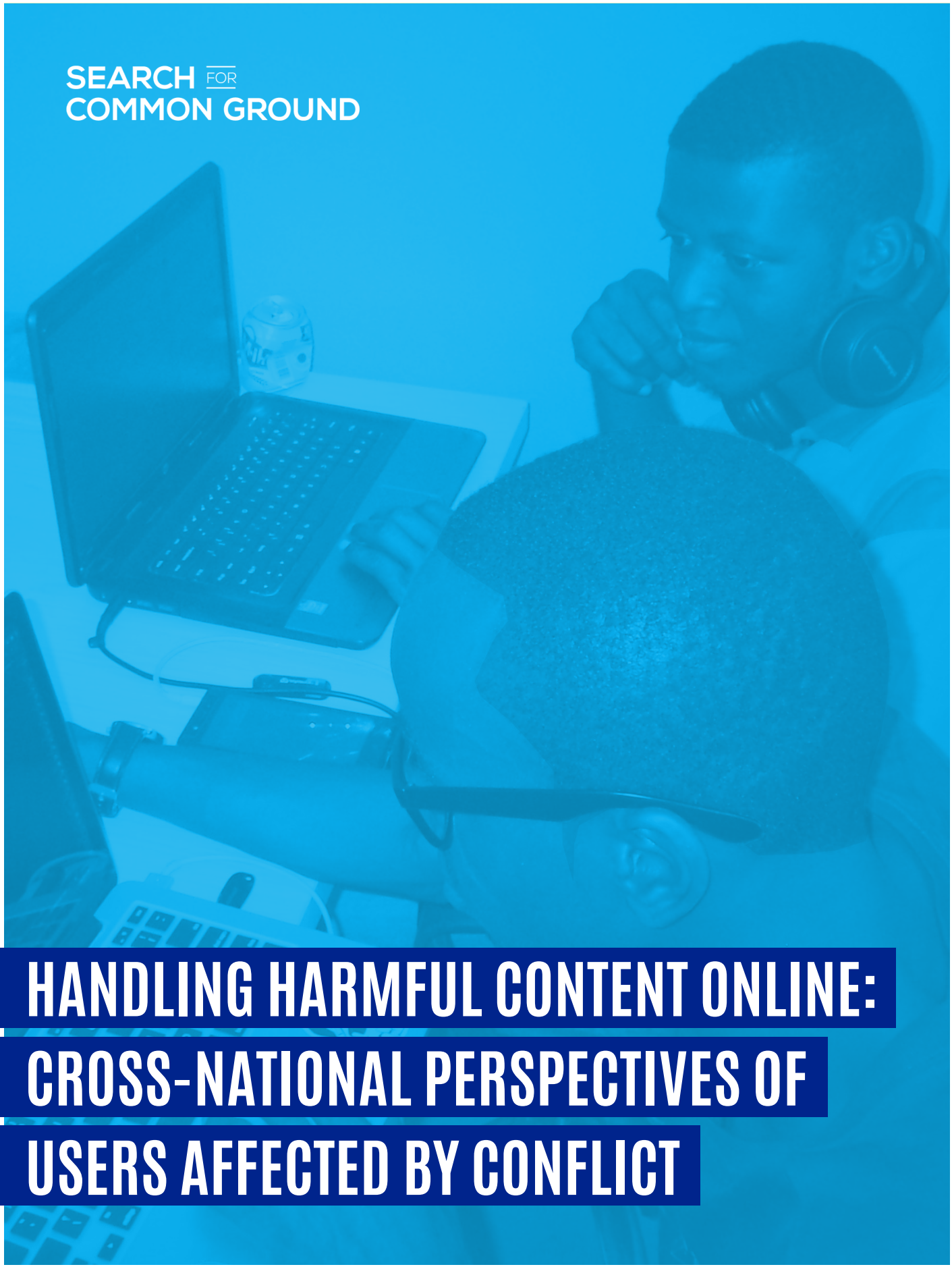
SEARCH FOR
COMMON GROUND



HANDLING HARMFUL CONTENT ONLINE: CROSS-NATIONAL PERSPECTIVES OF USERS AFFECTED BY CONFLICT

About

Search for Common Ground (Search) undertook this study with financial support from Facebook. The study aimed to identify barriers that users in conflict-affected societies face in using mechanisms, such as content reporting, to enforce online social norms. This report is the second in a series developed with support from Facebook. The following members of Search's Institutional Learning Team led the research design and authored this report: Aisalkyn Botoeva, Lilly Crown Wilder, Adrienne Lemon, and Marin O'Brien Belhoussein.

Acknowledgements

There are many who graciously participated in this research as interviewees or focus group participants. We appreciate their time and eagerness to take part in the study. We are grateful to all of the researchers who collaborated with us in collecting primary data: Harry Myo Lin, Jude Kallick, Fathima Azmiya Badurdeen, Vilma Guadalupe Portillo Cienfuegos, Temirlan Jailobaev, Kanykey Jailobaeva, and Nendo research company. We are also grateful for the continued contributions from Search's Global Affairs and Partnerships team, including Mike Jobbins, Sharmila Shewprasad, and Katie Smith.

To cite this report

Institutional Learning Team, "Handling Harmful Content Online: Cross-National Perspectives of Users Affected by Conflict." Search for Common Ground (April 2021).

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

Organized disinformation campaigns and malign actors use social media to increase polarization and incite violence around the world. Harmful online content—including hate speech, false news, cyberbullying, and inflammatory rumors—can spread quickly and reach millions. While research on how malign actors use social media is growing, it is still rare for researchers and policy makers to directly engage with end users in conflict settings. This is largely due to methodological challenges related to access, language barriers, political constraints, mobility limitations, and most importantly establishing trust among potential study participants. As a result, much less is known about how individuals and groups living in conflict settings respond to harmful content online.

Search for Common Ground (Search) aimed to address this knowledge gap by exploring the online experiences of social media users from three groups:

1. Direct participants in violence, including youth gang members, particularly from indigenous backgrounds in Guatemala and Honduras, former fighters from the al-Shabaab insurgency group in Kenya and Tanzania, and recent armed group recruits of the Arakan Army in Myanmar.

2. Those who socialize with direct participants of violence, including friends and family members, religious leaders, such as imams and Buddhist monks, as well as social workers and community organizers who have worked with vulnerable youth groups.

3. Active resistors of violence, including local youth and young professionals, civil society organizers and leaders, journalists, and NGO workers who have worked in interfaith dialogue and social media literacy initiatives.

We aimed to address the central research question: how do users in violent conflict settings experience and handle harmful content online?

We spoke to 68 individuals from these groups across four geographic regions: the Northern Triangle (El Salvador, Guatemala, Honduras), East Africa (Kenya and Tanzania), Central Asia (Kyrgyzstan), and Southeast Asia (Myanmar). With rising internet penetration rates in these conflict-affected contexts online violence is increasingly mirroring and complementing offline violence. It is important to understand the individual and group-level resiliencies towards these newer manifestations of violence in ongoing conflict settings.

The study found that:

1. The worst violent content, and most effective responses to it, appear in private groups and messages.

- The most violent content and criminal acts, including harassment, threats, and extortion, occurred in private groups and direct messages. However, many study participants also use private groups on Facebook to collectively report content to tackle misinformation and hate speech.

2. Users actively try to make their online spaces safer.

- Over one-third of the participants used official reporting channels to address hate speech but not all feel encouraged to continue using them. Direct participants in violence reported feeling uncomfortable with some content and wanted to take proactive actions to reduce their family members' exposure to such harmful content.
- Some users directly engaged with creators or boosters of harmful content via private messaging or settling disputes by 'fighting it out' in person.
- 'Exit strategies' including blocking, unfollowing, and deleting harmful accounts and pages are the most common way through which users seek to reduce their exposure to harmful content online.

This research presents three opportunities to create safer online spaces:

1. Encourage users already using 'exit strategies' to shift to active content reporting and moderation. Many users are not inclined to turn to official content reporting channels—or continue to use them—because they do not trust that reporting will lead to tangible results.

- Social media companies should improve transparency and feedback loops in formal content reporting features, through instant response features such as "You and sixty-three others reported this content as harmful" and explanations of what type of response can be expected.
- At the time of reporting harmful content, social media companies should direct users to additional ways to deal with harmful content online, referring them to resources on how to engage in non-adversarial communication with users of different beliefs or linking to a database of hate speech or misinformation management organizations. Such resources would provide concrete options for users to go beyond blocking and unfollowing.

2. Form partnerships with organizations in-country with deep understanding of conflict dynamics to help identify and transform cultural and social barriers to content reporting.

- Social media companies should assign a point-person within a country portfolio team to regularly engage with local civil

society, religious communities, youth groups, national security groups, and local influencers to share information, concerns and risks in the online space.

- Co-design and implement interventions in conflict-affected communities to transform the structural, social, and cultural barriers to mitigating harmful content online and contribute to a healthy online environment such as online activism campaigns.

3. Focus on making private platforms safer.

- Create central resources and training for group administrators and community moderators to uphold standards in private group settings such as Facebook groups and messaging platforms such as WhatsApp, Twitter DMs, and Signal.

- Facilitate networking or information-sharing channels amongst group administrators and community moderators of different groups to share best practices in mitigating harmful content and fostering positive dialogue in their groups.

- Create reporting mechanisms for closed group, conversation and message-level communications.

2 INTRODUCTION

Social media has created a more connected world. But along with the good, the biases, fake news, and prejudices that disrupt societies play out through Facebook, Instagram, WhatsApp, Twitter, YouTube, TikTok, and other platforms.

Malign actors, ranging from street gangs,[1] rebel groups,[2] and mafias[3] to terrorist networks[4] and authoritarian regimes,[5] have used social media platforms and digital technologies as means to sew animosity and hatred in societies, coordinate violent forms of collective action and conduct surveillance and oppression.[6] Dangerous online content can spread rapidly to millions of users and has amplified divides and catalyzed mass violence.[7] The threat of harmful content is particularly pronounced in fragile contexts that have a history of violent conflict.

While research on how malign actors use social media to advance their interests is growing, there has been little direct engagement with individuals and groups in close proximity to violent conflict, and understanding of how they handle harmful content online. Search for Common Ground (Search) set out to address this knowledge gap by engaging with individuals from countries with histories of long-standing divisions and violent conflicts.

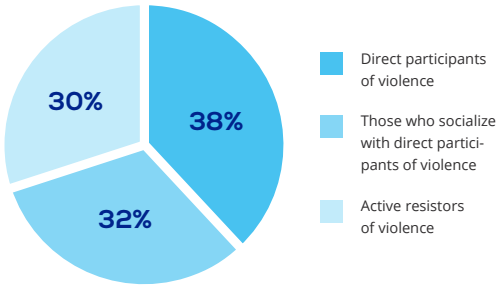
Our study revealed a spectrum of activities that individuals and groups in close proximity to violent conflict are taking to keep their digital spaces safe. While strategies vary, many are motivated by a desire for ownership and agency in their interactions with harmful content online.

Depending on their background, individuals feel comfortable using 'exit strategies', such as blocking, unfollowing, and deleting accounts and pages that they see as harmful. Individuals also settle disputes that spark online by 'talking it out'.

This study suggests a number of opportunities that social media companies and civil society organizations can leverage to encourage users to move along a continuum from using 'exit strategies' to utilizing online content reporting more proactively.

RESEARCH DESIGN

Search engaged with social media users from three groups for this study:



1) Direct participants in violence. This group primarily consisted of men, whose average age was 26-years old. Most of them come from a lower socio-economic background and reside in rural areas. Depending on the context, members of this group were either previously or are currently involved in youth street gangs or militant and armed groups. 38% of our study participants came from this background.

Illustrative profile

Aboud, a 40-year-old former al-Shabaab member from Mombasa, Kenya, directly participated in violence through his involvement with the group. He thinks that social media fuels divisions and tries to make sure his children are not exposed to harmful content that al-Shabaab members circulate as well as other content that he considers corrupting, by blocking, deleting, and unfollowing pages and accounts.

Read more about Aboud on Page 21

2) Those who socialize with direct participants of violence. This group was evenly split between men and women and included friends and family members, educators as well as social workers and community organizers who have worked with vulnerable youth groups. They tend to be skeptical of engaging in violence and worry about their family members, friends and other community members getting drawn into youth gangs, insurgency or armed groups. They are better educated and more likely to be employed in comparison to the first group. Users from this group represented 32% of study participants.

Illustrative profile

Mateo is a 27-year-old youth leader from Quiché, Guatemala who socializes with those who directly participate in violence. He almost joined a gang but managed to resist. He doesn't report harmful content due to his fear of retaliation by the person accused.

Read more about Mateo on Page 25

3) Active resistors of violence. This group was evenly split between men and women and included local youth and young professionals, civil society organizers and leaders, journalists and NGO representatives who have worked in interfaith dialogue, and social media literacy initiatives. This group primarily resides in capital or other larger cities and has higher educational attainment as well as socio-economic status. They represented 30% of study participants.

Illustrative profile

Maryam is a 27-year-old Rakhine-Muslim woman from Myanmar. She is an active resistor to violence through her work with youth on information management and her involvement in local politics. She reports content and believes that mobilizing like-minded people to mass report is effective.

Read more about Maryam on Page 15

ACCESS TO THE FIELD

In each context, the research team used a combination of purposive and snowball sample approaches. We followed a purposive sampling approach to identify direct participants of violence and then worked with local researchers to identify appropriate individuals with a range of perspectives on the topic. In purposely selecting the first group, we relied on Search's in-country teams and researchers' past experience as well as their established rapport with communities of interest. For example, our researcher in Kenya had completed extensive studies involving al-Shabaab members and had easy access to a number of individuals. Our researcher in Myanmar had connections to Arakan Army recruits, Buddhist and religious leaders, and youth in Yangon and Rakhine. Our researchers in Kyrgyzstan had done numerous studies in the South of the country with youth groups considered susceptible for recruitment to forces in Syria.

Once researchers completed interviews with the first group, they followed a snowball sampling approach and asked participants to refer their family members, neighbors, and peers in the community in order to recruit members of the second group. Finally, the researchers used a combination of both purposive and snowball sampling methods to recruit participants who matched characteristics of active resistors of violence. Throughout the entire research process, the research team made sure that our study did no harm to the participants or to the surrounding communities.

RESEARCH CONTEXT

We spoke to 68 individuals from these groups in December 2020 in seven countries across four geographic regions: the **Northern Triangle** (El Salvador, Guatemala, Honduras), **East Africa** (Kenya and Tanzania), **Central Asia** (Kyrgyzstan), and **Southeast Asia** (Myanmar). Each country has a history of long-standing social, political, and/or identity-based conflicts. These countries also experienced instances of mass violence against civilians in the last two decades. As internet penetration rates increase quickly in these and other conflict-affected contexts, the takeaways from this research are relevant for users from similar backgrounds who live in proximity to violence elsewhere around the globe.

To guarantee anonymity of and confidentiality to the individuals interviewed, the research team kept only the demographic data such as age, ethnicity, and religion. All other personal information such as names, names of family members, and others are changed. Pseudonyms are used throughout the report.

Northern Triangle Focus Countries:

El Salvador

Population: 6.5 million people

Internet Penetration: 50% penetration[8]

Internet Users: 3.3 million—420,000 new internet users between 2020 and 2021

Guatemala

Population: 18.1 million people

Internet Penetration: 65% penetration[9]

Internet Users: 11.8 million—217,000 new internet users between 2020 and 2021

Honduras

Population: 10 million people

Internet Penetration: 38% penetration[10]

Internet Users: 3.8 million—365,000 new internet users between 2020 and 2021

East Africa Focus Countries:

Kenya

Population: 54.4 million people

Internet Penetration: 40% penetration[11]

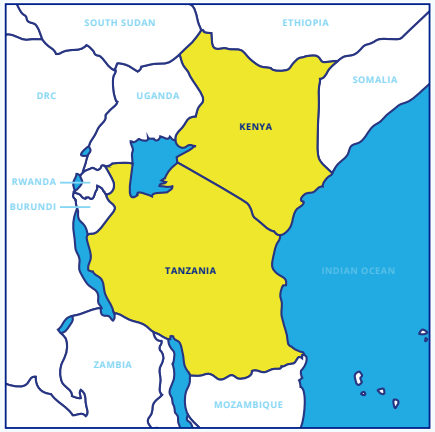
Internet Users: 21.8 million—435,000 new internet users between 2020 and 2021

Tanzania

Population: 60.6 million people

Internet Penetration: 25% penetration[12]

Internet Users: 15.2 million—435,000 new internet users between 2020 and 2021



Central Asia Focus Countries:

Kyrgyzstan

Population: 6.6 million people

Internet Penetration: 50% penetration[13]

Internet Users: 3.3 million—260,000 new internet users between 2020 and 2021

Southeast Asia Focus Countries:

Myanmar

Population: 54.6 million people

Internet Penetration: 43% penetration[14]

Internet Users: 23.7 million—2.5 million new internet users between 2020 and 2021

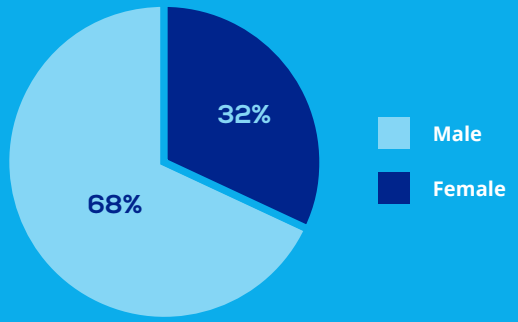
PARTICIPANT DEMOGRAPHICS

No. of Participants: 68

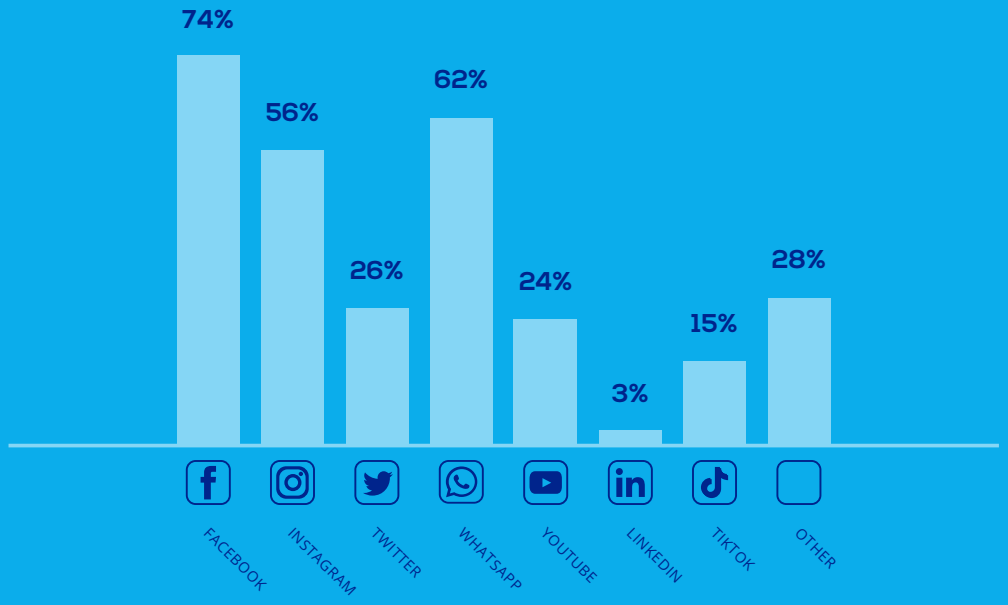
Age Range: 16-52

Average Age: 30

Gender Breakdown of Participants:



Percentage of Participants Using Various Social Media Platforms:



3 KEY FINDINGS

The majority of study participants—including those engaged in violent groups—did not want to interact with content that they consider harmful. Moreover, participants from all groups have taken steps to create safer online environments for themselves and their close ones. Although the ways in which individuals respond may differ, participants indicate that they want to feel a sense of ownership and agency when they handle harmful content.

In the sections that follow, we discuss the main findings of the study related to participants' experiences of harmful content and three main tactics that they have relied on to tackle harmful content online. We discuss the factors that have encouraged some to use official content reporting channels and prompted others to rely on 'exit strategies' instead.

The worst violent content, and most effective responses to it, appears in private groups and messages.

For respondents in contexts like El Salvador, the most violent content (including violent imagery and rhetoric, extremist, and violent ideologies) and most direct violence (such as harrassment, extortion, and intimidation) happened through private groups and direct messages. At the same time, respondents in Myanmar, Kyrgyzstan, and Tanzania spoke of the importance and power of private groups to create safer online spaces. These private groups provided positive community and safe spaces to discuss experiences while also enabling burden-sharing for reporting content and tackling misinformation and hate speech.

Not surprisingly, closed Facebook groups and private messaging applications such Messenger, WhatsApp, and others serve as spaces through which organized criminal groups recruit new members, terrorize communities and deploy their extortion tactics. A respondent in Honduras noted that gang members recruit new members among youth, by engaging them in casual conversations through closed groups and chats and stated, "When young people engage, they're already involved in an initiation ritual." Another respondent from Honduras who lives in San Pedro Sula observed that, "Because poverty is still on the rise and therefore crime as well. In the area where we live, more and more minors are recruited."

Gang cliques and organized criminal groups gather information about people through their personal profiles on social media and use the information against them to extort money. A respondent in Honduras shared that they often use information about family members, place of work and children's school and threaten sending messages like: "I know

MARYAM'S PERSPECTIVE:

Maryam is a 27-year old Rakhine-Muslim woman working with youth on information management.

She also is highly active in local politics. She said that she has engaged in mobilization efforts through which like-minded people have come together in Facebook groups to report certain kinds of harmful content. She believes that this strategy is efficient because the social media companies review the posts and decide what to do based on the numbers of people who report the same content.

Private Facebook groups where users know and trust other members allow for individuals to overcome certain barriers. Particularly amongst Rakhine-Muslims, like Maryam, who are a minority within a minority in Myanmar, participation in these private reporting groups is common. Nearly half of the people from Myanmar who participated in the study referenced these groups and their utility.

where you work, I know your name. You have two children, they are in such a school. I need you to give me 10,000 lempiras a week or I'm gonna kill them." Respondents in El Salvador have similarly shared that the gang members are known to keep a close tab of other community members' posts online, using this information further to threaten or extort money.

Yet in other contexts, resistors of violence have mobilized and tackled harmful content online through closed groups on Facebook. Grassroots activists in Myanmar, Kyrgyzstan, and Tanzania all described coordinating their reporting of posts and accounts because of the belief that a critical mass will influence social media companies to deal with harmful content. These activists form private Facebook groups to flag content easily to their peers which is then reported en masse to get the attention of the social media platform.

Users across all three groups reported that they had tried to make their online spaces safer.

There are three main types of tactics that study participants use to tackle harmful content online: official content reporting channels, settling disputes that spark online through direct engagement, and reliance on 'exit strategies'. Each tactic is described below, followed by a discussion of why users from different backgrounds prefer one or another.

Over one-third of the study participants, across all three groups, reported that they had used official reporting channels on social media platforms. They are not all inclined to use them--or continue to use them--however. A discussion of why users from different backgrounds feel discouraged to utilize official reporting channels appears in the section on 'exit strategies'. *Read more about this dynamic on page 12.*

Some users prefer to directly engage with perpetrators of harmful content. Citing harmful content that incites violence based on ethnic or religious divides, some users shared that they tried to 'talk it out' by striking-up a conversation through Messenger or 'fight it out' by setting-up a meeting in person. Users who socialize with direct participants of violence shared that engaging with perpetrators of harmful content online is their way of 'taking matters into one's own hand'. *Read more about this perspective on page 13.*

Users from all three groups cited 'exit strategies' as the most common way through which they try to reduce their own and their close ones' exposure to harmful content online. Study participants often leave groups that they dislike on WhatsApp, unfriend and unfollow accounts that they consider sources of wrongful or harmful content on Twitter and Facebook, and ban or block people on Facebook and other platforms to regulate their digital spaces. 'Exit strategies' offer users a clear and concrete way for a user to end his or her engagement with harmful content. Unlike reporting, the result is guaranteed as well as immediate. Users are able to draw on their own judgement and take action to control their environment. *Read more about this dynamic on page 15.*

Users are more likely to turn to reporting mechanisms when they are able to see the consequences of the actions

Our study found that those who responded the most positively about the efficacy of reporting and use reporting mechanisms most frequently are the same ones who have reported content and seen tangible consequences, such as removal of content or accounts. Matthew, a 32-year old activist who works for a local Kenyan organization that aims to prevent violent extremism, reflected on the first time that he reported a post on behalf of his NGO: *"It is amazing how Facebook reacts to these reports."* The report was in response to propaganda being spread in advance of the December 2020 bi-election in Msambweni, a town on the southeastern coast. Matthew said the poster was trying to fuel tensions by claiming that one of the politicians is gay and that *"the only reason that he reached his position is because he is sleeping around with people in power."* Other messages shared made accusations about the politician's family members in an attempt to convince others to inflict harm on them. The post Matthew flagged was removed soon after by a Facebook administrator. Matthew's experience resonated with study participants in other countries, most of whom worked as journalists, civil society actors, youth leaders, and others who identified themselves as active resistors of violence in their communities.

Active resistors of violence commonly use content reporting mechanisms for specific material and also have contributed to initiatives on digital media literacy to tackle the problem broadly. Notably, some active resistors of violence have delivered social media literacy training sessions to empower community members to tackle harmful content. Such respondents pointed out that these initiatives primarily have targeted self-selected audiences, however, and need to be scaled-up if they are to reach at-risk youth and other groups that are most susceptible to harmful content.

For Gulmira, a 32-year old journalist from Kyrgyzstan, it was clear that she needed to contribute to broader and more systematic efforts if she wanted to tackle harmful content in her community and country. Two years ago, she started working on a project that engaged journalists, educating them on techniques of content reporting and effective use of social media platforms. Eventually, the project expanded and now targets Kyrgyz-speaking audiences primarily outside of the capital city. While she is proud of the project, Gulmira did caution that it has not yet yielded the desired results:

> *To be totally honest, I wouldn't rate it as a success. Because people appear to skip parts of our training sessions that touch on topics of how to avoid propaganda, or how not to fall for extremist information. Also, for some reason, our people seem to watch a lot of content that has violence and extremist content. And it is very difficult to reach out to them. That's why I'm not fully satisfied with the project, although it is a necessary step in the right direction. One of the ways to potentially reach youth who are susceptible to harmful content is to introduce similar sessions in school curriculum. We only reach adults who are self-selected because they want to be taught. However, the real impact can be made through schools.*

Speaking of their activist stance, those who identify themselves as active resistors of violence, emphasized that training sessions on social media literacy need to be scaled-up. Andrea, a 17-year old indigenous youth influencer in Guatemala echoed Gulmira's points and noted that the young people

AZIM'S PERSPECTIVE:

Azim is a young Uzbek entrepreneur who lives in Osh, in Kyrgyzstan's south.

Here the most recent deadly conflict occurred in 2010 among Kyrgyz and Uzbeks. The violence has left deep divisions and mistrust between these groups. Although many Uzbek-owned businesses suffered disproportionately in comparison to the Kyrgyz-owned businesses, and were attacked, looted or appropriated in the aftermath of the 2010 conflict, younger generations of entrepreneurs like Azim are finding ways to generate income. He uses Instagram and other platforms to boost his business opportunities by following motivational accounts and pages of famous people in Uzbekistan and Russia, and by advertising his services. Azim shared that he recently observed an argument between an Uzbek blogger and Kyrgyz social media users. A group of users who identified themselves as Kyrgyz were harassing and threatening the Uzbek blogger with ethnic slurs online. Things escalated to a point where Azim and his peers decided to meet the perpetrators of the hateful posts in person and 'fight it out' in order to settle the dispute. After all, in Azim's words, *"everyone can post in social media, but what you say online can have consequences too."* Azim shared how incitement of real-world violence is common, in particular on Telegram, WhatsApp, and TikTok. He saw this 'fight it out' response as being the most direct way in which users in his community could move from online arguments and harassment to offline engagement.

particularly need to be trained on how to identify harmful content and how to avoid falling prey to gangs and other malign groups who extensively use social media: *"[T]here is a need to explain prevention measures, such as not publishing misinformation and sharing sensitive information about themselves. Young people need to have such training sessions at their schools. I feel that it would be good to teach them about the risks of social media that they don't know about and don't know how to remedy later on."*

The users' goal in utilizing official reporting channels was to counter harmful content that stigmatizes and harasses groups that belong to a certain ethnicity, religion, or support a particular political agenda. Not all users saw tangible results of their content reporting, however, and therefore have relied on alternative tactics discussed below.

Some users prefer to directly engage with perpetrators of harmful content

Users want to be in charge of policing their offline and online communities. Emphasizing this point, some study participants reported that they have settled disputes that sparked online by engaging in-person with the perpetrators of harmful content to ensure there were consequences for online actions.

Users who socialize with direct participants of violence expressed that engaging with perpetrators of harmful content online is their way of 'taking matters into one's own hand.' Having lived through violent conflicts themselves, they want to minimize the effect of online fake news or wrongful accusations based on ethnic or religious grounds on the social dynamics within their communities. Asmin, a 38-year old Buddhist monk from Myanmar shared his observations of what content tends to fuel conflicts: *"Tension goes up when people discuss minority ethnic groups. Also there is a lot of propaganda using religion. If fuels a lot of conflict. Comments under BBC and Voice of America when they report about religious groups contain a lot of hate speech."* During the pandemic, harmful content reflects the public concerns around health as well as the long-seated ethnic and religious hatred in his observations: *"[I]n the COVID times, what also is widely circulating is content portraying monks as Satans for organizing large gatherings. It went viral and fuels so much tension".* Instead of merely ignoring or seeking 'exit strategies,' Asmin chooses to engage directly with some of the users who post harmful content. He shared that he used to reach out to people through Messenger but now also uses his phone when he can: *"If I only write then people do not pay much attention, or just respond with accusations. So, I try to set up personal connections and take a good amount of time to persuade them. Only very few people explore the root causes of conflicts, so I engage with them to help them in their learning journey."*

Joseph, a 23-year old Kenyan artist and youth activist similarly chooses to engage with those who circulate harmful content personally. Speaking of content that has in the past incited violence on religious grounds in his community, such as wrongful accusations of Muslims or negative portrayals of the Prophet Mohammad, Joseph shared:

> *If I know the person personally, I take it upon myself to talk to them and make them understand the problem. I'd really emphasize that the vulnerable are the teenagers because when a teen gets such information he or she doesn't*

think of the results. He just shares it. Then the issue gets escalated. If he gets aggravated because of the drawing of the Prophet, he doesn't really think there may be repercussions of what he or she shares. If he is frustrated with his religion such as that of the drawing of the Prophet, he will get really angry. He doesn't really think that there might be any repercussion of what he or she shares. So I take it upon myself to make them understand and talk to them. The elderly also like our parents and our grandparents. They also get such information from different people on their WhatsApp. I think we should take it upon ourselves to talk to them and make them understand.

As can be seen from the experiences of Asmin and Joseph, tactics of engaging with users who circulate harmful content provides a sense of agency and ownership. Yet, such tactics can take a dangerous turn, particularly among youth. A young Uzbek entrepreneur's case from Kyrgyzstan illustrates how verbal assaults online can end in fights offline. While one-off fist fights may not pose a great threat to a community as a whole, they can contribute to the escalation of violence between already volatile groups who can then become inspired to commit acts of mass violence.

Users like Asmin, Joseph, and Azim tried content reporting mechanisms in the past but were discouraged from continuing to report because of the lack of feedback from social media companies and lack of repercussions for the perpetrators of harm.

Specific political, social, and cultural factors shape users' experiences offline and online, prompting them to choose 'exit strategies' to control their own online experience

The type of violent content experienced by participants takes many forms, from direct extortion to stereotypes and hate speech, and is shared in different ways, including public posts, direct messages, and private groups. The type of content sometimes affects their willingness to use official reporting systems provided by the platform. Participants cited a number of specific political, social, and cultural barriers to using content-reporting channels.

Political factors are significant, particularly for users from ethnic and religious minority backgrounds as well as those who have directly participated in violence. Individuals from these groups described fearing for their security and anonymity online, thereby mostly relying on 'exit strategies' to regulate their online spaces. Lack of trust in institutions is another factor that contributes to underutilization of content reporting channels. Some users who reported harmful content online in the past reverted to the use of 'exit strategies' over time as they lost faith that the perpetrators of harmful content will be held accountable. Cultural norms are a third factor that shape users' perceptions around content reporting. For example, some study participants associated content reporting with whistle blowing or telling on someone and described stigmas against such actions.

Study participants, particularly from ethnic and religious backgrounds, reported low levels of trust in social media and fear of government censorship and retaliation. They avoid using content reporting channels, partly because they assume that their online activities are monitored by perpetrators of

ABOUD'S PERSPECTIVE:

We spoke with Aboud, a 40-year-old former al-Shabaab member from the Indian Ocean port city of Mombasa, Kenya's second-largest city.

Due to his previous affiliation with the group, he is highly aware of the real dangers that exist from hate speech and other types of harmful content online. He discussed how, whenever he joins a new social media platform, ranging from the professional (LinkedIn), to a dating app (Gogo), he is able to clearly identify fake accounts that are al-Shabaab recruiters. This situation made him increasingly skeptical of how well social media companies address dangerous activity occurring on their platforms. He said he doesn't trust social media platforms and monitors everything his four children look at and do online very closely.

Aboud believes that *"social media is really trying to divide communities."* In terms of using in-platform content reporting tools, he chooses not to because, *"When you report on Facebook then you don't see your report addressed. How are they going to take action against something happening between Kenyans? Hence, I don't see the need to report it."*

KHADIJA'S PERSPECTIVE:

Khadija is a Muslim woman, age 44, in Lashio, Myanmar, who began spreading extremist messages after her family was attacked by Buddhist youth.

She described how she was susceptible to messaging from leaders of ethnic armed groups who sympathized with her. This engagement led her to share harmful content herself.

Khadija now believes that reporting contributes to exacerbating violence. She says when people create private groups with the intended purpose of mass reporting, it is another way for tensions to escalate and can lead to religious violence. For example, she has seen posts where influential leaders spread hate speech or calls to attack another group. Others have expressed with strong support to the posts in the comments as well as in messenger groups that are aimed at recruiting people to join the violence. Even though she knows this content is harmful and dangerous, she doesn't report out of fear. She says that she doesn't trust social media platforms and what they will do once she contacts them. Instead, she prefers to only use social media platforms as a source of finding information and *"explaining the truth to her friends about the news."*

violence or government authorities. Some social media users, like Aboud and Khadija, discussed in the perspective profiles, assume that their content reporting activities are traced and avoid flagging or complaining about harmful content so that they do not get entangled with the police. These users believe that reporting would bring more trouble than relief.

Direct participants of violence and those from ethnic and religious minority backgrounds prefer 'exit strategies' for an additional reason--to disengage with potential perpetrators of violence in an effort to ensure their own security. A 30-year old current al-Shabaab member, Wadeen, shared:

> *If I don't like what someone is sending me, I just block and unfollow. If I don't, they might escalate and somehow find my phone number. So right when I see those messages, I just block. I even said, let me just do away with Facebook for a little while and quit my account.*

In Myanmar, a 42-year old Kyi similarly stated, *"If I report or engage with those who spread hate speech [online], I will get attacked."* Fernando, a 28-year old in El-Salvador who lives in a gang-controlled area, talked about the persistent violence that his community faces every day. He described his tactic to minimize violence and hatred online by using 'exit strategies':

> *In Whatsapp I just ignore, delete, and block the group of contact that is sending me things I don't want to be seeing. On Facebook, when I see unsuitable videos, evangelical pastors talking violently of others, I just ignore and delete. I also hide offensive things.*

For those in close proximity to violence, threats of violence online have direct repercussions offline, and 'exit strategies' offer a means to mitigate such threats.

Low level of trust in institutions of social media is the next factor that prompts users to underutilize content reporting channels and default to 'exit strategies.' Users often have different expectations of what they think the 'results' of reporting will be. When the response, or lack thereof, doesn't meet the expectation, users become disillusioned with the reporting process overall. Participants in our study who have socialized with those directly involved in violence most commonly reported that they have tried using content reporting channels in the past. They were discouraged, however, and are not inclined to report in the future because they did not see tangible results in the form of reported accounts being taken down or updates and reports from social media companies that added to the transparency of the reporting process.

When users see how much harmful content gets left up on social media, they have the impression that social media companies do not care about the issue. Users discussed how populist and autocratic regimes have gained traction globally and use social media to advance divisive rhetoric without repercussions. Additionally, users across geographic regions have grown exceedingly disappointed that social media companies have not actively curbed sexist, nationalist, racist, and otherwise divisive messages that some government leaders deploy. Given the visibility of many of these accounts, and the fact they have been allowed to stay online, people are less certain that reporting harmful content would make a difference.

Albert is a 35-year-old social worker from El Salva-

dor who is also a member of the LGBTQ community. He spoke about how Facebook has allowed the president of the country to build a following based on violent and hateful messaging:

> *I feel that he has a lot of followers in networks, so... In fact, to a great extent, the success of his campaigns in different periods—when he was a Mayor and then the president—has been due to social networks and his messages on Twitter and Facebook, and even TikTok. He is a populist with a lot of capacity to understand the behavior of digital spaces, and people! He spreads his messages that are not in favor of peace, are not in favor of equality, are not in favor of human rights, but quite the opposite. He questions the work of human rights organizations, and of women's work, and he demonizes sexual diversity.*

Like Albert, many users would like to see perpetrators of harm also be held accountable. This desire is especially true in relation to organized criminal groups and malign actors who have inflicted significant harm both online and offline in some communities. Echoing Albert's sentiments, some users referred to numerous cases when extremist or criminal groups stated their names, affiliations, and crimes committed online but did not face any punishment. These experiences foster a sense of helplessness. Many users choose to disengage because they do not see how it will contribute to altering the current situation.

Another notable element of the social factor that influences the use of 'exit strategies' includes immediate personal relationships. Users from all three groups discussed their connections with extended families, friends and colleagues, reminding us that such personal relationships have an important effect on how people use technology. Some users see 'exit strategies', including muting or unfollowing, as a better or less extreme way to disapprove of content posted by people they care about, particularly compared to formally reporting the content. More importantly, 'exit strategies' are less disruptive to the various, everyday social situations in which people are embedded. Our respondents shared that they may not want to report someone if they are a family member, relative, former high school friend, or other acquaintance. For example, some respondents in Kyrgyzstan shared that they see harmful rumors and potentially conflict-inciting content circulating through WhatsApp groups with their former classmates or relatives from their villages. Similar sentiments about using 'exit strategies' were true in all seven countries, as people maintain multiple layers of kinship, friendship, and collegial connections, and do not want their online activities to hurt relationships in which they are invested, sometimes deeply.

Finally, cultural norms of retribution against those who blow the whistle on others prevent many community members from utilizing reporting channels. In contexts where retribution against those who blow the whistle on others is common, users fear that reporting content will threaten their physical security. These fears are not baseless and are deeply rooted in experiences that then shape local cultural norms of what it means to 'tell on or report someone.'

Retaliation may occur within interpersonal relationships. Some users who report hate speech, for instance, fear retaliation from other users who

MATEO'S PERSPECTIVE:

Mateo is a 27-year-old youth leader from Quiché, Guatemala, who doesn't report online due to the fear of retaliation and a lack of trust in social media.

Mateo almost joined a gang but managed to resist. He now works to help prevent other youth from being recruited. He demonstrated a deep understanding of ethnic conflict that is cultivated in online spheres in his area which is the current manifestation of long-term tensions in the Ixil region between urban Guatemalans and indigenous Mayan communities. Mateo expressed his confusion about what is at stake by allowing hate speech and other harmful content to remain on the social media platforms, but said he does not feel comfortable using in-platform reporting mechanisms. He feels that, if he reports, *"the social media platform will close my account, and maybe they'll inform the person I complained against, and they will be able to take action against me."* He described how he doesn't believe that he is alone in this thinking either: *"Many people use social networks but are not completely aware of their rights and what could happen to them after they report."* He also discussed how there is a lack of social norms in the culture surrounding complaining. In general, making complaints is not acceptable behavior, which translates into the online space. Mateo's experience as a youth leader and working at the grassroots level has shown him that young people in particular do not feel empowered to reject hate speech or racist content, and Mateo feels that these attitudes need to change at a cultural level.

may engage in cyberbullying or doxing. In other instances, retaliation can take the form of punishment from government or security actors. In many of the contexts examined in this study, there are tense relationships between government authorities and civilians. Some users fear that their online activity will be used to punish them.

Some users describe that reporting can lead to direct harm. In the Northern Triangle, users were the least active in content reporting. Those living in gang-held areas across Guatemala and El Salvador, in particular, expressed that they did not feel comfortable utilizing reporting mechanisms.

Diana is a 39-year-old domestic worker who lives in a gang-held area and often sees images and videos of violent acts shared by gang members to incite responses from their rivals. Though she knows it is wrong, she doesn't feel comfortable reporting them. She said: *"I may be punished for denouncing someone. I think it's part of our culture, and this is reflected in our social media practices. I do not report; I just move along. If you react and report, then the problems begin."*

When respondents faced these challenges, many relied instead on 'exit strategies' such as blocking, unfollowing, and deleting accounts and pages that they see as harmful. This category of action offers immediate relief from harmful content. The users' experiences discussed above show that aside from awareness and knowledge about formal content reporting channels, users would need to overcome multiple political, cultural, and social barriers if they are to more proactively engage in moderating their digital spaces.

4 KEY RECOMMENDATIONS

We suggest three main sets of online and offline solutions that can drive systemic, collective efforts to make digital spaces safer for existing and new users.

1. Encourage users already utilizing 'exit strategies' to engage in active content reporting and moderation.

Social media companies should focus their efforts on engaging those already acting to make their online spaces safer. Users across all four geographic regions, particularly those who are connected to direct participants of violence, are aware of the content reporting mechanisms on social media platforms. They are not inclined to use them - or continue using them - because they do not know if reporting leads to tangible results. Instead, they favor other responses such as blocking, unfollowing, or offline engagement such as direct confrontation, reports to police, or turning to trusted community figures.

- **Improve transparency and feedback loops in formal content reporting features.** Options could include instant response features that tally the number of other complaints of harmful content on this post and provide users immediately with the steps that the company will take if the content is deemed harmful. Additional options to 'report anonymously' or 'report with feedback' could help to reduce barriers to reporting. Examples could be "You and sixty-three others have reported this content as harmful" or include country-based reporting data such as: "Two hundred videos flagged as 'hate speech' have been removed this month in your country."

- **Direct users to additional ways to deal with harmful content online.** Action in this area could include the development of links to resources on how to engage in non-adversarial communication with users of different beliefs, a link to a database of hate speech or misinformation management organizations, and other opportunities to make online spaces safer that go beyond block and unfollow this user.

2. Form partnerships with organizations in-country with deep understanding of conflict dynamics to help identify and transform cultural and social barriers to content reporting.

Effectively addressing conflict dynamics originating from and exacerbated by social media activity requires a coordinated effort between headquarters teams, regional or country experts, host-country governments, and local civil society. Success ultimately requires the buy-in and ownership from those who are most affected by the crises and those who will be critical to long-term success. Social media companies need to have the right kind of personnel and the strategic vision to work with local government and groups working on hate speech to address the multi-layered dimensions of conflict that affect the dissemination and reporting on violent content.

- **Assign a civil society point-person assigned to a country portfolio** and create regular meetings for local civil society, religious communities, youth groups, national security groups, and aid workers to share intelligence, concerns and risks, and figure out who was best placed to respond to challenges and opportunities presented there. This should be reflected in agreeing to a shared definition of purpose and benchmarks for success, involving communities, local government, and the whole of the company itself.

- **Co-design and implement interventions in conflict-affected communities in partnership with local organizations to transform the structural, social, and cultural barriers** to mitigating harmful content online and contribute to a healthy online environment. This could include online activism campaigns that emphasize the transformative potential of users who strive to make online spaces safer. Such collaborations are crucial to respond to the conflict dynamics of today as well as the conflict ecosystem of tomorrow as new users in conflict-affected communities continue to come online.

3. Focus on making private communication channels safer.

According to direct participants of violence, the most dangerous activities typically occur in closed Facebook groups and messaging platforms such as WhatsApp, Twitter DMs and Signal.

- **Create guidance and resources for administrators and moderators on upholding community standards in private group settings.** Create central resources and training for administrators and moderators of closed channels on how to manage violence and conflict within their communities. Such resources and training are necessary but not sufficient to address the issues related to violent content in closed groups. It is also important to understand the existing incentives for administrators and moderators to access resources and participate in training then seek to expand these incentives.

- **Facilitate networking or information sharing channels amongst administrators and moderators of different groups to share best practices** in mitigating harmful content and fostering positive dialogue in their groups. In consultation with local civil society, local government partners, and country experts, as appropriate, map the key administrators and moderators of private online groups. Explore ways in which to create networks among these stakeholders and foster information sharing both within country contexts and across borders.

- **Broadly publicize strategies for reporting hateful or violent content shared through closed groups and private messages.** Create reporting mechanisms for closed group, conversation, and message-level communications. If such channels are already in place, make them more visible and educate users on how to proactively report such harmful content at each level.

This study shows the importance of changing the narrative around content reporting to reflect the reality that people closest in proximity to violent conflicts generally want to make their online spaces safer by using different tactics. Future initiatives to promote safe digital spaces should be informed by these perspectives.

5 SOURCES

1. Behrman, Megan. “When gangs go viral: Using social media and surveillance cameras to enhance gang databases.” Harv. JL & Tech. 29 (2015): 315.

2. Carroll, Rory. (2013). “Blog del Narco: Author Who Chronicled Mexico’s Drugs War Forced to Flee,” 27 April. https://www.theguardian.com/world/2013/may/16/blog-del-narco-mexico-drug-war

3. Rudner, Martin. “‘Electronic Jihad ‘: The Internet As Al Qaeda’s Catalyst for Global Terror.” Studies In Conflict & Terrorism, vol. 40, no. 1, 1 Jan. 2017, pp. 10 - 23.

4. MacAskill, Ewen. (2015). “British Army Creates a Team of Facebook Warriors,” January 31 https://www.theguardian.com/uk-news/2015/jan/31/british-army-facebook-warriors-77th-brigade

5. Roberts, Tony, and Gauthier Marchais. “Assessing the role of social media and digital technology in violence reporting.” Contemporary Readings in Law & Social Justice 10, no. 2 (2018).

6. Kemp, Simon. “Digital El Salvador: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 11, 2021. https://datareportal.com/reports/digital-2021-el-salvador.

7. Kemp, Simon. “Digital Guatemala: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 11, 2021. https://datareportal.com/reports/digital-2021-guatemala.

8. Kemp, Simon. “Digital Honduras: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 11, 2021. https://datareportal.com/reports/digital-2021-honduras.

9. Kemp, Simon. “Digital Kenya: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 11, 2021. https://datareportal.com/reports/digital-2021-kenya.

10. Kemp, Simon. “Digital Tanzania: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 12, 2021. https://datareportal.com/reports/digital-2021-tanzania.

11. Kemp, Simon. “Digital Kyrgyzstan: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 23, 2021. https://datareportal.com/reports/digital-2021-kyrgyzstan.

12. Kemp, Simon. “Digital Myanmar: All the Statistics You Need in 2021 - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, February 12, 2021. https://datareportal.com/reports/digital-2021-myanmar.

SEARCH FOR
COMMON GROUND

www.sfcg.com