employ locally informed creative writers, who “weaponize popular vernaculars to maximize the reach of social media posts.”²²
- 2) **Leveraging digital influencers:** Anonymous influencers and key opinion leaders (celebrities, pundits, etc.) commanding between 50,000 and 2 million followers) capitalize on popular culture trends and disseminate manufactured narratives through Twitter (via trending rankings) and Facebook.²³
- 3) **Amplifying through community-level fake operators:** Sub-contracted workers amplify messaging and localize narratives using pre-drafted, script-based messaging, predetermined schedules for media blasting and click strategies.
- 4) **Engaging grassroots intermediaries:** Fan page moderators, unpaid volunteers and members of political organizations drive real grassroots engagement with disinformation by manufacturing “illusions of engagement”.

Impacts and implications

Despite efforts by Rappler and others to shore up free speech and provide counter-narratives to the Duterte propaganda machine, Duterte continues to use misinformation and coordinated disinformation campaigns to obfuscate controversial policies and practices, such as the war on drugs. The Philippines is now teeming with fake news, and other political agents are adopting, adapting and scaling up the digital disinformation model, which thrives on an unregulated digital advertising industry, exploits the mechanisms of social media to control narratives, and harms not only journalists and political opponents but also local workers used as active agents of disinformation.

Case Study 3: Digital Hate Speech — Intercommunal Violence in Myanmar

Social media platforms can act to amplify **hateful, dangerous speech** in fragile contexts, where rumors, misinformation and disinformation can play a role in inciting intercommunal, electoral or other forms of violence. Extra-factual sources of information contribute to this problem and are often amplified by social media. Digital hate speech has driven anti-Muslim sentiment in Myanmar and been directly implicated in fomenting intercommunal violence.

Weaponization via digital hate speech

While hate or dangerous speech has traditionally been propagated by traditional media such as radio or television or through in-person gatherings, the rapid proliferation of mobile phones and internet connectivity and the inherent technological and psychological features of social media platforms magnify these risks.²⁴ In today’s digital environment, every individual has the capacity and agency to develop, disseminate and


²² Ong and Cabañes, 45.

²³ Ibid, 34.

²⁴ Svetoka, 5-6; Brandon Paladino, “Democracy Disconnected: Social Media’s Caustic Influence on Southeast Asia’s Fragile Republics” (Brookings Institute: 2018), 7-8; Eran Fraenkel, “A Critical Analysis of Digital Communications and Conflict Dynamics in Vulnerable Societies” (Internews: 2014), 2, 10-11; Nils Weidmann, “Communication, Technology, and Political Conflict: Introduction” *Journal of Peace Research* (2015)264; Deen Freelon, “Personalized Information Environments and Their Potential Consequences for Disinformation” in *Understanding and Addressing the Disinformation Ecosystem* (Annenberg School of Communication: 2017), 38, 60; Norman Vasu, et al, “Fake News: National Security in the Post-Truth Era” (RSIS: 2018),10-11; Gregory Asmolov, “The Disconnective Power of Disinformation Campaigns,” “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018), 32

consume potentially fabricated or misleading information on digital platforms with the power to increase communication speed, volume (of output and input), variety (of content), reach and coverage. Social media amplifies hate at scale. Finally, the inherent design of social media begets selective exposure, information bubbles, homogeneous echo-chambers, confirmation bias, and hyper-personalized, hyper-sensory and hyper-insular information environments that reduce our cognitive capacity to objectively evaluate information.²⁵

Digital hate speech and virulent rumors warrant a unique aggregation of environmental factors, malicious strategies, and inadvertent actions in a logical narrative to know when thresholds for violent conflict are reached. In considering the patterns, conditions, features and drivers above, Mercy Corps, Do No Digital Harm and Peacebuilding have developed a theory of harm:



When certain underlying conditions are present, we would expect social media to have an amplifying effect on conventional conflict dynamics. In situations of security-related anxiety, rumors – especially if they conform to pre-existing worldviews and emotionally-relevant narratives, and especially if audiences are repeatedly exposed to them – can perpetuate unfounded threat claims, amplify in-group/out-group tensions, and motivate rational actors to engage and justify collective violence in the name of self-defense.²⁶

Techniques and tactics

Digital communication amplifies conflict dynamics through the following **extra-factual sources of information**²⁷:

- 1) **Rumor**: Unverified information that is transmitted from one person to others. Rumors can be true, false or a mixture. At their core, mis- and disinformation are rumors.²⁸
- 2) **Hate speech**: Any form of expression (speech, text, images) that “demeans or attacks a person or people as members of a group with shared characteristics such as race, gender, religion, sexual orientation, or disability.”²⁹
- 3) **Dangerous speech**: Speech that has a special capacity to catalyze or amplify violence by one group against another.³⁰

25 Padalino; 7-8; Fraenkel 10-11; Freelon, 38.

26 Greenhill and Oppenheim, 2-3; Benesch, 2014 3, 21; Padalino, 9.

27 According to Greenhill and Oppenheim, extra-factual sources of information are (a) unverified at the time of transmission, (b) serve as a source of actionable knowledge, (c) intended to influence recipients’ attitudes or behavior, (d) are emotionally resonant, and (e) are framed in ways that fit pre-existing societal narratives. See Kelly Greenhill and Ben Oppenheim, “Rumor Has It: The Adoption of Unverified Information in Conflict Zones,” in *International Studies Quarterly* (2017), 2

28 See John Bugge, “Rumour Has It: A Practice Guide to Working with Rumours:” (Communicating with Disaster Affected Communities (CDAC): 2017), 8; and Greenhill and Oppenheim, 2. Importantly, a rumor can also take on multiple forms over time: “For example, a human trafficker can spread a rumor amongst refugees ... with the intent to deceive (disinformation), and a refugee can then pass this rumor to his friends and family not intending to deceive them (misinformation).” (Bugge, 8).

29 See Robert Faris et al., “Understanding Harmful Speech Online” Berkman Klein Center for Internet and Society (2016), 5-6. See also Kagonya Awori and Susan Benesch, “Umati: Kenyan Online Discourse to Catalyze and Counter Violence” (Conference Paper: IFIP 2013), 470.

30 Awori, 470; Theo Dolan et al, “Youth and Radicalization in Mombasa, Kenya: A Lexicon of Violent Extremist Language on Social Media” (PeaceTech: 2018), 6. This kind of speech is predicated upon the risk of violence (e.g. instilling fear by warning of impending threats, or by making an incitement to violence).

Susan Benesch's (2014) *Dangerous Speech Guidelines*³¹ include a range of contextual factors and series of hallmarks that collectively estimate the capacity of speech to inspire violence. Dangerousness can be estimated according to the following five factors:

- 1) A powerful speaker with a high degree of influence over an audience most likely to react
- 2) An audience with grievances and/or fears that the speaker can cultivate
- 3) A speech act understood by the audience as a call to violence
- 4) A social or historical context propitious for violence
- 5) An influential means of dissemination

Impacts and implications

In Myanmar, both misinformation in the form of **organic rumors** and speculation, and deliberate **disinformation** have played a significant role in amplifying grievances and triggering violence between groups of differing ethnic and religious identities. Anti-Muslim sentiment and intercommunal violence against Muslim identity groups have been the most visible examples and are linked to the country's deep Buddhist nationalist project.

Buddhist nationalists such as the 969 movement and Ma Ba Tha have exploited social media (particularly Facebook) "to stoke fear, normalize hateful views and facilitate actors of violence" against identity groups (particularly Muslims, or the ethnic Rohingya) who are perceived and promoted as enemies of Buddhism, or of the State.^{32,33} Research has documented narratives of Muslim people (particularly the Rohingya) as illegal immigrants, terrorists and rapists, among other fabricated and incendiary messaging, reflecting the hallmarks of Benesch's *Dangerous Speech*.³⁴

C4ADS documents how virulent rumors and online hate speech triggered the Mandalay riot of July 2014, in which approximately 20 people were injured and two people were killed.³⁵ More recently, the Myanmar military has carried out systematic clearance operations in 2017 against the Rohingya people in response to the Arakan Rohingya Salvation Army attacks, another situation that was largely amplified by digital hate speech and the propagation of unverified rumors.³⁶ Hundreds of thousands of Rohingya have fled to Bangladesh as a result, with many reports documenting systematic rape by security forces and affiliated militia groups, and over 6,000 civilian deaths. United Nations officials and human-rights organizations have characterized the Rakhine State security operations as ethnic cleansing.

31 Benesch, 2014, 7-8.

32 C4ADS, "Sticks and Stones: Hate Speech Narratives and Facilitators in Myanmar" (2016)

33 Fink writes, "The speakers are highly regarded. The society has struggled with mistrust and violence, and Facebook has become the primary medium of communication. By creating and disseminating images of adversaries through the mass media—and in Myanmar's case, social media—a group can generate widespread support for the idea that such adversaries cannot remain..." Christina Fink, "Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar" in "Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms" *Journal of International Affairs* Vol. 71, No. 1.5 (2018). Fink also points out the emotionally resonant language used, the widespread reach of Facebook, and lack of space and resources for critically accessing information (i.e. weak media institutions).

34 C4ADS, 2016.

35 C4ADS, 11.

36 According to Paladino, "In the wake of the ARSA attacks, the Facebook group for Ma Ba Tha supporters registered a significant spike in anti-Rohingya messages, illustrating the powerful tendency of such online associations to amplify and reinforce the thoughts of its individual members." (Paladino, 9).

Case Study 4: Radicalization and Recruitment — the ISIS media jihad

Digital propaganda is a central aspect of the Islamic State’s approach to contemporary jihad. ISIS leaders go so far as to elevate the production and dissemination of propaganda *as a form of worship*.³⁷ In doing so, ISIS has been particularly successful in exploiting defining elements of social media — peer-to-peer communication, near real-time user-generated content and low-cost dissemination of multimedia content — to spread extremist propaganda; target, manipulate, and seek to recruit supporters; and coordinate tactical operations.

Weaponization via radicalization and recruitment

ISIS has rebranded the image of radical extremism through messages of inclusiveness and belonging — many of which are disseminated online. An initial focus on gaining currency among extremist and fringe demographics has been replaced by a broader approach to appeal to wider audiences.³⁸

The affordances of social media allow for near-instantaneous access to emotionally resonant narratives, reduce costs associated with participation in formal organizations, offer a relatively risk-free entry point for potential recruits to find like-minded individuals, and create a social environment in which extremist views are normalized. In short, “social media provides cheaper and more accessible pathways to radicalization.”³⁹ These features of the contemporary information landscape are perfectly aligned with the franchised nature of modern terrorist organizations and operations.⁴⁰



Helmand, Afghanistan | Toni Greaves for Mercy Corps

Techniques and Tactics

ISIS adapts conventional recruiting techniques to the digital information environment, allowing for more targeted campaigning, more emotionally resonant messaging, and more personalized exchanges between operatives and potential recruits. Specific tactics include:

- 1) **Targeting younger, tech-savvy millennials who feel isolated from society.** They often lack a strong sense of identity or purpose, and are frustrated with their economic, familial or interpersonal situations.⁴¹

37 Winter, 11, 17.

38 Koerner, 2016.

39 Zeitzoff, 9.

40 Theohary and Rollins, 4.

41 See Dylan Gerstel, “ISIS and Innovative Propaganda: Confronting Extremism in the Digital Age,” *Swarthmore International Relations Journal* Issue 1 (2017) pg. 2; and Lydia Wilson, “Understanding the Appeal of ISIS” *New England Journal of Public Policy* Vol 29 No. 1 (March 2017), pg. 8.

- 2) **Highlighting themes of openness, inclusion and participation.**⁴² Through coordinated messaging campaigns, ISIS recruits are shown a sense of purpose, collective identity and meaning.⁴³
- 3) **Moving recruits toward radicalization one step at a time.** Exposure to one set of ideas can open the door for other, more radical thinking to take root.
- 4) **Mobilizing and coordinating operations.** Social media are also leveraged for mobilizing supporters, sharing logistical and training information, and coordinating tactical operations.

Impacts and implications

While ISIS is not the first militant group to use social media for information activities and gaining support, their use of social media is distinct in three ways:

- 1) **The focus on propaganda elevates the status, importance and role of social media users,** including online volunteers, influencers, film makers, graphic designers and other technical specialists within the greater ISIS network.⁴⁴
- 2) **ISIS has set the standard for strategic communications innovation among violent extremist networks,** with activities such as developing bespoke software and its own mobile application for disseminating and amplifying its propaganda.
- 3) **ISIS is adaptive and persistent, despite coordinated efforts among technology companies, governments and civil society to counter them.** When suspected accounts are de-platformed, blocked users come back online using alternative handles, and operatives have become skilled at undermining detection efforts and ensuring the group's digital survival.

While the exact contribution of social media in increasing recruitment performance is unclear, ISIS' innovative use of social media, along with its adaptability in seeking to maximize the use of social media to create harms, sets a dangerous precedent for other violent political organizations.

42 See Charlie Winter, "Media Jihad: The Islamic State's Doctrine for Information Warfare," International Centre for the Study of Radicalization and Political Violence (ICSR), (2017), pg. 15-16.

43 Gerstel, 2.

44 Winter, 12; Koerner, 2016.

Conflict Analysis

The weaponization of social media drives conflict in powerful ways that intersect with and exacerbate existing issues in specific contexts. For individuals, social media’s amplification power and highly targeted, personalized nature can exploit fundamental cognitive processes to implant dangerous information and influence susceptible people with greater efficiency than other means of communication. On a broader scale, these same qualities can polarize groups of people and lend credibility to rumors, further dividing and inciting violence between groups at risk of conflict. These drivers are evolving quickly — faster than traditional approaches to addressing information-borne threats.

The following examples outline the various relationships between weaponization phenomena and conflict typologies. Their diversity illustrates the need for a detailed analysis in each context that takes stock of types of conflict, root causes, profiles of the actors involved and the role of weaponization.

Conflict Typology ⁴⁵	Potential roles of social media weaponization	Example
Interstate warfare: a conflict between two or more governments	As part of a conventional warfare strategy, social media can provide a vector for state-supported information operations, including coordinated disinformation campaigns designed to influence public perception, confuse adversaries and weaken their internal relationships.	Russia’s coordinated disinformation campaigns in Ukraine ⁴⁶
Civil war/ state-formation conflict: a conflict between a government and non-government party	Non-state actors use social media platforms to organize violent opposition against states, publicize their causes and secure resources from international allies. Online discourse is partitioned along ideological, identity or linguistic lines, generating echo chambers that maintain stereotypes, polarization and grievance.	The role of social media in the initiation and maintenance of Syria’s civil war ⁴⁷
Popular protest (and riots): popular demonstrations, often involving a spontaneous action by unorganized, unaffiliated members of society	Social media platforms have enabled citizens to mobilize more effectively in opposition to oppressive regimes. This mobilization, even if it is nonviolent in intention, can bring populations into violent conflict with the security organs of the state, who may target opponents identified online or shut down the internet.	The use of social media to organize popular protests in Egypt in 2010-2011
Intercommunal conflict: violence between non-state groups that occur along shared identities.	Hate speech or dangerous speech purposefully or unintentionally generated by members of one identity group is spread and amplified on social media, driving animosity and triggering violence against members of another identity group.	Hate speech and intercommunal violence in Myanmar
Electoral Violence: violence used by political operatives and supporters to achieve their	Social media platforms amplify political competition between constituencies divided along ideological or identity lines, in some cases inciting hatred and violence.	Electoral violence in Kenya since 2008

45 Conflict types rarely have universally accepted definitions. The descriptions here have been drawn from the Uppsala Conflict Data Program (UCDP), the Armed Conflict Location and Event Data (ACLED) Project, the United States Institute of Peace (USIP), and www.understandingconflict.org

46 Boyte, K.J. (2017) An analysis of the social-media technology, tactics, and narratives used to control perception in the propaganda war over Ukraine. *Journal of Information Warfare*. 16:1

47 Lynch, M., Freelon, D., Aday, S. (2014) Syria’s socially mediated civil war. United States Institute of Peace. Available: <https://www.files.ethz.ch/isn/176084/PW91-Syrias%20Socially%20Mediated%20Civil%20War.pdf>

Violent Extremism: violence used to achieve political, ideological, religious or social goals

Social media platforms are used as international publicity tools for spreading extremist propaganda; targeting, recruiting and radicalizing potential supporters to extremist causes; and for coordinating tactical operations among geographically dispersed operatives.

The recruitment of Western and non-Western combatants by the Islamic State

Existing conflict drivers

Weaponization phenomena intersect with a diverse range of societal predispositions to conflict, increasing tensions and the risk of violence. Root causes of conflict (also called structural causes or underlying causes) are context-specific, long-term or systemic causes of violent conflict that have become built into the norms, structures and policies of a society.⁴⁸

Different types of root causes likely present varying degrees of susceptibility to malicious or inaccurate information on social media, but further research on whether some causes (i.e. identity-based and attitudinal/normative) are particularly susceptible to exacerbation by weaponized social media would be useful. Constructivist perspectives, particularly concerning the social construction of identity and the roles of language, norms, knowledge and symbols in the initiation and maintenance of conflict, might be useful in analyzing the influence of weaponized social media in more depth.⁴⁹

Proximate digital drivers of conflict

Weaponization of social media produces digital proximate drivers of conflict,⁵⁰ or triggers, with distinct causal relationships and the potential to influence more people across broader geographies and with immediate impact. These drivers are relatively short-term catalysts that accentuate a conflict's underlying causes and promote the escalation of violence.⁵¹ The following descriptions demonstrate how social media weaponization acts as an amplifier, speeding up or increasing the potential for conflict, beyond what traditional media forms have had the power to do.

Social media platforms increase multiple dimensions of communication power. The international reach and ease of access of information communication technologies create much larger potential audiences for malicious and/or inaccurate information compared with legacy communication technologies. Online, this information can reach large segments of a population faster, enabling activists to organize violent or non-violent protests in real time. Violent extremists using social media platforms can easily spread messages internationally, drawing new recruits to their causes, or generating fear and sowing discord in other societies. Non-state armed groups or those engaged in intercommunal violence or armed conflicts against the state can more easily popularize their activities globally and attract support from foreign sympathizers, which can complicate and sustain internal conflicts.

48 Herbert, Siân (2017) Conflict Analysis. Governance and Social Development Resource Center. Available: <https://gsdrc.org/topic-guides/conflict-analysis/core-elements/>

49 Jackson, Richard (2009). Constructivism and conflict resolution, in "The Sage Handbook of Conflict Resolution," eds. Bercovitch, J., Kremenyuk, V., and Zartman, W. London: Sage Publications.

50 So-called conflict triggers (single events that can rapidly change the intensity or direction of violent conflict) are not presented in detail in this analysis, though the speed and reach of social media platforms, and the fertile ground that they provide for rumours (especially in times of crisis) strongly supports the activating effect of trigger events.

51 Ibid (GSDRC).

Personalization of social media content improves targeting. Social media platforms provide information tailored to each user's preferences. Users personalize their social media experience across platforms by selecting who or what they follow. Machine learning takes that personalization further, serving up content based on individuals' previous choices and social media metadata, which helps content producers more accurately tailor and target content. It's necessary for revenue from both advertising and data mining, which have made huge investments to set up and make social media platforms commercially viable.⁵²

However, personalization technologies also allow malicious information to spread via social media to influence people efficiently, which can lead to attitudinal and behavior changes conducive to conflict. This happens via various individual cognitive processes that are susceptible to exploitation. For example:

- *The primacy effect:* Users at a formative state with respect to an issue, such as young people and violent extremism, may form conclusive opinions on the basis of the information that is first acquired, regardless of its veracity.⁵³
- *The illusory truth effect:* Because of personalization, users that demonstrate an interest in violent extremism are more likely to see additional content that promotes radicalization. When people are exposed and re-exposed to information, they tend to believe that it is more truthful because they cannot remember the original source.⁵⁴
- *The availability heuristic and confirmation bias:* On social media, individuals' preferences and algorithmic targeting will expose them to content that further confirms rather than discounts their views.⁵⁵

Increasing polarization exacerbates conflict risk (especially in vulnerable societies). Though the internet can theoretically "democratize information" by connecting people across geographic and identity divisions, the personalization of social media platforms (and even search algorithms) instead functions to parse users' information access into relatively uniform groups that share preferences and demographic characteristics. When there is dissonance between people online, cutting ties via unfriending or unfollowing others is relatively easy compared with real-world social networks, encouraging further segregation.

The resulting "social media bubbles" or "echo chambers" have implications for conflict, particularly in settings with volatile root causes. In these contexts, discourse, attitudes, norms and "facts" begin sorting along ethnic, ideological, linguistic or other societal cleavages. Individuated social media news streams provide user groups with content and *social proof* that justifies their attitudes, beliefs and prejudices, fueling disagreement and eroding the shared language and knowledge through which people could otherwise find common ground.⁵⁶ *Identity protective cognition* provides one explanation for why groups polarize, and describes a form of reasoning in which people selectively credit and dismiss factual information to protect their status within an affinity group.⁵⁷

Social media fills the trust gap and intensifies rumor dynamics. The conflict-driving effect of personalized information is exacerbated by the low levels of trust that are prevalent in settings with histories

52 Mark Hachman, "The Price of Free: How Apple, Facebook, Microsoft and Google Sell You to Advertisers," PCWorld, October 1, 2015, <https://www.peworld.com/article/2986988/privacy/the-price-of-free-how-apple-facebook-microsoft-and-google-sell-you-to-advertisers.html>.

53 Vasu Et Al. (2018) Fake News: National Security In The Post-truth Era. S. Rajaratnam School Of International Studies: Policy Report.

54 Ibid.

55 Heshmat, S. (2015) What is confirmation bias? Psychology Today. Available: <https://www.psychologytoday.com/us/blog/science-choice/201504/what-is-confirmation-bias>

56 Allcott, H. and Gentzkow, M. (201) "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives 31, no. 2: 230.

57 NATO (2017) Digital Hydra: Security Implications Of False Information Online. Available: <https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online>

of violent conflict. Low intergroup trust is more commonly found when groups have a history of competition or violence. Low trust in institutions is linked to a history of state governance failures, and particularly relevant to the weaponization of social media, when independent media has been restricted by the state.⁵⁸ When these types of trust are low, and when alternative sources of verifiable truth are less readily available, people tend to rely upon and believe in information that is provided to them by people on social media who share friendship, kinship, ethnic or religious social ties.⁵⁹ This phenomenon exacerbates the creation and implications of echo chambers, while limiting the veracity of information on social media. Inaccurate information can be highly inflammatory from a conflict perspective, particularly when tensions are running high.

Social media provides a fertile breeding ground for the spread of rumors, a potent example of the conflict risks of inaccurate information spread between friends. Rumors tend to be particularly pervasive in threatening and ambiguous situations, such as periods of social unrest, riot or war. Rumors operate as a kind of collective problem solving in the face of uncertainty, but tend to overemphasize threatening information that promotes defensive or retaliatory actions against perceived villains.⁶⁰ These types of information have been shown to resonate with others online, distorting and hardening potentially incendiary perceptions of threat (including who is deemed to be the cause).

Speed and decentralization make weaponized information of social media difficult to police. Social media communication is a highly decentralized person-to-person technology, with billions of users continuously producing content every hour of every day. The swarm-like nature of social media allows for conversations to evolve very quickly compared with legacy media, propelled by a high number of one-to-many interactions, in which it is very hard to identify and respond to risks promptly or hold any one entity accountable for the nature or impact of the conversation. The conflict risks posed by social media discourse have proven difficult to regulate against, in part because they have evolved faster than government regulators can understand them, and also because of the complexities (both technological and contextual) involved in recognizing contextually specific harmful content, assigning responsibility for its dissemination, and determining the best pathway for addressing it. There are also inherent challenges with relying on the social media platforms to regulate themselves, and ambiguity around how they approach it. These factors combine to limit traditional mechanisms that detect, assess and seek responses to the risks of harmful information, such as watchdog organizations, media associations, public regulators, investigative branches of government and concerned citizens.

58 Pew Research Center (2017). Spring 2017 Global Attitudes Survey. <http://www.pewglobal.org/2018/01/11/publics-globally-want-unbiased-news-coverage-but-are-divided-on-whether-their-news-media-deliver/>

59 NATO (2017) Digital Hydra: Security Implications Of False Information Online. Available: <https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online>

60 Difonzo, N. (2010) Ferreting Facts or Fashioning Fallacies? Factors in Rumor Accuracy. *Social and Personality Psychology Compass*. 4(11):1124 - 1137.

Responding to Weaponization

The disruptive nature of weaponized social media demands a response from a broad array of actors, including governments, technology companies, media organizations and other private companies, and also international humanitarian, development and peacebuilding organizations — the focus of this assessment— that are affected by these phenomena or have mandates to design and implement programming in response.

A new response framework: Phases and entry points for programming

Digital weaponization phenomena challenge the way organizations develop responses to problems in international humanitarian, development and peacebuilding domains. In particular, these phenomena add a layer of complexity to traditional conflict drivers and the ways in which peacebuilding and violence prevention efforts seek to address those drivers. The actors and pathways of weaponized social media are cross-sectoral, trans-national and evolving at rates that outpace the international system's current response. Effective alternatives require organizations willing to navigate new terrain.

The **response framework** introduced below is based on analysis of how weaponization phenomena typically unfold and metastasize in different phases over time. The framework alludes to unique response entry points that correspond, in some cases, with how public, private and practitioner organizations are already responding. In reality, each case study of weaponization does not conform to a consistent lifecycle, and some activities are ongoing rather than sequential. This is a model and as such, it is imperfect and inexact (i.e., the categories aren't mutually exclusive and the examples aren't exhaustive.)

The growing and multidisciplinary field of responding to the weaponization of social media in the context of violence prevention still requires more evidence to demonstrate the effectiveness or relevance of certain strategies or activities, as well as:

- A common understanding of the collective solutions, pathways, resources and capacities needed for this work
- Shared criteria or indicators to evaluate the effectiveness of various approaches
- A comprehensive understanding of the various disciplines (e.g., cybersecurity, communications studies, cognitive and behavioral science, network theory) that can contribute to a comprehensive response framework

Examples shared within the framework simply illustrate some of the efforts undertaken to date and help clarify the complementary purposes of each response category.

Prevention

Activities of **prevention** are intended to reduce the incidence of weaponization. These include regulations developed and enforced by governments, multinational bodies or industry associations, such as legislation or regulations concerning transparency, user data protection and accountability mechanisms, as well as punitive efforts that might deter future harms. 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 and establishes rights for citizens and means for redress.⁶¹

Prevention also includes the policies and technical initiatives of tech companies that impact the prevalence of weaponization on social media platforms, such as the development of company community standards, product updates that promote feedback mechanisms or the improvement of practices to remove inauthentic accounts. Civil society-led advocacy activities to influence either regulations or technology company practices also contribute to prevention. Many of these activities and the relationships required to promote prevention are relatively novel for international humanitarian, development and peacebuilding organizations.

Monitoring, detection, and assessment of threats

A wide variety of stakeholders, from intelligence organizations to civil society activists, play a role in **threat monitoring, detection and assessment**. However, these activities have not traditionally been a central concern of most international humanitarian, development and peacebuilding organizations. Activities under this category include information and threat mapping, the development of open-source rumor monitoring and management systems, identification and analysis of online hate speech, and social network monitoring, analysis and reporting.

An example of creating and facilitating a misinformation monitoring and management system is the Sentinel Project's Una Hakika program in Kenya's Tana Delta. The program was designed to counter rumors that contributed to inter-ethnic conflict by creating a platform for community members to report, verify and develop strategies to address misinformation.⁶² Using another strategy, PeaceTech Lab's Hate Speech Lexicon workstream seeks to identify and analyze online hate speech in specific contexts.⁶³ In South Sudan, for example, PeaceTech Lab worked with local citizens and experts over time to develop a series of monitoring reports that outlined hateful speech and provided predictions for future conflict based on online hate speech.

Building resilience to threats

Building resilience, or increasing the ability of societies to resist weaponized social media's worst impacts, includes both online and offline responses and aligns with traditional aspects of peacebuilding and conflict-sensitive development or humanitarian action. Strategies for building resilience to the impacts of social media include training in digital media literacy, general awareness-raising campaigns on the ways in which social media can be weaponized and general social cohesion building. The Digital Storytelling initiative in Sri Lanka is an example of a combined approach that seeks to build skills in citizen storytelling to balance some of the polarized online rhetoric, while also increasing digital literacy within communities to be more responsible consumers of online information.⁶⁴

Peacebuilding organizations have developed a range of tools and strategies to build social cohesion within and between communities in or at risk of conflict, as well as between communities and government

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

62 "How It Works: Una Hakika." Sentinel Project. <https://thesentinelproject.org/2014/02/17/how-it-works-una-hakika/>

63 Combating Hate Speech. PeaceTech Lab. <https://www.peacetechlab.org/combating-online-hate-speech-main>

64 Digital Literacy Project. <https://www.linkedin.com/school/digitalstorytelling/about/>

institutions, increasing trust and improving relationships, and strengthening the social contract. For 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.⁶⁷ In another example, Search for Common Ground's Social Cohesion Framework for Myanmar presents a range of activities to address conflict drivers, including those on social media, to build stronger communities.⁶⁸

Mitigation of impacts

Mitigation occurs when weaponized information has already manifested and seeks to minimize harm, particularly during crises. Like building resilience, mitigation activities might take place offline or online — either traditional and agnostic to the technology of weaponization, or as some new form of activity made possible by new technology. Types of responses include integrating referral or warning and response components into monitoring systems described above, establishing weaponization of social media crisis and response plans, and addressing and countering prevalent online polarization and hate speech, as well as radical or violent extremist narratives, through online or offline means. Organizations have produced numerous methodologies and toolkits for mitigating the impacts of weaponized social media to guide responses by security actors and non-governmental actors.⁶⁹

An example is Interaction's Disinformation Toolkit, which among other strategies, outlines potential responses for organizations to counter online rumors and their advantages and disadvantages.⁷⁰ In an example of mitigating the impact of weaponized social media on conflict-affected communities, the Dangerous Speech Project's Nipe Uwell in Kenya project provided public information on dangerous speech

RESPONDING TO DIGITAL DRIVERS OF VIOLENT EXTREMISM IN JORDAN

Mercy Corps' programming in Jordan sought to build resilience among youth to cyber crimes and online violent extremist recruitment by supporting youth coaches to build digital literacy among adolescents and other youth. Youth participants then led initiatives to raise awareness among their peers about internet safety in hotspot areas.⁶⁵ In another program, youth civic engagement activities included youth-led workshops on propaganda and media and supported youth to create films about peace, tolerance and positive role models to mitigate impacts of online recruitment.⁶⁶ Throughout this programming, mothers and fathers of adolescents and youth participated in awareness-raising sessions so they could better understand and support youth's ability to resist negative online messaging.

65 Programs include Youth Advancement for Peaceful and Productive Tomorrow, funded by the European Commission in 2016-2018 and its follow-on in 2018 - present funded by the US Department of State.

66 This program is Nubader: Advancing Adolescents and Youth in Jordan, funded by the Canada Global Affairs Commission from 2016 - present.

67 Mercy Corps' tested peacebuilding work in Nigeria, the USAID-funded Engaging Communities for Peace in Nigeria program, for example, helped increase trust, perceptions of security and positive interactions between farmers and herders in conflict. <https://www.mercycorps.org/research/does-peacebuilding-work-midst-conflict> This program complements Mercy Corps and others' programs that seek to build trust while also supporting communities to understand and address growing concerns around online hate speech in Nigeria's Middle Belt.

68 Social Cohesion Framework: Social Cohesion for Stronger Communities. Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2017/02/SC2_Framework-copy.pdf

69 Examples include Build Up's The Commons: A Pilot Methodology for Addressing Polarization Online: <https://howtobuildup.org/wp-content/uploads/2016/04/The-Commons-A-pilot-methodology-for-addressing-polarization-online-2-27-18.pdf>; the Social Media Hate Speech Mitigation Field Guide from #defyhatenow: <https://openculture.agency/launching-the-social-media-hate-speech-mitigation-field-guide/>; and the Nigeria Stability and Reconciliation Programme's How-to Guide: Mitigating Dangerous Speech: <http://www.nsrp-nigeria.org/wp-content/uploads/2017/12/NSRP-How-to-Guide-Mitigating-Hate-and-Dangerous-Speech.pdf>.

70 Oh, Sarah, and Travis L. Adkins. 2018. "Disinformation Toolkit." Interaction. <https://www.interaction.org/documents/disinformation-toolkit/>

as well as mechanisms to report and remove such speech from online platforms in the height of electoral tensions.⁷¹

The following graphic outlines specific activities that could fall within each phase of the response framework.

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 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 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 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



Initial insights about responses

Exploring the sequence in which weaponized social media evolves and how organizations have been responding leads to several conclusions.

Mitigation is an urgent need in the short term, both in response to new organizational risks borne of weaponization, as well as threats of weaponized social media to populations of concern. Security and reputational hazards abound from weaponization and yet most global NGOs do not have policies, frameworks, protocols or response plans in place to respond appropriately and systematically. Until preventative measures are better developed, organizations in the peacebuilding space will be increasingly expected to respond to online threats to social cohesion or violence triggers.

Prevention work stretches the knowledge capacities and relationships of humanitarian, development and peacebuilding organizations. Prevention work involves legal, regulatory and oversight work (with

⁷¹ Dangerous Speech Project. Nipe Ukweli. <https://dangerousspeech.org/nipeukweli/>

governmental stakeholders), or product interface development (with the technology industry) — which take place primarily in Silicon Valley, Geneva, Brussels and New York. Engaging with prevention work challenges the traditional reach of humanitarian, development and peacebuilding organizations and calls for engagement with relatively new stakeholders, not least of all the technology industry. That said, there remains value in peacebuilding organizations continuing to implement traditional activities with established partners, including advocacy, awareness raising and training, and conflict mitigation and social cohesion activities.

Effective response requires integration, bridging domains and working at local and global levels. The problems of weaponized social media stem partly from the disconnect between how the value of technology platforms was originally conceived in the West, and how their functions actually manifest in the spectrum of global contexts. The architects of these platforms, or the personalization algorithms by which they were monetized, never intended adverse outcomes, yet they are susceptible to hijacking for nefarious ends. Responses to weaponized social media need to mirror these transnational and local-



Iraq | Nigel Downes for Mercy Corps

to-global relationships. There is value in a “brokerage” role between the NGO/civil society and private sectors to bridge the gap between detection and assessment (a specialty of the private sector) and resilience building and mitigation work (strengths of NGOs and civil society). There is also a need for bridging knowledge gaps across geographies. For example, organizations like Mercy Corps often have a longstanding presence, deep contextual knowledge and relationships in places where weaponized social media is causing problems, whereas technologies and relevant policies are developed in relatively insular and mostly Western locations (e.g., Silicon Valley). Influential stakeholders — technology companies chief among them — are often unaware of the extent of the harm they are creating. Given the slow pace of government in the immediate and dynamic evolution of these problems, humanitarian, development and peacebuilding organizations can play important roles in filling the knowledge gaps in private companies and encouraging more accountability and responsiveness to weaponization threats.

Implications for humanitarian, development and peacebuilding organizations

Weaponized social media has a range of implications that might drive organizations to develop programmatic responses.

Impact on communities at risk of or experiencing crisis

Weaponization phenomena directly or indirectly impact the populations that humanitarian, development and peacebuilding organizations serve, by causing harm in a variety of forms, including:

- **Physical:** when an individual or group is targeted by violent action
- **Psychological:** when an individual or group is harassed, intimidated or exploited
- **Social:** when an individual or group is ostracized or defamed by their community

These phenomena might intersect with peacebuilding or conflict mitigation priorities. Weaponization issues also increase exposure to other risks and vulnerabilities. Examples include if vulnerable displaced populations in need of humanitarian assistance are inundated with intentionally misleading information about life-saving services and resources, or if weaponization-related attacks lead to negative coping mechanisms such as self-isolation.

Operational Implications

Weaponization phenomena also affect civil society organizations' ability to effectively implement programs across a range of work areas, such as:

- when misinformation erodes situational awareness
- damaging critical systems or relationships that are vital for communicating or for coordinating resource distribution
- diverting attention and/or resources

Development work to improve the functioning of state institutions, for example, can be undermined by political disinformation campaigns that sow discord about the impartiality of state institutions.

The prevalence of misinformation on social media has significant implications for refugees and asylum seekers, placing them at increased risk of violence and exploitation. Impacts on the viability of existing humanitarian programming due to the evolution of these threats has prompted some organizations to integrate social media strategies and platforms alongside traditional humanitarian programming.⁷²

Security implications

Humanitarian, development and peacebuilding organizations have been targets of coordinated disinformation operations and other kinds of social-media related information attacks that have significant security implications in highly fragile contexts. Research in the context of Syria provides an example whereby coordinated and deliberate online defamation campaigns significantly delegitimize the White Helmets' status in an attempt to make them a legitimate target for kinetic attacks and undermine their capacities to serve affected populations.

Although security examples that have proven disastrous for humanitarian, development and peacebuilding organizations are rare, many have faced at least minor adverse impacts by unintentional misinformation, or an intentionally defamatory social media campaign. This warrants that organizations focus on the importance of social media threats vis a vis conventional security and strategic communications verticals. Interaction's [Disinformation Toolkit](#) (2018), CDAC Network's [Rumor Has It](#) report (2017), and Tactical Tech's [Holistic Security Manual](#) are excellent strategic resources for thinking about context analysis, risk assessment and threat modeling approaches in this regard.

Exposure to unintended harm

Populations affected by complex humanitarian emergencies and situations of armed conflict are particularly vulnerable to digital threats and risks, as refugees, besieged populations and other marginalized communities increasingly rely on modern information systems and digital platforms to meet their basic needs. Coupled with a concerted effort in the humanitarian sector to deploy ever more information-based

⁷² See for example www.refugee.info

and data-driven services at scale — yet without requisite standards, tools and capacities for risk management, data ethics and digital security — such efforts may unintentionally expose affected populations to additional risk.

Weaponization phenomena therefore raise very serious questions regarding the Tech4Good, Open Data, Humanitarian Innovation and Digital Transformation agendas, among others, within the sector. If we do not address our own failure to account for the range of emerging negative externalities associated with digital data-driven interventions, we risk turning ourselves into threat actors and eroding the very principles that provide the foundation for our work.

Improving programmatic responses to weaponized social media

This assessment identifies a range of needs and approaches to enhance the quality of programmatic responses to weaponized social media.

- 1) A working theory of harm must be developed in order to prevent, mitigate or counter weaponization phenomena.** Since January 2019, Mercy Corps and The Do No Digital Harm Initiative have been working together to develop such a theory, based on systems approaches to information environments. We have learned that for matters regarding the weaponization of information — whether malicious and systematic disinformation operations, or the inadvertent propagation of viral rumors — the propensity of harmful or otherwise violent outcomes is a function of the dynamic interplay of a handful of environmental factors that characterize either robust and resilient information ecosystems on one end of the spectrum or weak, volatile, asymmetric or hostile information ecosystems on the other. The latter have a latent capacity for harm when misinformation, disinformation, propaganda and the like are introduced. These environmental factors include:

- **Foundations for digital harm:** The environmental conditions that make a particular context more (or less) susceptible to digital harms.⁷³
- **Pathways to digital harm:** The ways in which such harms unfold in a particular context, whether intentionally (in a malicious way) or unintentionally (in an organic way), or a combination of the two.⁷⁴
- **Signals of digital harm:** Indicators (direct or indirect), symptoms, or early warning signs that alert us that weaponization of social media issues are actively unfolding in a particular context.

To effectively address weaponization of social media at scale, any intervention must take into account the dynamic, interdependent and nonlinear nature of this system — and the confluence of these factors.

⁷³ This includes, for example, lack of reliable information, high levels of ambient fear, a history of group grievance, distrust of local authorities to provide justice and/or security, etc.

⁷⁴ For example, intentionally means that there are malicious actors actively deploying strategies for leveraging information to cause harm. The strategies and tactics used by Russia's Internet Research Agency (IRA), cyber-operations divisions, and affiliated cyber-mercenary groups to deceive, distort, and disrupt information environments in targeted contexts is one such example. Unintentionally means that there are dynamics around the flow of information that exacerbate or amplify the propensity for harm, whether or not there are coordinated campaigns. For example, in Ethiopia, while social media fill a need for critical information in the face of limited press freedoms, periodic internet blackouts, and restriction of mobile data, in recent years, these platforms have inadvertently amplified rumors, misinformation, and dangerous speech that have played a significant role in promoting radicalization and extremism, fomenting ethnic tensions, and inciting violence during a tenuous period of transition and reform.

2) **Organizations must commit to building local capacities for digital resilience** as a means by which to prevent, mitigate and counter the most urgent and harmful effects of weaponized social media.⁷⁵ All actions within the response framework must be designed to support the digital resilience of affected communities as a cornerstone of the humanitarian, development and peacebuilding initiatives. Our efforts must resolve the tensions and trade-offs between remote-based, technologically driven capabilities (for example, those afforded by today’s digitally driven, multisourced, hyper-permeable information landscape) and the *actual* localization of protection, prevention and conflict resolution capacities of affected communities.

3) **There is a need for much improved evidence of what does and does not achieve desired impacts.** It is not clear, for example, if innovation processes such as hackathons are or are not producing technological solutions that are effective in addressing weaponization phenomena. There is a lack of evidence if programming responses such as media literacy training or counter-speech are achieving impacts that match the scale of the problem. The same question arises regarding the effectiveness of systems for reporting and mitigating electoral manipulation and violence. Response organizations pursuing further programs should invest more in and carefully design monitoring and evaluation of these programs, to establish proof of concept and credibility in the eyes of donors and potential partners.



Unity State, South Sudan | Mathieu Rouquette for Mercy Corps

4) **Lessons for countering weaponized social media can be drawn from the factors that make the problem itself so virulent.** It’s clear, for example, that countering malicious and/or inaccurate information requires maximizing the *saliency* of information alternatives among target populations. Saliency (and ultimately uptake and practical usage) of information has increased by crafting content with human dimensions and emotional appeal, based on human-centered design principles. In the humanitarian space, for example, employing content creators and advisers that were refugees themselves supports the creation of content that is timely, linguistically accurate and *culturally relevant*. Facebook targeting by location, language and interests allows information to be delivered to the right people in the right place in a *timely* manner. Having dedicated moderators that allow for two-way communication, ensures that social media strategies for peace, development or humanitarian action are more *responsive* to individual concerns in a timely manner.

5) **Addressing the weaponization of social media should build on existing good peacebuilding practice where possible,** while recognizing that the unique challenges wrought by social media will require new and creative approaches. The spread of malicious or inaccurate information in general is not a new phenomenon — it has been used frequently to incite or stoke conflict using traditional media. Social media-related drivers, as with all drivers in conflict contexts, must be analyzed and framed within the complex web of underlying and proximate conflict causes. Peacebuilders should

⁷⁵ As InterNews notes, because “information is vital to community resilience,”—defined as “the capacity of individuals, communities, and systems to survive, adapt, grow, and transform in the face of change, stress, shocks, and disruption”—we can posit that “a community with strong information ecosystem is a more resilient one.”

examine their existing toolkit of responses for addressing those conflict drivers and determine how those responses need to be adapted to the changing realities presented by social media specifically. As noted above, digital drivers will not always necessitate digital solutions. However, in some cases, due to the qualitatively different dynamics that social media introduces to conflict environments — including the introduction of unconventional and non-local actors, the difficulty of detecting digital harms that are far “upstream” to violent conflict, and the ubiquitous access to disseminate and receive information, among others — peacebuilders will need to design and test groundbreaking approaches. These will require new types of engagement with private sector actors and new ways of thinking, working and learning.

- 6) Humanitarian, development and peacebuilding organizations engaging with weaponization problems will likely require an unconventional skill mix across teams to design and implement programs that take advantage of social media’s incredible power to reach and influence people. Teams sometimes include lawyers, journalists, refugees, creative content producers and coders, alongside more traditional civil society practitioner roles.

The **program development pathway will likely build on and amplify current trends in traditional programming**, including concepts and calls to “never stop iterating,” conceiving of interventions as “test kitchens” that must “adapt and innovate” and achieve “proof of concept” before they can be “taken to scale”. Humanitarian, development and peacebuilding organizations have much to learn from their private-sector counterparts about how to respond more quickly and effectively to dynamic socio-technological processes, and will again need a different mix of organization capacities and types of internal processes.

Conclusion

The weaponization of social media is a highly disruptive set of socio-technological phenomena that cut across various domains of knowledge. The act of producing this assessment was itself a highly interdisciplinary endeavor, drawing upon people and knowledge from diverse academic, practitioner and policy domains. The complexity of related themes, both for understanding the nature of problems and configuring response, is compounded by heterogeneity in how these issues surface at varying levels, from headquarters to the field (i.e. community, operational and programmatic, strategic and coordination, and legal/regulatory and policy). Perspectives, questions, dependencies and entry points abound at each.

The disruptive nature of weaponization has significant implications for how organizations respond. On a macro level, the novelty and influence of these phenomena exposes new players, as well as various inabilities and mismatches in the network of rules and organizations that we would traditionally rely upon to respond to these risks. In this space, tech companies have emerged as entities with much influence, for example, but have not been regulated (or do not self-regulate) commensurate with the scale of harm that their platforms can generate. Governments and institutions of media find themselves struggling to keep pace with the speed at which these technologies evolve. Civil society organizations working with affected populations must consider new partnerships, employees, programs and funding sources to respond effectively.

This assessment has presented a cycle of different response categories, means by which organizations can determine the relevance of these phenomena to their organization, and some preliminary findings about the organizational and programmatic needs for effective response. While further research and testing are needed to advance work in this space, the risks and opportunities that the weaponization of social media presents create an urgent need for collective critical analysis and creative solutions for promoting online and offline peace.

BIBLIOGRAPHY

We have compiled a number of hyper-linked resources (journal articles, case studies, reports, evaluations, policy papers, investigative news articles, and other online resources) we feel would be helpful for peace-building, humanitarian protection, and development practitioners. These resources can be used for introductions into key issues that characterize “weaponization of social media” phenomena in contexts of concern to international NGOs, UN agencies, and international organizations (IOs). They represent a broad landscape of practitioners, organizations, and initiatives working on these issues, across multiple domains and disciplines perhaps not familiar to peace-building and humanitarian practitioners.

Information Operations and Hybrid Warfare (with focus on Syria and Ukraine)

Asmolov, Gregory. “The Disconnective Power of Disinformation Campaigns,” “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018). Available [here](#).

Brooking, Emerson T., and P.W. Singer. “War Goes Viral: How social media is being weaponized across the world.” *The Atlantic* (November 2016). Available [here](#).

Calabresi, Massimo. “Inside Russia’s Social Media War on America.” *Time Magazine* (May 2017). Available [here](#).

CB Insights. “Memes that Kill: The Future of Information Warfare.” *Research Briefs* (May 2018). Available [here](#).

Congressional Research Service (CRS). *Information Warfare: Issues for Congress* (March, 2018) Available [here](#).

Dias, Nic. “The Era of Whatsapp Propaganda Is Upon Us: The future of fake news is messaging apps, not social media. And it’s going to be even worse.” *Foreign Policy* (August 2017). Available [here](#).

DFRLab staff. “#SyriaHoax, Part Two: Kremlin Targets White Helmets.” *Digital Forensics Research Lab* (February 2018). Available [here](#).

DiResta, Renee; Dr. Kris Shaffer; Becky Ruppel; David Sullivan; Robert Matney; Ryan Fox; Jonathan Albright; and Ben Johnson. “The Tactics & Tropes of the Internet Research Agency,” *New Knowledge* (December 2018). Available [here](#).

DiResta, Renee; John Little; Jonathan Morgan; Lisa Maria Neudert; and Ben Nimmo. “The Bots That Are Changing Politics: A taxonomy of politibots, a swelling force in global elections that cannot be ignored.” *MOTHERBOARD* (November 2017). Available [here](#).

Earle, Sam. “Trolls, Bots and Fake News: The Mysterious World of Social Media Manipulation.” *Newsweek* (October 2017). Available [here](#).

Legatum Institute “From how to start a revolution to how to beat ISIS.” Beyond Propaganda series. (November 2015). Available [here](#).

Lucas, Edward, and Peter Pomeranzev. “Winning the Information War: Techniques and Counter-strategies to Russian Propaganda in Central and Eastern Europe” (CEPA: 2016). Available [here](#).

Giles, Keir. “Countering Russian Information Operations in the Age of Social Media” *Council on Foreign Relations*. (November 2017). Available [here](#).

Bertolin, Giorgio. “Digital Hydra: Security Implications of False Information Online,” (NATO StratCom COE: May 2016) Available [here](#).

Golovchenko, Yevgeniy; Mareike Hartmann; and Rebecca Adler-Nissen. “State, Media and Civil Society in the Information Warfare over Ukraine: Citizen Curators of Digital Disinformation,” *International Affairs*, Volume 94, Issue 5, September 2018. Available [here](#). Graphika. “Killing the Truth: How Russia is fueling a disinformation campaign to cover up war crimes in Syria.” The Syria Campaign (2017) Available [here](#).

Greenberg, Andy. “Russian Hacker False Flags Work – Even After They’re Exposed.” *WIRED Magazine*. (February 2018). Available [here](#).

Greenberg, Andy. “Russian Hackers are Using ‘Tainted’ Leaks to Sow Disinformation.” *WIRED Magazine*. (May 2017). Available [here](#).

Gregory, Sam. “Deepfakes and Synthetic Media: Survey of Solutions against Malicious Usages.” *WITNESS* (2018). Available [here](#).

Haines, John R. “Russia’s Use of Disinformation in the Ukraine Conflict,” *Foreign Policy Research Institute*, (February, 2015). Available [here](#).

Johnson, Eric. “Tech is now a weapon for propaganda and the problem is way bigger than Russia.” *ReCode* (January 2018). Available [here](#).

Kreko, Peter. “The Authoritarian Capture of Social Media.” *Power 3.0* (November 2017). Available [here](#).

Lange-Ionatamishvili, Elina, and Sanda Svetoka. “Strategic Communications and Social Media in the Russia Ukraine Conflict,” NATO Cooperative Cyber Defence Centre of Excellent Tallin, Estonia (2015). Available [here](#).

Lapowsky, Issie. “Inside the Research Lab Teaching Facebook about its Trolls,” *WIRED Magazine* (August 2018). Available [here](#).

Meserole, Chris, and Alina Polyakova. “Disinformation Wars: The United States and Europe are ill-prepared for the coming wave of “deep fakes” that artificial intelligence could unleash.” *Foreign Policy* (May 2018). Available [here](#).

Morgan, Jonathan, and Renee DiResta. “Information Operations are a Cybersecurity Problem: Toward a New Strategic Paradigm to Combat Disinformation.” *Just Security* (July 2018). Available [here](#).

Nissen, Thomas Elkjer. #TheWeaponizationOfSocialMedia: @Characteristics_of_Contemporary_Conflicts. Royal Danish Defense College (Copenhagen: March 2015) Available [here](#).

Patrikarakos, David. "Ukraine: Dissident Capabilities in the Cyber Age," in *Cyber Propaganda*

PBS Frontline "The Facebook Dilemma" (2018). Available [here](#).

Pesenti, Marina, and Peter Pomerantsev. "How to Stop Disinformation: Lessons from Ukraine for the Wider World," *Beyond Propaganda* series. Legatum Institute (August, 2016). Available [here](#).

Polakow-Suransky, Sasha. "For Whom the Cell Trolls," *Foreign Policy* (February 2018). Available [here](#).

Pomerantsev, Peter. "Brave New War: A new form of conflict emerged in 2015—from the Islamic State to the South China Sea." *The Atlantic*. (2015) Available [here](#).

Pomerantsev, Peter. "Russia and the Menace of Unreality: How Vladimir Putin is revolutionizing information warfare." *The Atlantic*. (September 2014). Available [here](#).

Priest, Dana; James Jacoby; and Anya Bourg. "Russian Disinformation on Facebook Targeted Ukraine Well Before the 2016 U.S. Election." FRONTLINE PBS. (October 2018). Available [here](#).

Singer, Peter, and Emerson T. Brooking. "The Machines that Will Fight the Social Media Wars of Tomorrow." *Gizmodo*. (October 2018). Available [here](#).

Solon, Olivia. "How Syria's White Helmets became victims of an online propaganda machine." *The Guardian* (December 2017). Available [here](#).

Snegovaya, Maria. "Putin's Information Warfare in Ukraine: Soviet Origins of Russia's Hybrid Warfare" *Institute for the Study of War*. (September 2015). Available [here](#).

Starbird, Kate. "Content Sharing within the Alternative Media Echo-System: The Case of the White Helmets" *Medium*. (May 2018). Available [here](#).

Svetoka, Sandra. "Social Media as a Tool of Hybrid Warfare," (NATO StratCom COE: May 2016) Available [here](#).

Van Niekerk, Brett, and Manoj Maharaj. "Social Media and Information Conflict" *International Journal of Communication* Vol. 7 (2013), 1162–1184. Available [here](#).

Wehner, Markus. "Hacking, propaganda and electoral manipulation: Moscow's information war on the West." *Eurozine*. (July 2017). Available [here](#).

Digital Hate Speech, Rumors, and Violent Conflict (with focus on Myanmar)

Awori, Kagonya, and Susan Benesch. “Umati: Kenyan Online Discourse to Catalyze and Counter Violence” (Conference Paper: IFIP 2013) Available [here](#).

Bailard, Catie Snow. “Ethnic conflict goes mobile: Mobile technology’s effect on the opportunities and motivations for violent collective action,” in *Journal of Peace Research* Vol. 52(3) (2015) 323–337. Available [here](#).

Benesch, Susan. “Countering Dangerous Speech: New Ideas for Genocide Prevention,” (Dangerous Speech Project: 2014) Available [here](#).

Benesch et al, “Web of Hate: Tackling Hateful Speech in Online Social Spaces” *Berkman Klein Center for Internet and Society* (2016) Available [here](#).

Benesch, Susan; Derek Ruths; Kelly P. Dillon; Haji Mohammad Saleem; and Lucas Wright. “Counterspeech on Twitter: A Field Study,” Kanishka Project (2016). Available [here](#).

Bugge, John. “Rumour Has It: A Practice Guide to Working with Rumours:” (Communicating with Disaster Affected Communities (CDAC): 2017) Available [here](#).

C4ADS, “Sticks and Stones: Hate Speech Narratives and Facilitators in Myanmar” (2016) Available [here](#).

Clark, Doug Bock. “Myanmar’s Internet Disrupted Society – and Fueled Extremists,” WIRED Magazine (September 2017). Available [here](#).

Douek, Evelyn. “Facebook’s Role in the Genocide in Myanmar: New Reporting Complicates the Narrative” (October 2018) Available [here](#).

Faris, Robert, et al. “Understanding Harmful Speech Online” *Berkman Klein Center for Internet and Society* (2016) Available [here](#).

Fink, Christina. “Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar” in “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018). Available [here](#).

Fraenkel, Eran. “A Critical Analysis of Digital Communications and Conflict Dynamics in Vulnerable Societies” (Internews: 2014) Available [here](#).

Greenhill, Kelly, and Ben Oppenheim. “Rumor Has It: The Adoption of Unverified Information in Conflict Zones,” in *International Studies Quarterly* (2017) Available [here](#).

Hogan, Libby, and Michael Safi. “Revealed: Facebook Hate Speech Exploded in Myanmar During Rohingya Crisis” *The Guardian* (April 2018) Available [here](#).

Kiersons, Steven. “New Paradigms of Violence: Are Burma’s Rohingya Facing a Hybrid Genocide?” The Sentinel Project (2015) Available [here](#).

Mozer, Paul. “A Genocide Incited on Facebook, with Posts from Myanmar’s Military,” (October 2018) Available [here](#).

Reventlow, Nani Jansen; Jonathon Penney; Susan Benesch; Urs Gasser, et al. “Perspectives on Harmful Speech Online: a Collection of Essays” Berkman Klein Center for Internet & Society Research Publication (2017). Available [here](#).

Sellars, Andrew F. “Defining Hate Speech,” The Berkman Klein Center for Internet & Society Research Publication Series. Research Publication No. 2016-20 (December 2016). Available [here](#).

Stecklow, Steve. “Hatebook: Inside Facebook’s Myanmar Operation” *Reuters Special Report* (August 2018). Available [here](#).

Weidmann, Nils. “Communication, Technology, and Political Conflict: Introduction” *Journal of Peace Research* (2015) Available [here](#).

Computational Propaganda and Political Polarization (with focus on Philippines)

Acker, Amelia. “Data Craft: The Manipulation of Social Media Metadata” Data & Society (2018). Available [here](#).

Alba, Davey. “Connecting Hate: How Duterte Used Facebook to Fuel the Drug War” *Buzzfeed News* (September 2018) Available [here](#).

Bradshaw, Samantha, and Philip N. Howard, “Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation” Computational Propaganda Research Project Working Paper No. 12 (2017). Available [here](#).

Caplan, Robyn; Lauren Hanson; and Joan Donovan, “Dead Reckoning: Navigating Content Moderation After ‘Fake News’” Data & Society (February 2018). Available [here](#).

Derakhshan, Hossein and Claire Wardle. “Information Disorder: Definitions” in *Understanding and Addressing the Disinformation Ecosystem* (Annenberg School of Communication: 2017) Available [here](#).

Derakhshan, Hossein and Claire Wardle. “Information Disorder: Toward an interdisciplinary framework for research and policy making.” Council of Europe report (September 2017). Available [here](#).

Etter, Lauren. “What Happens When the Government Uses Facebook as a Weapon?” *Bloomberg Businessweek* (December 2017) Available [here](#).

Freelon, Dean. “Personalized Information Environments and Their Potential Consequences for Disinformation” in *Understanding and Addressing the Disinformation Ecosystem* (Annenberg School of Communication: 2017) Available [here](#).

Gabriel, Mariya. "A multi-dimensional approach to disinformation report of the independent High Level Group on fake news and online disinformation." European Commission, Directorate-General for Communication Networks, Content and Technology (March 2018). Available [here](#).

Ghosh, Dipayan, and Ben Scott. "#DigitalDeceit: The Technologies Behind Precision Propaganda on the Internet (New America Foundation: 2018) Available [here](#).

Ghosh, Dipayan, and Ben Scott. "Digital Deceit II: A Policy Agenda for Fight Disinformation on the Internet" (New America Foundation: 2018). Available [here](#).

Gregory, Sam. "Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening." Conference Report. WITNESS & FIRST DRAFT NEWS (July 2018). Available [here](#).

Hofileña, Chay F. "Facebook Accounts, Manufactured Reality on Social Media" (October 2016) Available [here](#).

Ireton, Cherilyn, and Julie Posetti. "Journalism, 'Fake News' & Disinformation: Handbook for Journalism Education and Training." UNESCO Series on Journalism Education. (2018). Available [here](#).

Lazer, David; Matthew Baum; Yochai Benkler; Adam Berinsky; Kelly Greenhill; Filippo Menczer; Miriam Metzger; Brendan Nyhan; Gordon Pennycook; David Rothschild; Michael Schudson; Steven Sloman; Cass Sunstein; Emily Thorson; Duncan Watts; and Jonathan Zittrain, "The Science of Fake News: Addressing fake news requires a multidisciplinary effort," in *Science* VOL 359 ISSUE 6380 (March 2018). Available [here](#).

Lewis, Paul. "Fiction is outperforming reality': how Youtube's algorithm distorts truth." *The Guardian* (February 2018). Available [here](#).

Livingston, Steven. "Forward" in "Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms" *Journal of International Affairs* Vol. 71, No. 1.5 (2018) Available [here](#).

Marwick, Alice and Rebecca Lewis, "Media Manipulation and Disinformation Online," (Data & Society: 2017) Available [here](#).

Nadler, Anthony; Matthew Crain; and Joan Donovan. "Weaponizing the Digital Influence Machine: the Political Perils of Online Ad Tech," Data & Society (October 2018). Available [here](#).

National Endowment for Democracy, "Issue Brief: Distinguishing Disinformation from Propaganda, Misinformation, and Fake News," (October 2017). Available [here](#).

Newton, Casey. "How autocratic governments use Facebook against their own citizens" *The Verge* (September 2018) Available [here](#).

Oh, Sarah and Travis Adkins. "Disinformation Toolkit" (InterAction: June 2018) Available [here](#).

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) Available [here](#).

Paladino, Brandon. “Democracy Disconnected: Social Media’s Caustic Influence on Southeast Asia’s Fragile Republics” (Brookings Institute: 2018) Available [here](#).

Ressa, Maria A. “How Facebook’s Algorithms Impact Democracy” (October 2016) Available [here](#).

Ressa, Maria A. “Propaganda War: Weaponizing the Internet” *Rappler*. (October 2016) Available [here](#).

Tucker, Joshua A.; Andrew Guess; Pablo Barbera; Cristian Vaccari; Alexandra Siegel; Sergey Sanovich; Denis Stukal; and Brendan Nyhan. “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.” Hewlett Foundation. (March 2018). Available [here](#).

UCSB Center for Information, Technology, & Society (CITS). “A Citizen's Guide to Fake News” (2018). Available [here](#).

Vasu, Norman, et al. “Fake News: National Security in the Post-Truth Era” (RSIS: 2018) Available [here](#).

Walther, Joseph B. “The Merger of Mass and Interpersonal Communication via New Media: Integrating Metaconstructs,” in *Human Communication Research* 43 (2017) 559–572. Available [here](#).

Woolley, Samuel and Philip N. Howard. “Computational Propaganda Worldwide: Executive Summary” Computational Propaganda Research Project, (2017). Available [here](#).

Online Radicalization and Violent Extremism (with focus on ISIL/Daesh)

Alarid, Maeghin. “Recruitment and Radicalization: The Role of Social Media and New Technology,” *Center for Complex Operations*. (Washington DC: 2016). Available [here](#).

Alexander, Audrey. “How to Fight ISIS Online” *Foreign Affairs* (April, 2017) Available [here](#).

Alexander, Audrey. “How to Fight ISIS Online” *Foreign Affairs* (March 2017) Available [here](#).

Alkhouri, Laith and Alex Kassirer, “Tech for Jihad: Dissecting Jihadists’ Digital Toolbox” (Flashpoint: July 2016)

Alva, Seraphin; Divina Frau-Meigs; and Ghayda Hassan. “Youth and Violent Extremism on Social Media: Mapping the Research” United Nations Educational, Scientific and Cultural Organization (UNESCO: 2017). Available [here](#).

Berger, J.M. “The Toxic Mix of Extremism and Social Media” *Intel Wire* (September 2016) Available [here](#).

Callimachi, Rukmini. “Not ‘Lone Wolves’ After All: How ISIS Guides World’s Terror Plots From Afar” *The New York Times* (February 2017) Available [here](#).

Cobo, Juan Cristóbal. “A Physicist Who Models ISIS and the Alt Right,” *Quanta Magazine* (August 2017) Available [here](#).

Cohen, Adi. "Volunteer Anti-ISIS Fighters Join Up On Facebook" *Vocativ* (2016) Available [here](#).

Cohen, Jared. "Digital Counterinsurgency." *Foreign Affairs*, 94(6), (2015) Available [here](#).

Conway, Maura. "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research," in *Studies in Conflict & Terrorism* 40:1, 77-98 (2017) Available [here](#).

Dastmalchi, Jilla. "The Rise and Fall of IS Propaganda Machine," *BBC News* (February 2017) Available [here](#).

DiResta, Renee. "How ISIS and Russia Won Friends and Manufactured Crowds," *Wired Magazine* (March 2018) Available [here](#).

Dolan, Theo, et al. "Youth and Radicalization in Mombasa, Kenya: A Lexicon of Violent Extremist Language on Social Media" (PeaceTech: 2018) Available [here](#).

Elledge, Katrina. "From Mullahs to Moscow: Propaganda in the Social Media Age," in *Cyber Propaganda*

Cyber Propaganda: From how to start a revolution to how to beat ISIS. Beyond Propaganda series. Legatum Institute (November 2015). Available [here](#).

Gates, Scott, and Sukanya Podder. "Social Media, Recruitment, Allegiance and the Islamic State." *Terrorism Research Initiative*, Vol 9, No 4 (2015) Available [here](#).

Gerstel, Dylan. "ISIS and Innovative Propaganda: Confronting Extremism in the Digital Age," *Swarthmore International Relations Journal* Issue 1 (2017) Available [here](#).

Johnson, Neil. "The Secret Behind Online ISIS Recruitment" *The New Republic* (June 2016) Available [here](#).

Journal of World Affairs, 20(2), (2014) Available [here](#).

Knoll, David. "How ISIS Endures by Innovating" *Foreign Affairs* (2016) Available [here](#).

Koehler, Daniel. "The Radical Online: Individual Radicalization Processes and the Role of the Internet." *The Journal For Deradicalization* (Winter 2014). Available [here](#).

Koener, Brendan. "Why ISIS is Winning the Social Media War" *Wired Magazine* (April 2016) Available [here](#).

Koerner, Brendan. "#jihad: Why ISIS Is Winning the Social Media War" *WIRED Magazine* (March 2016) Available [here](#).

Martins, Ralph. "Anonymous' Cyberwar Against ISIS and the Asymmetrical Nature of Cyber Conflicts," *The Cyber Defense Review*. (Fall 2017). Available [here](#).

Menkhaus, Ken. "Al-Shabaab and Social Media: A Double-Edged Sword." *Brown Journal of World Affairs*, 20(2), (2014) Available [here](#).

Miller, Greg, and Souad Mekhennet. "Inside the Surreal World of the Islamic State's Propaganda Machine" *The Washington Post* (November 2015) Available [here](#).

Paul, Christopher; Colin P. Clarke; Michael Schwille; Jakub P. Hlávka; Michael A. Brown; Steven S. Davenport; Isaac R. Porche III; and Joel Harding. Chapter 10 "Al-Qaeda" in *Lessons from Others for Future U.S. Army Operations in and Through the Information Environment: CASE STUDIES*, Rand Corporation (2018). Available [here](#).

Paul, Christopher; Colin P. Clarke; Michael Schwille; Jakub P. Hlávka; Michael A. Brown; Steven S. Davenport; Isaac R. Porche III; and Joel Harding. Chapter 11 "ISIL/Daesh" in *Lessons from Others for Future U.S. Army Operations in and Through the Information Environment: CASE STUDIES*, Rand Corporation (2018). Available [here](#).

Pauwels, Lieven; Fabienne Brion; Nele Schils; Julianne Laffineur; Antoinette Verhage et al. *Explaining and Understanding the Role of Exposure to New Social Media on Violent Extremism: an Integrative Quantitative and Qualitative Approach*. Academia Press. (Gent; 2014) Available [here](#).

Pearson, Elizabeth. "Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media." *Studies in Conflict & Terrorism* 41:11 (2017). Available [here](#).

Ramsay, Gilbert. "Consuming the jihad: An enquiry into the subculture of Internet jihadism." University of St. Andrews (2011) Available [here](#).

Shultz, David. "Predicting #Terrorism with Social Media" *Pacific Standard* (June 2016) Available [here](#).

Silke, Andrew. "The Internet & Terrorist Radicalization: The Psychological Dimension." *Terrorism and the Internet: Threats – Target Groups – Deradicalization Strategies* (Vol. 67, pp. 27-39). Available [here](#).

Theohary, Catherine, and John Rollins. "Terrorist Use of the Internet: Information Operations in Cyberspace" (Congressional Research Service (CRS): March, 2011) Available [here](#).

University of Miami. "Analyzing how ISIS recruits through social media: Researchers apply the laws of physics to study how terrorist support groups grow online, and how law enforcement can track activities." *ScienceDaily*. (June 2016) Available [here](#).

Vidino, Lorenzo, and Seamus Hughes, "ISIS in America: From ReTweets to Raqqa" (Program on Extremism GWU: December 2015) Available [here](#).

Von Behr, Ines; Anais Reding; Charlie Edwards; Luke Gribbon. "Radicalisation in the Digital Era: The use of the internet in 15 cases of terrorism and extremism." *RAND Europe* (2013). Available [here](#).

Walsh, Declan, and Suliman Ali Zway. "Facebook War: Libyans Battle on the Streets and on Screens" *The New York Times* (September 2018) Available [here](#).

Walter, Barbara F. "Internet Propaganda and the Recruitment and Radicalization of Foreign Citizens," conference proceedings: 2016 Peace Science Workshop: The Impact of Intra-war Processes on Post-conflict Outcomes (UNC Chapel Hill: 2016). Available [here](#).

Wilson, Lydia. "Understanding the Appeal of ISIS" *New England Journal of Public Policy* Vol 29 No. 1 (March 2017) Available [here](#).

Winter, Charlie. "Inside the Collapse of Islamic State's Propaganda Machine" *Wired Magazine* (December 2017) Available [here](#).

Winter, Charlie. "Islamic State Propaganda: Our Response to the Competition," in *Cyber Propaganda*

Winter, Charlie. "Media Jihad: The Islamic State's Doctrine for Information Warfare," *International Centre for the Study of Radicalization and Political Violence (ICSR)*, (2017) Available [here](#).

APPENDIX

Methodology

This assessment followed two phases: **problem analysis** and **analysis of responses**. Data was drawn from a literature review and expert interviews. The following questions guided data collection and analysis:

- What is the nature, scope, prevalence, and impact of the “weaponization of social media” phenomena? Are there, and what are the distinct typologies of weaponization?
- What is the role of social media in instances of weaponization of information?
- Where do we see these phenomena happening? Are there particular contexts (locations, incidents, platforms, or perspectives) that are exemplary of this problematic?
- What actors are implicated in the weaponization of social media? What are their goals? What techniques and tactics do they employ?
- What are the drivers, vulnerabilities, and other conditions that interact with weaponized information so as to amplify or exacerbate negative outcomes for populations of concern?
- What common patterns manifest in the initiation and development of weaponized social media phenomena?
- What are various actors doing to identify and respond to weaponized social media?

The literature review canvassed peer-reviewed studies as well as non-academic reports, book chapters, independent evaluations, policy documents, case studies, news articles, and other documentation related to the weaponization of social media.⁷⁶ Ninety-three published resources and over 270 news articles, investigations, opinion pieces, and other web-based content were identified.⁷⁷ A short list of 55 key literature sources and 30 online resources were reviewed in depth for the problem analysis phase. Approximately 15 additional sources were consulted for the analysis of responses phase.

Expert interviewees were drawn from within Mercy Corps and other humanitarian, peacebuilding and development organizations that might directly respond to these issues. Twelve external experts across various academic and practitioner domains were interviewed specifically in relation to the problem analysis phase.⁷⁸ A further 12 interviews and a **strategic workshop** with headquarters and field-based humanitarian, development and peacebuilding practitioners within Mercy Corps provided data for the analysis of responses phase.

⁷⁶ Sources were identified using a semi-systematic method based on combinations of search terms relating to: social media, disinformation, misinformation, computational propaganda, fake news, hate speech, dangerous speech, rumors, information warfare, information operations, hybrid warfare, polarization, radicalization, unrest, violence, elections, riots, social unrest, and upheaval. We also used back-chaining method to identify and prioritize key studies commonly cited by key scholars and practitioners, and took note of news stories and reports coming out of areas where weaponization of information issues are known to be prevalent, such as: Philippines, Myanmar, South Sudan, Nigeria, India, and Sri Lanka, as well as Syria, Ukraine, and Russian meddling in the U.S. 2016 Presidential election (although this last context was largely left out).

⁷⁷ This information was derived from the following disciplines, fields, and domains: Information and Computer Science; Technology & Society Studies; Data and Network Science; Communications and Media Studies; Counterterrorism or CVE Studies; Cybersecurity, Threat Intelligence, and Information Security; Cognitive Science and Human Behavioral Science; Political Economy, Comparative Politics, and Social Science; Conflict Studies and International Security.

⁷⁸ We are grateful to our interviewees: Alex Warofka (Facebook), Anahi Ayala Iacucci (Internews); Ben Nimmo (Digital Forensics Research Lab); Chris Tuckwood (The Sentinel Project); David Madden (Phandeevar Labs); Helena Puig Larrauri and Maude Morrison (Build Up); Hugh Brooks (Omelas); Nanjira Sambuli (Web Foundation); Sanjana Hattotuwa (ICT4Peace Foundation); Steven Livingston (The George Washington University); and Theo Dolan (PeaceTech Labs).

Data analysis

Data for the problem analysis was analyzed and is presented firstly through a collection of **case studies**. These include:

- Information operations — Russia’s targeting of the White Helmets in Syria
- Political manipulation — Elections in the Philippines
- Digital hate speech — Intercommunal violence in Myanmar
- Online radicalization and recruitment — the ISIS media jihad

The case studies, along with additional literature and expert sources, provided data to develop a **conflict analysis of social media weaponization (See page XX)**. This analysis used a common practitioner framework of conflict root causes and drivers to articulate the role and interaction between weaponized social media and societal conflict pre-dispositions, according to various conflict typologies.

The aforementioned strategic workshop provided a means of validating findings and assessing the practical implications of social media weaponization for organizations involved in or accountable to development, peacebuilding, and humanitarian processes. Several methods for determining relevance and calibrating effective response were designed in this workshop, and are presented in this assessment as potential tools for organizations mobilizing in response to these phenomena.

Chief among these tools is a **response framework**, which characterizes typical intervention types and potential entry points for these interventions. This framework was produced by mapping how government, the technology industry, UN/NGOs, civil society and academia are responding to phenomena related to the weaponization of social media. Across different domains (for example, security vs media communications vs sociology) over 100 examples of responses (strategies, policies, programs, tools, etc.) were identified that are addressing weaponization phenomena. These were mapped against a commonly experienced sequence of phases (and entry points) in which weaponization of social media threats unfold and can be responded to, according to the expertise of those interviewed for this assessment.

In-depth Case Studies

Case Study 1: Information Operations – Russia’s targeting of the White Helmet in Syria

In the digital era, coordinated disinformation operations have re-emerged as a central component of Russia’s information warfare strategy. Modern battlegrounds include areas of the former Soviet Union (such as Ukraine), but also places like Syria, a country of significant geostrategic importance for Russia that has been plagued by one of the worst refugee crises in modern history. Russia’s Internet Research Agency (IRA), cyber-operations divisions and affiliated cyber-mercenary groups have used social media to deceive, distort and disrupt information environments in targeted contexts. Their modern approaches to hybrid warfare will increasingly influence conflict in these (and likely other) contexts in the foreseeable future.⁷⁹

Weaponization via information operations

Information operations —defined as “the integrated employment...of information-related capabilities in concert with other lines of operations to influence, disrupt, corrupt, or usurp the decision-making of adversaries”—is a central component of Russia’s Information Warfare strategy.⁸⁰ In such situations, conflict is not declared overtly, and most activities are carried out below the threshold of conventional means. Clashes are “contactless,” using precision capabilities that target non-combatants (i.e., civilian populations, news media, and/or the private sector).⁸¹

Carried out in cyberspace and through social media platforms and networks, **psychological operations (PSY-OPS)** influence the emotions, motives, objective reasoning and ultimately the behavior of foreign governments, organizations, groups or individuals without firing a single shot.⁸² This has the effect of **denigrating or disrupting critical decision-making capabilities** (at multiple levels of societal governance), and **eroding horizontal cohesion between citizen groups, as well as vertical cohesion between citizens and government** through psychological subversion of social institutions and horizontal linkages, and through the amplification of uncertainty in information ecosystems.⁸³

When information operations are carried out on a systematic basis, they achieve a sort of **reflexive control**, or a means of influencing a partner or opponent to take a specific, desired action by conveying specially prepared information to incline him/her to act voluntarily.⁸⁴ Targets of Russian information operations have found their capacity for strategic decision making and collective action crippled.

79 See: Steven Livingston, “Forward” in “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018) ([Link here](#)); Congressional Research Service (CRS), *Information Warfare: Issues for Congress* (March, 2018); Sandra Svetoka, “Social Media as a Tool of Hybrid Warfare,” (NATO StratCom COE: May 2016); and Edward Lucas and Peter Pomeranzev, “Winning the Information War: Techniques and Counter-strategies to Russian Propaganda in Central and Eastern Europe” (CEPA: 2016).

80 CRS, *Information Warfare*, 3 quoting Joint Chiefs of Staff: 3-13, “Information Operations” (November 27, 2012).

81 Svetoka, 9-11; Lucas and Pomeranzev, 11.

82 CRS, 4). Psychological Objectives: Increase the targets suggestibility, gain control over information environment, create doubt or a sense of powerlessness, create strong emotional responses to a target, heavy intimidation. See Svetoka, 9.

83 Lucas and Pomeranzev, 12.

84 Lucas and Pomeranzev, 7-8

Techniques and tactics

While there is some variation in the descriptions of the specific steps taken to implement IO, we've identified a central sequence of practices across these sources that make up the "Digital Disinformation Playbook."⁸⁵

- 1) **Targeting:** Propagators of disinformation operations carry out **intelligence collection** on their target audiences via open-source channels on the web and analysis gathered by digital advertising agencies. This information is used to develop highly granular understanding of potentially receptive demographics.⁸⁶
- 2) **Content creation:** Operatives create and curate emotionally resonant or otherwise inciteful content (audio/visual, text-based information) for weaponization. This includes building narratives based on:
 - *Propaganda:* an idea or narrative-which can be misleading, but true, that is intended to influence
 - *Misinformation:* the spreading of unintentionally false information by individuals believing the information to be true
 - *Disinformation:* the spreading of intentionally false information by individuals seeking to manipulate others⁸⁷
- 3) **Dissemination:** Narratives are systematically disseminated through multiple channels, fusing together social and traditional media and offline contexts such as printed materials or public rallies.⁸⁸
- 4) **Amplification:** Propagated narratives (whether manipulated, misleading, fabricated or unverified) are then amplified by the following:
 - *Botnets:* pieces of software designed to create content and interact on social media platforms.⁸⁹
 - *Inauthentic accounts*
 - *Influencers:* willing or unwilling individuals or public personas who command large numbers of followers.
 - *#hashtag hijacking:* using an existing hashtag on social media for a different purpose than was originally intended
 - *Astrourfing:* imitating grassroots actions using coordinated, inauthentic accounts⁹⁰

85 See T.E. Nissen Framework – Svetoka, 11; Facebook "Information Operations," 2016; Lucas and Pomeranzev, 6; Sarah Oh and Travis Adkins, "Disinformation Toolkit" (InterAction: June 2018), 11; Giorgio Bertolin, "Digital Hydra: Security Implications of False Information Online," (NATO StratCom COE: May 2016), 8-9; Hossein Derakhshan and Claire Wardle, "Information Disorder: Definitions" in Understanding and Addressing the Disinformation Ecosystem (Annenberg School of Communication: 2017); Livingston, "Contentious Narratives" 2018; CRS 9-10; Lucas and Pomeranzev, "Winning the information War" 2018.

86 User signatures and **inferential metadata** from social media can be pieced together to create profiles on behavior, transactions and interactions, movements, personal details and preferences, and personal relationships and networks of any individual or group that might be targeted for disinformation campaigns.

87 Disinformation includes deliberately false news stories, manufactured protests, doctored content (such as photos or videos), and tampering with private communications before release.

88 While **mixed media information campaigns** use multiple social media channels and website-based platforms to perpetuate and amplify the reach of a single narrative, **cross-media campaigns** leverage a central channel around which the campaign is built and hyperlinked to. Both are extremely effective as masking inauthenticity. See Bertolin, 40-42.

89 Increasingly, bots are used for political reasons: to inflate the numbers of followers a politician has; to spread propaganda; to subtly influence political discourse; and to aggregate and broadcast content..." See Alice Marwick and Rebecca Lewis, "Media Manipulation and Disinformation Online," (Data & Society: 2017), 38, 39.

90 Russian IO fuse **astrourfing** with **hybrid-trolling** (deliberately provocative behavior that aims to distort online discussions and cause conflict among participants in order to advance ideological, political, or military objectives). (Bertolin, 29).

- *Trading up the chain*: planting a story with a small or local news outlet, from where it can then be amplified.⁹¹ It is here that disinformation can become misinformation, as unwitting recipients, themselves, can act as propagators simply by interacting with or sharing unverified, virulent content.⁹²

5) **Distraction**: All actors within the system work together to prevent objective sense-making within the target zone of operations. This can be achieved through saturating or flooding the information environment with “noise,” by disrupting telecommunications infrastructure, or banning the use of certain social media platforms.

Impacts and implications

The Syrian Civil Defense, the official name of the White Helmets, is a Nobel-prize nominated humanitarian organization made up of 3,400 volunteers who are credited with saving thousands of lives in Syria. They’ve also documented and shared thousands of hours of first-hand video footage of alleged war crimes and other atrocities, which has been used by UN war crimes investigators for advocacy and legal accountability work. [Graphika](#), the [University of Washington](#), the [Digital Forensics Research Lab](#) (DRFL) and [The Guardian](#) have detailed how the Russian government has made systematic use of information operations to amplify manufactured claims and false accusations against the White Helmets in the context of armed conflict in Syria, labeling them a terrorist organization with links to al-Qaeda and ISIS. So far, Graphika estimates that “bots and trolls linked to other Russian disinformation campaigns *have reached an estimated 56 million people on Twitter* with posts related to the White Helmets during ten key news moments of 2016 and 2017.”

These online defamation campaigns attempt to delegitimize the White Helmets’ status as a neutral and impartial humanitarian actor in an attempt to make them a legitimate target for kinetic attacks.⁹³ Over 210 white helmet volunteers have been killed since 2013. Their centers “have been hit by missiles, barrel bombs and artillery bombardment 238 times between June 2016 and December 2017.”⁹⁴

*“This propaganda ... paints a wrong image of the White Helmets, an image that is the opposite of what we really are...this gave the air forces an alleged reason to attack our centers and target us while on rescue missions. We have lost so many colleagues because of this.”*⁹⁵

“False accusations, abusive language and violent threats [also] chip away at the volunteers’ morale,” and “are designed to undermine the evidence they collect,” according to Graphika.

As a consequence, **the operational capacities of the White Helmets and their partners are eroded**, as they navigate a sustained defamation campaign that cripples morale, diverts attention away from live-saving activities, intimidates affiliates of the organization and puts affected populations in harm’s way. The

91 According to Marwick and Lewis, “media manipulators are able to **trade stories “up the chain”** of media outlets...by planting a story with a small or local news outlet who may be too understaffed or financially strained to sufficiently fact-check it. If the story performs well enough...it gets amplified beyond its current scope.” (Marwick and Lewis, 38-39).

92 Bertolin, 18-19; Oh and Adkins, 11.

93 For example, “KARMA IS A BITCH -> #WhiteHelmets killed. That will teach you to kill innocent children to fake #syriachemicalattack!! #SyriaHoax #MFANews.” (@BinsakSB). Furthermore “Those at the heart of these conspiracy theories, such as Vanessa Beeley, call for the White Helmets to be killed as legitimate military targets.” See Graphika, “Killing the Truth: How Russia is Fuelling a Disinformation Campaign to Cover up War Crimes in Syria (The Syria Campaign:2017), 13.

94 Graphika., 2017.

95 Graphika., 2017.

disinformation campaign has a net effect of helping to cover up the war crimes Syrian and Russian forces are allegedly committing on the ground.

Case Study 2: Political Manipulation – Elections in the Philippines

Philippine President Duterte has proven adept at exploiting social media for political gain, leveraging Facebook to reinforce positive narratives about his campaign, and to defame and silence opponents and critics.⁹⁶ The Philippine-based online news website Rappler has been the target of coordinated and sustained disinformation campaigns, after it exposed the systematic use of paid trolls,⁹⁷ bots,⁹⁸ networks of fake accounts and contracted influencers propagating pro-Duterte narratives (including mis- and disinformation) during the 2016 presidential election.

Here's how it works, according to Ong and Cabañes:

“In order to achieve emotionally resonant messaging and authentic branding that would trigger grassroots support, chief disinformation architects need to collaborate with influencers much more fluent in popular vernaculars when planning creative executions of digital campaigns. Strategists and influencers then harness the support of both community-level fake account operators tasked to generate momentum and energy for a campaign and create “illusions of engagement”. Then, unpaid grassroots intermediaries and “real” supporters ... amplify original campaign messages through shares and likes.”⁹⁹

Weaponization via political manipulation

Political manipulation is similar to information operations, but within the context of a single community or state. Political discourse is systematically manipulated by networked disinformation campaigns modeled after digital advertising approaches and operationalized through exploitative strategies and incentive structures. These practices have the power to set agendas, propagate ideas, debase political discourse and silent dissent, ultimately seeking to change the outcome of political events.

These disinformation campaigns play out in three phases:

- 1) **Design:** Establishing objectives, branding, core narratives, etc.
- 2) **Mobilization:** Onboarding influencers, fake account operators, and grassroots intermediaries, and preparing media channels
- 3) **Implementation:** Disseminating and amplifying messages and implementing other tactics such as digital black ops, #trending and signal scrambling.¹⁰⁰

96 Paladino, 16-17.

97 In internet slang, trolls refers to people who post inflammatory content online in order to cause argument or harass an individual or organization, either for their own amusement, or for another form of gain. Trolls are also associated with the presentation of extraneous information that sows or normalizes tangential conversations or narratives.

98 An internet bot, also known as a web robot, robot or simply bot, is a software application that runs automated tasks (scripts) over the Internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone.

99 Ong and Cabañes, 28.

100 Much like in the Information Operations case study 1, these political messaging campaigns also make use of **#hashtag hijacking**, **astroturfing**, and **trading up the chain** tactics.

Technique and tactics

The architecture of operatives who design strategies and implement tactics in disinformation campaigns includes:

- 1) **Employing PR strategists and creatives:** Elite strategists use marketing techniques to align campaign objectives with consistent messaging through “branding” and employ locally informed creative writers, who “weaponize popular vernaculars to maximize the reach of social media posts.”¹⁰¹
- 2) **Leveraging digital influencers:** Anonymous influencers and key opinion leaders (celebrities, pundits, etc) commanding between 50,000 and 2 million followers) weaponize popular culture trends and disseminate manufactured narratives through Twitter (via trending rankings) and Facebook.¹⁰²
- 3) **Amplifying through community-level fake operators:** Sub-contracted workers amplify messaging and localize narratives using pre-drafted, script-based messaging, predetermined schedules for media blasting, and click strategies.
- 4) **Engaging grassroots intermediaries:** Fan page moderators, unpaid volunteers and members of political organizations drive real grassroots engagement with disinformation by manufacturing “illusions of engagement”.

Impacts and implications

Despite efforts by Rappler and others to shore up free speech and provide counter-narratives to the Duterte propaganda machine, Duterte continues to use misinformation and coordinated disinformation campaigns to obfuscate controversial policies and practices, such as the war on drugs. The Philippines is now teeming with fake news, and other political agents are adopting, adapting and scaling up the digital disinformation model. This ought to be concerning for organizations working in international development and peacebuilding for various reasons, including:

1. There is an underlying system that propagates political disinformation in the Philippines. It thrives on an unregulated and highly profitable industry of digital advertising that incentivizes and exploits labor arrangements through “digital sweatshops” for political deception work. This digital economy of disinformation cannot be decoupled from the digital data industry. According to analysts at New America Foundation:

“Every post, click, search, and share is logged to a user profile, grouped into a segmented audience, and fed into machine learning algorithms. This data allows advertisers to infer an individual’s preferences, behaviors, and beliefs—all of which inform highly targeted digital advertising campaigns. Accumulated data is ..used not to drive purchasing decisions but to influence sentiment, political views, and voting behavior through precision propaganda..... Political disinformation succeeds because it follows the structural logic, benefits from the products, and perfects the strategies of the broader digital advertising market.”¹⁰³

2. These examples flip the “democratization” argument on its head, as proponents of so-called “Liberation Technologies” struggle to contend with the ways in which elites exploit the very nature and mechanisms of social media to control narratives and manipulate mass-based appeals for political

101 Ong and Cabañes, 45.

102 Ibid, 34.

103 Ghoash and Scott, 3-5.

mobilization. Social media in some cases has strengthened, not up-ended this relationship.¹⁰⁴

3. The distribution of harm in this system is multi-layered and multi-directional. While journalists and political opponents are silenced, intimidated, coerced and physically threatened by an authoritarian system, the political economy of disinformation sheds new light on the incentive structures that exploit local workers as active agents of disinformation, as well as the psychological harm that this system engenders at increasing scale through “race-to-the-bottom” work arrangements and the emotionally traumatic nature of this work.

Case Study 3: Digital hate speech – Intercommunal violence in Myanmar

Social media platforms can act to amplify *hateful, dangerous speech* in fragile contexts —such as Myanmar — where rumors, misinformation and disinformation can play a role in the incitement of intercommunal, electoral or other forms of violence. Extra-factual sources of information contribute to this problem, often amplified by social media. Digital hate speech has driven anti-Muslim sentiment in Myanmar and been directly implicated in the foment of intercommunal violence.

Weaponization via digital hate speech


While hate or dangerous speech has traditionally been propagated by traditional media such as radio or television or through in-person gatherings, the rapid proliferation of mobile phones and internet connectivity and the inherent technological and psychological features of social media platforms magnify these risks.¹⁰⁵ In today’s digital environment, every individual has the capacity and agency to develop, disseminate and consume potentially fabricated or misleading information on digital platforms with the power to increase communication speed, volume (of output and input), variety (of content), reach and coverage. Social media amplifies hate at scale. Finally, the inherent design of social media begets selective exposure, information bubbles, homogeneous echo-chambers, confirmation bias, and hyper-personalized, hyper-sensory, and hyper-insular information environments that reduce our cognitive capacity to objectively evaluate information.¹⁰⁶

Digital hate speech and virulent rumors warrant a unique aggregation of environmental factors, malicious strategies, and inadvertent actions in a logical narrative to know when thresholds for violent conflict are reached. In considering the patterns, conditions, features and drivers above, Mercy Corps, Do No Digital Harm and Peacebuilding have developed a theory of harm:

104 Zeitzoff, 6.

105 Svetoka, 5-6; Brandon Paladino, “Democracy Disconnected: Social Media’s Caustic Influence on Southeast Asia’s Fragile Republics” (Brookings Institute: 2018), 7-8; Eran Fraenkel, “A Critical Analysis of Digital Communications and Conflict Dynamics in Vulnerable Societies” (Internews: 2014), 2, 10-11; Nils Weidmann, “Communication, Technology, and Political Conflict: Introduction” *Journal of Peace Research* (2015)264; Deen Freelon, “Personalized Information Environments and Their Potential Consequences for Disinformation” in *Understanding and Addressing the Disinformation Ecosystem* (Annenberg School of Communication: 2017), 38, 60; Norman Vasu, et al, “Fake News: National Security in the Post-Truth Era” (RSIS: 2018),10-11; Gregory Asmolov, “The Disconnective Power of Disinformation Campaigns,” “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018), 32)

106 Paladino; 7-8; Fraenkel 10-11; Freelon, 38.



When such conditions are present, then, we would expect social media to have an amplifying effect on conventional conflict dynamics. In situations of security-related anxiety, rumors – especially if they conform to pre-existing worldviews and emotionally relevant narratives, and especially if the audiences are repeatedly exposed to them – can perpetuate unfounded threat claims, amplify in-group/out-group tensions, and motivate rational actors to engage and justify collective violence in the name of self-defense.¹⁰⁷

Techniques and tactics

Digitally transmitted communication amplifies conflict dynamics through the following **extra-factual sources of information**¹⁰⁸:

- 1) **Rumor:** *Unverified information that is transmitted from one person to others.* Rumors can be true, false or a mixture. At their core, mis- and dis-information are rumors.¹⁰⁹
- 2) **Hate speech:** Any form of expression (speech, text, images) that “demeans or attacks a person or people as members of a group with shared characteristics such as race, gender, religion, sexual orientation, or disability.”¹¹⁰
- 3) **Dangerous speech:** *Speech that has a special capacity to catalyze or amplify violence by one group against another.*¹¹¹

While some scholars claim that rumors “are nearly universal before the outbreak of riots and other forms of political violence,” and might even be “an independently sufficient cause of ethnic conflict,” others report weaker links between incidents of offline violence and online hate speech,¹¹² or frame the relationship between exposure to digital mis- and dis-information and violence as mutually-constituted and multi-directional.¹¹³

Nevertheless, we *can* point to a range of factors, conditions, and drivers that might predispose particular contexts to the negative effects of virulent and hateful digital speech. Susan Benesch’s (2014) *Dangerous Speech Guidelines*¹¹⁴ include a range of contextual factors and series of hallmarks that collectively estimate

107 Greenhill and Oppenheim, 2-3; Benesch, 2014 3, 21; Padalino, 9.

108 According to Greenhill and Oppenheim, extra-factual sources of information are (a) unverified at the time of transmission, (b) serve as a source of actionable knowledge, (c) intended to influence recipients’ attitudes or behavior, (d) are emotionally resonant, and (e) are framed in ways that fit pre-existing societal narratives. See Kelly Greenhill and Ben Oppenheim, “Rumor Has It: The Adoption of Unverified Information in Conflict Zones,” in *International Studies Quarterly* (2017), 2.

109 See John Bugge, “Rumour Has It: A Practice Guide to Working with Rumours” (Communicating with Disaster Affected Communities (CDAC): 2017), 8; and Greenhill and Oppenheim, 2. Importantly, a rumor can also take on multiple forms over time: “For example, a human trafficker can spread a rumor amongst refugees ... with the intent to deceive (disinformation), and a refugee can then pass this rumor to his friends and family not intending to deceive them (misinformation).” (Bugge, 8).

110 See Robert Faris et al., “Understanding Harmful Speech Online” Berkman Klein Center for Internet and Society (2016), 5-6. See also Kagonya Awori and Susan Benesch, “Umati: Kenyan Online Discourse to Catalyze and Counter Violence” (Conference Paper: IFIP 2013), 470.

111 Awori, 470; Theo Dolan et al., “Youth and Radicalization in Mombasa, Kenya: A Lexicon of Violent Extremist Language on Social Media” (PeaceTech: 2018), 6. This kind of speech is predicated upon the risk of violence (e.g. instilling fear by warning of impending threats, or by making an incitement to violence).

112 See Dolan et al, 4.

113 In other words, that rumors, disinformation, hateful speech (etc) are both the outcomes and drivers of polarization, radicalization, and violent behavior. Of course, many scholars point to the need for holistic, multivariate model to explain complex social behavior.

114 Benesch, 2014, 7-8.

the capacity of speech to inspire violence. Dangerousness can be estimated according to the following five factors:

- 1) a powerful speaker with a high degree of influence over an audience most likely to react
- 2) an audience with grievances and/or fears that the speaker can cultivate
- 3) a speech act understood by the audience as a call to violence
- 4) a social or historical context propitious for violence
- 5) an influential means of dissemination

Coupled with these factors, conditions for conflict are further maximized in **information-poor environments** with high levels of ambient fear, anxiety and/or uncertainty, and where civil society institutions and the rule of law are weak or nonexistent.¹¹⁵

Impacts and implications

In Myanmar, both misinformation in the form of **organic rumors** and speculation, and deliberate **disinformation** have played a significant role in amplifying grievances and triggering violence between groups of differing ethnic and religious identities. Anti-Islamic sentiment and intercommunal violence against Islamic identity groups has been the most visible example, and is linked to the country's deep Buddhist nationalist project.

Buddhist nationalists such as the 969 movement and Ma Ba Tha have exploited social media (particularly Facebook) “to stoke fear, normalize hateful views and facilitate actors of violence,” against identity groups (particularly Muslims, or the ethnic Rohingya) who are perceived and promoted as enemies of Buddhism, or of the State.^{116,117} Research has documented narratives of Islamic people (particularly the Rohingya) as illegal immigrants, Muslim terrorists and rapists, among other fabricated and incendiary messaging, reflecting the hallmarks of Benesch’s Dangerous Speech.¹¹⁸

Drivers and amplifiers include **the explosion in access to smartphones and the internet** — largely through reduction in price of mobile SIM cards and increased telecommunications coverage throughout the country, in addition to Facebook’s controversial internet.org initiatives (to provide low-cost basic internet service) and its zero-rating Free Basics package, essentially ensuring that “Facebook is the internet here.”¹¹⁹

There are **several instances of hate speech and misinformation implicated in violence in Myanmar**. For example, C4ADS documents how virulent rumors and online hate speech triggered the Mandalay riot of July 2014, in which approximately 20 people were injured, and 2 people were killed.¹²⁰ More recently, the

115 See Oh and Adkins, 13-14; Bugge, 9-11; and Greenhill and Oppenheim, 2.

116 C4ADS, “Sticks and Stones: Hate Speech Narratives and Facilitators in Myanmar” (2016).

117 Fink writes, “The speakers are highly regarded. The society has struggled with mistrust and violence, and Facebook has become the primary medium of communication. By creating and disseminating images of adversaries through the mass media—and in Myanmar’s case, social media—a group can generate widespread support for the idea that such adversaries cannot remain...” Christina Fink, “Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar” in “Contentious Narratives: Digital Technology and the Attack on Liberal Democratic Norms” *Journal of International Affairs* Vol. 71, No. 1.5 (2018). Fink also points out the emotionally resonant language used, the widespread reach of Facebook, and lack of space and resources for critically accessing information (i.e. weak media institutions).

118 C4ADS, 2016.

119 “By 2018, there were an estimated 16 to 30 million Facebook accounts.4 Today, Facebook is the internet for most people in Myanmar, and it has had a transformative effect on their lives. It has provided them with newfound freedom to obtain information, express themselves, and connect with others.” Fink, 26. See also Paladino, 6-8.

120 C4ADS, 11.

Myanmar military has carried out systematic clearance operations in 2017 against the Rohingya people in response to the Arakan Rohingya Salvation Army attacks, another situation that was largely amplified by digital hate speech and the propagation of unverified rumors.¹²¹ Hundreds of thousands of Rohingya have fled to Bangladesh as a result, with many reports documenting systematic rape by security forces and affiliated militia groups, and over 6,000 civilian deaths. United Nations officials and human-rights organizations have characterized the Rakhine State security operations as ethnic cleansing.

Case Study 4: Radicalization and Recruitment – the ISIS media jihad

In contrast to most violent extremist organizations, digital propaganda is a central aspect of the Islamic State’s approach to contemporary jihad. ISIS leadership go so far as to elevate the production and dissemination of propaganda as a form of worship.¹²² In doing so, ISIS has been particularly successful in exploiting defining elements of social media — peer-to-peer communication, near real-time user-generated content and low-cost dissemination of multimedia content — to spread extremist propaganda; target, manipulate, and seek to recruit supporters; and coordinate tactical operations.

Weaponization via radicalization and recruitment

ISIS has rebranded the image of radical extremism through messages of inclusiveness and belonging — many of which are disseminated online. An initial focus on gaining currency among extremist and fringe demographics has been replaced by a broader approach to appeal to wider audiences.¹²³ By rebranding itself, ISIS has set the bar for strategic communications efforts among violent extremist networks, elevating the importance of inclusive narratives, crowd-sourced content, multi-modal media and strategic counter-speech in the arsenal of asymmetrical warfare.

The affordances of social media allow for near-instantaneous access to emotionally resonant narratives, reduce costs associated with participation in formal organizations, offer a relatively risk-free entry point for potential recruits to find like-minded individuals, and create a social environment in which extremist views are normalized. In short, “social media provides cheaper and more accessible pathways to radicalization.”¹²⁴ These features of the contemporary information landscape are perfectly aligned with the franchised nature of modern terrorist organizations and operations.¹²⁵

Techniques and tactics

ISIS adapts conventional recruiting techniques to the digital information environment, allowing for more targeted campaigning, more emotionally-resonant messaging, and more personalized exchanges between operatives and potential recruits. Specific tactics include:

Targeting younger, tech-savvy millennials who feel isolated from society, lack a strong sense of identity or purpose, and are frustrated with their economic, familial, or interpersonal situations.¹²⁶ These

121 According to Paladino, “In the wake of the ARSA attacks, the Facebook group for Ma Ba Tha supporters registered a significant spike in anti-Rohingya messages, illustrating the powerful tendency of such online associations to amplify and reinforce the thoughts of its individual members.” (Paladino, 9).

122 Winter, 11, 17.

123 Koerner, 2016.

124 Zeitsoff, 9.

125 Theohary and Rollins, 4.

126 See Dylan Gerstel, “ISIS and Innovative Propaganda: Confronting Extremism in the Digital Age,” *Swarthmore International Relations Journal* Issue 1 (2017) pg. 2; and Lydia Wilson, “Understanding the Appeal of ISIS” *New England Journal of Public Policy* Vol 29 No. 1 (March 2017), pg. 8.

individuals are particularly susceptible to ISIS' re-branded, inclusive messaging and viral, accessible content.¹²⁷ Through the two main production companies within the Islamic State (*Al-Hayat* and *Al-Furqan*), media operatives produce propaganda films (such as the so-called "mujatweets") that are shared on social platforms, circulate monthly magazines (such as *Dabiq*), live stream updates from the front lines, design video-games, and even circulate job advertisements.¹²⁸

Highlighting themes of openness, inclusion, and participation.¹²⁹ Through coordinated messaging campaigns, ISIS recruits are shown a sense of purpose, collective identity and meaning.¹³⁰ ISIS media operatives also place stories of ordinary fighters front and center, humanizing extremists through relatable messaging and contextualizing radicalization within local grievances and community perspectives.¹³¹ Personal videos and photos from the front lines are effective recruitment and "symbolic currencies" that are used to drive user engagement and radicalization.¹³²

Moving recruits toward radicalization one step at a time. Exposure to one set of ideas can open the door for other, more radical thinking to take root. This incremental conditioning process, known as "**red-pilling**" in certain internet subcultures, normalizes extremist views over time and erodes the capacity for objective reasoning.¹³³ Once potential recruits and supporters have been inundated with pro-ISIS content, recruiters move to more **private communications channels**, such as encrypted messaging applications (i.e., WhatsApp or Telegram, Kik, Threema, etc.) or Skype for more targeted, personalized and intimate one-on-one outreach that makes targets feel included, valued and part of a community.¹³⁴

Mobilizing and coordinating operations. Social media are also leveraged for mobilizing supporters, sharing information and coordinating tactical operations. The use of social media for these purposes "is important for non-state actors such as insurgent groups, particularly if these groups lack formal structure or are dispersed over large geographical areas."¹³⁵ In this way, the information environment is a key resource, staging ground and site of conflict for extremist efforts.

Violent extremist groups have leveraged social media for **targeting and intelligence gathering** and **sharing logistical information** between geographically dispersed groups. According to NATO, these platforms, channels and networks are used to identify potential targets for military actions, as user-generated content, participatory maps and user metadata become resources for actionable intelligence in conflict zones and other inaccessible environments.

Social media are also a conduit for **sharing training or instructional materials**. This includes information pertaining to local travel conditions, how to stage attacks and how to mask communications from law enforcement.¹³⁶ Operatives share step-by-step instructions on how to build and deploy weapons such as

127 Wilson, 3, 8; Gerstel, 1; and Bertolin, 22.

128 Wilson, 5.

129 See Charlie Winter, "Media Jihad: The Islamic State's Doctrine for Information Warfare," International Centre for the Study of Radicalization and Political Violence (ICSR), (2017), pg. 15-16.

130 Gerstel, 2.

131 See Saltman and Winter, 43; Brendan Koerner, "#jihad: Why ISIS Is Winning the Social Media War" WIRED Magazine (March 2016); and Wilson, 7.

132 Wilson, 3.

133 Marwick and Lewis, 29. Importantly, people that become red-pilled in one ideological area are more likely to be red-pilled in others: susceptibility for red-pilling is an indicator for and entry point of Islamic radicalization (Freelon, 18).

134 Gerstel, 3.

135 Svetoka, 15-16.

136 Zeitzoff, 10; Catherine Theohary and John Rollins, "Terrorist Use of the Internet: Information Operations in Cyberspace" (Congressional Research Service (CRS): March, 2011), 12-13; Lorenzo Vidino and Seamus Hughes, "ISIS in America: From ReTweets to Raqqa" (Program on Extremism GWU: December 2015), 20-21

explosive devices or even malware via social media.¹³⁷ Social media can also be a vehicle to facilitate both kinetic and digitally-derived forms of violence, for example **social cyber attacks**, in which cyber militias have engaged in online defamation campaigns and have weaponized rumors and false information to incite panic and/or violence.

Impacts and implications

While ISIS is not the first militant group to use social media for information activities and gaining support, their use of social media is distinct in three ways:

- 1) **The focus on propaganda elevates the status, importance and role of online volunteers, influencers, film makers, graphic designers and other technical specialists within the greater ISIS network.**¹³⁸ Digital activists, dedicated freelancers, online influencers, keyboard warriors, and internet recruiters play a central role within the organization, lowering the bar for participation, blurring the distinction between “support” and “membership,” and challenging conventional notions of affiliation and hierarchy. A GW report, *ISIS in America*¹³⁹ makes note of 4 key roles:
 - **Nodes:** Leading voices within an online community that function primarily as content creators for ISIS propaganda and pool new followers by bridging social networks within and between platforms.
 - **Amplifiers:** Popular platform users (anonymous or known) that disseminate content by re-tweeting, liking, favoriting and sharing ISIS-related materials created by other users.
 - **Shout outs:** Connectors and curators that introduce potential recruits to the ISIS community, facilitate introductions to authentic accounts, and promote newly created accounts of previously suspended users, thereby playing a pivotal role in the resilience of ISIS’s online communities in response to account suspensions and takedowns.
 - **Spotters:** Operatives that scan the social media landscape to identify potential candidates for radicalization and engage them via more private channels.
- 2) **ISIS has set the standard for strategic communications innovation among violent extremist networks.** ISIS is the first extremist group to have developed bespoke software for disseminating and amplifying its propaganda. Its mobile application, “Dawn of Glad Tidings” allowed for automated posting and re-tweeting of pre-drafted content via remote control of the users’ Twitter account. ISIS media operatives could exploit the accounts of anyone downloading the app to amplify their media content, allowing media campaigns to proliferate rapidly, spaced out to avoid Twitter’s spam-detection algorithms, and increasing anonymity for those individuals actually running the accounts.¹⁴⁰
- 3) **ISIS is adaptive and persistent, despite coordinated efforts among technology companies, governments and civil society to counter them.** When suspected accounts are de-platformed, blocked users come back online using alternative handles, which are authenticated among the remaining network through geo-location and list-based means.¹⁴¹ Operatives also use signaling techniques such to authenticate accounts, as well as URL obfuscation, encrypted messaging apps, username manipulation and cross-platform migration to undermine detection efforts and ensure digital survival.¹⁴²

137 Theohary and Rollins, 4, 7.

138 Winter, 12; Koerner, 2016.

139 Lorenzo Vidino and Seamus Hughes, “ISIS in America: From ReTweets to Raqqa” (Program on Extremism GWU: December 2015) pg. 23-26.

140 Saltman and Winter, 41; Gerstel, 4; Wilson, 4.

141 Svetoka, 34-35.

142 See Audrey Alexander, “How to Fight ISIS Online” Foreign Affairs (April, 2017)

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