

[Back to Newsroom](#)

Meta

Understanding Social Media and Conflict

June 20, 2019

By Samidh Chakrabarti, Director of Product Management, Civic Integrity; and Rosa Birch, Director of Strategic Response



Update on May 21, 2020 at 3:52PM PT:

We've talked a lot about the proactive detection technology we use to identify hate speech and other abusive content before people report it to us. On [May 12](#), we announced some of the progress that we've made. But hate speech is always changing and new trends emerge. To stay ahead of these trends, our team has been enhancing this technology to identify new forms of potentially inflammatory speech that hasn't been reviewed for possible removal from our platform. In countries at risk of conflict, we may demote this content to reduce the risk of it going viral or inciting violence or hatred, taking local context into account. We also update our list of slurs prohibited by our Community Standards. We seek input from experts and civil society organizations everywhere as part of the process. We're doing this in many languages, including Burmese and Sinhala, as we work to keep Facebook safe for people everywhere.

Originally published on June 20, 2019 at 2:00PM PT:

At Facebook, a dedicated, multidisciplinary team is focused on understanding the historical, political and technological contexts of countries in conflict. Today we're sharing an update on their work to remove hate speech, reduce misinformation and polarization, and inform people through digital literacy programs.

Last week, we were among the thousands who gathered at [RightsCon](#), an international summit on human rights in the digital age, where we listened to and learned from advocates, activists, academics, and civil society. It also gave our teams an opportunity to talk about the work we're doing to understand and address the way social media is used in countries experiencing conflict. Today, we're sharing updates on: 1) the dedicated team we've set up to proactively prevent the abuse of our platform and protect vulnerable groups in future instances of conflict around the world; 2) fundamental product changes that attempt to limit virality; and 3) the principles that inform our engagement with stakeholders around the world.

About the Team

We care about these issues deeply and write today's post not just as representatives of Facebook, but also as concerned citizens who are committed to protecting digital and human rights and promoting vibrant civic discourse. Both of us have dedicated our careers to working at the intersection of civics, policy and tech.

Last year, we set up a dedicated team spanning product, engineering, policy, research and operations to better understand and address the way social media is used in countries experiencing conflict. The people on this team have spent their careers studying issues like misinformation, hate speech, and polarization. Many have lived or worked in the countries we're focused on. Here are just a few of them:

Ravi, Research Manager

With a PhD in social psychology, Ravi has spent much of his career looking at how conflicts can drive division and polarization. At Facebook, Ravi analyzes user behavior data and surveys to understand how content that doesn't violate our Community Standards — such as posts from gossip pages — can still sow division. This analysis informs how we reduce the reach and impact of polarizing posts and comments.

Sarah, Program Manager

Beginning as a student in Cameroon, Sarah has devoted nearly a decade to understanding the role of technology in countries experiencing political and social conflict. In 2014, she moved to Myanmar to research the challenges activists face online and to support community organizations using social media. Sarah helps Facebook respond to complex crises and develop long-term product solutions to prevent abuse — for example, how to render Burmese content in a machine-readable format so our AI tools can better detect hate speech.

Abhishek, Research Scientist

With a masters in computer science and a doctorate in media theory, Abhishek focuses on issues including the technical challenges we face in different countries and how best to categorize different types of violent content. For example, research in Cameroon revealed that some images of violence being shared on Facebook helped people pinpoint — and avoid — conflict areas. Nuances like this help us consider the ethics of different product solutions, like removing or reducing the spread of certain content.

Emilar, Policy Manager

Prior to joining Facebook, Emilar spent more than a decade working on human rights and social justice issues in Africa, including as a member of the team that developed the African Declaration on Internet Rights and Freedoms. She joined the company to work on public policy issues in Southern Africa, including the promotion of affordable, widely available internet access and human rights both on and offline.

Ali, Product Manager

Born and raised in Iran in the 1980s and 90s, Ali and his family experienced violence and conflict firsthand as Iran and Iraq were involved in an eight-year conflict. Ali was an early adopter of blogging and wrote about much of what he saw around him in Iran. As an adult, Ali received his PhD in computer science but remained interested in geopolitical issues. His work on Facebook's product team has allowed him to bridge his interest in technology and social science, effecting change by identifying technical solutions to root out hate speech and misinformation in a way that accounts for local nuances and cultural sensitivities.

Focus Areas

In working on these issues, local groups have given us invaluable input on our products and programs. No one knows more about the challenges in a given community than the organizations and experts on the ground. We regularly solicit their input on our products, policies and programs, and last week we published the principles that guide our continued engagement with external stakeholders.

In the last year, we visited countries such as Lebanon, Cameroon, Nigeria, Myanmar, and Sri Lanka to speak with affected communities in these countries, better understand how they use Facebook, and evaluate what types of content might promote depolarization in these environments. These findings have led us to focus on three key areas: *removing* content and accounts that violate our Community Standards, *reducing* the spread of borderline content that has the potential to amplify and exacerbate tensions and *informing* people about our products and the internet at large. To address content that may lead to offline violence, our team is particularly focused on combating hate speech and misinformation.



Removing Bad Actors and Bad Content

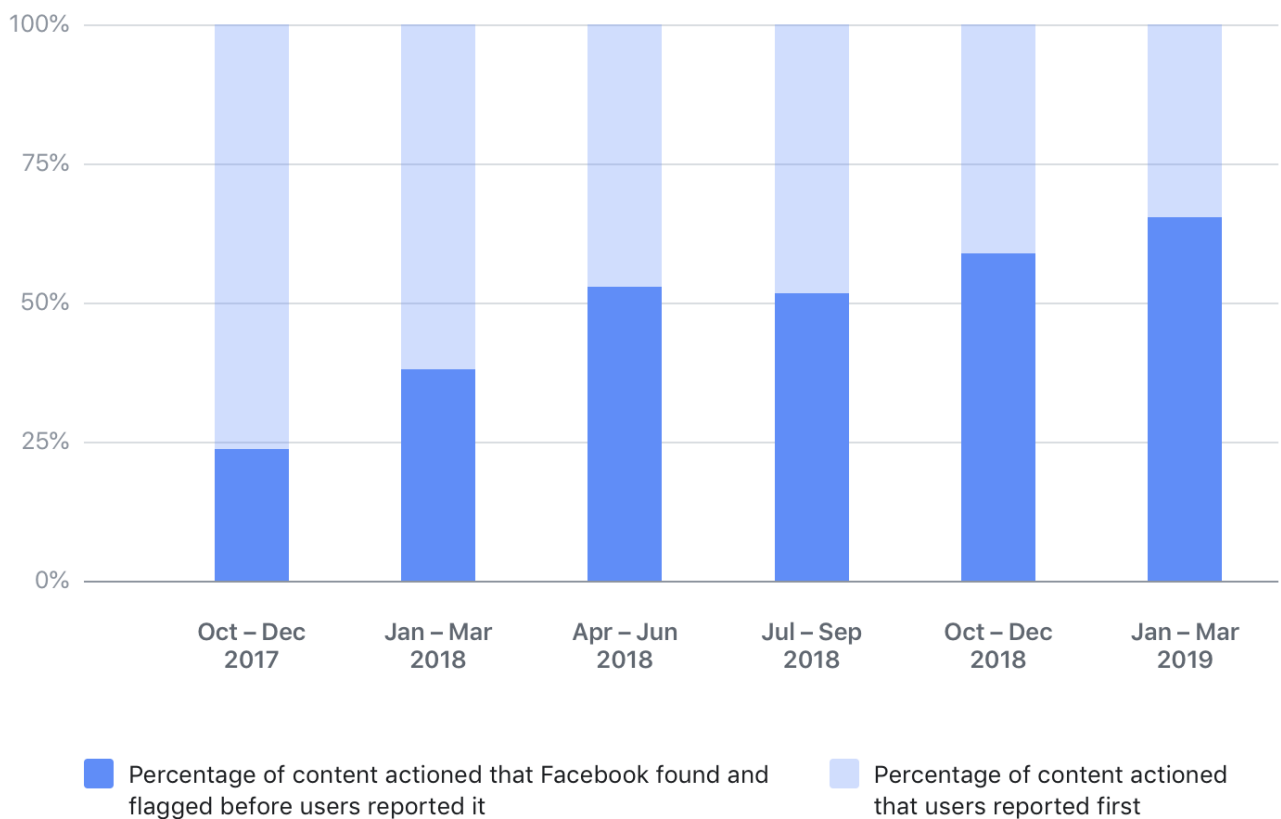
Hate speech isn't allowed under our Community Standards. As we shared last year, removing this content requires supplementing user reports with AI that can proactively flag potentially violating posts. We're continuing to improve our detection in local languages such as Arabic, Burmese, Tagalog, Vietnamese, Bengali

and Sinhalese. In the past few months, we've been able to detect and remove considerably more hate speech than before. Globally, we increased our proactive rate — the percent of the hate speech Facebook removed that we found before users reported it to us — from 51.5% in Q3 2018 to 65.4% in Q1 2019.

PROACTIVE RATE

Of the violating content we actioned, how much did we find before users reported it?

Facebook increased our proactive rate from 58.8% in Q4 2018 to 65.4% in Q1 2019.



We're also using new applications of AI to more effectively combat hate speech online. Memes and graphics that violate our policies, for example, get added to a photo bank so we can automatically delete similar posts. We're also using AI to identify clusters of words that might be used in hateful and offensive ways, and tracking how those clusters vary over time and geography to stay ahead of local trends in hate speech. This allows us to remove viral text more quickly.

Still, we have a long way to go. Every time we want to use AI to proactively detect potentially violating content in a new country, we have to start from scratch and source a high volume of high quality, locally relevant examples to train the

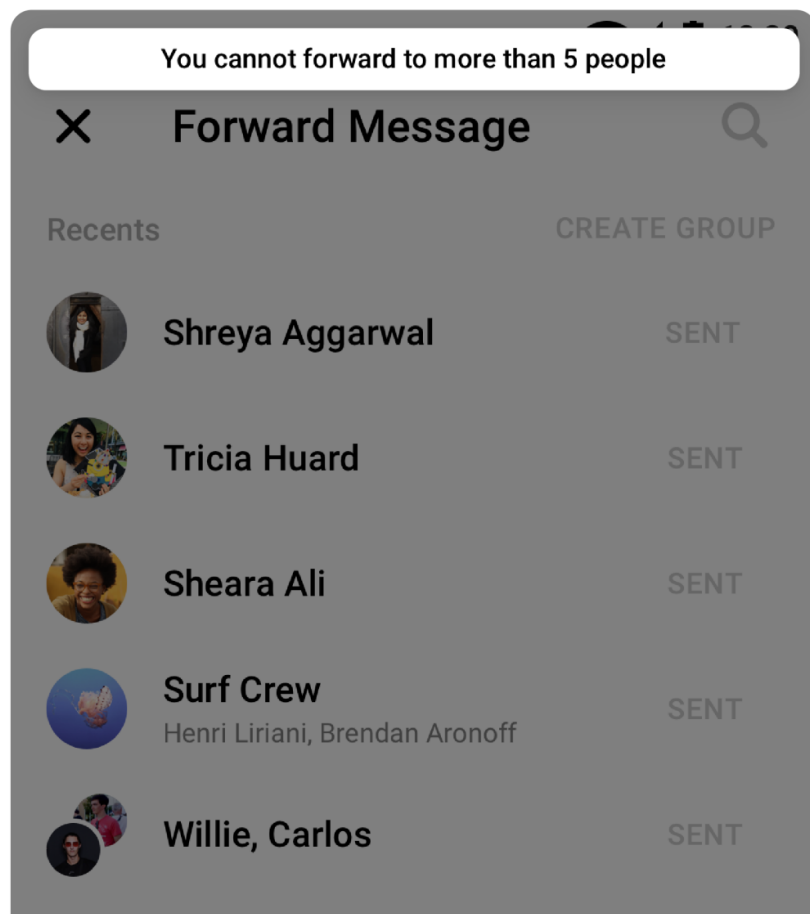
algorithms. Without this context-specific data, we risk losing language nuances that affect accuracy.

Globally, when it comes to misinformation, we reduce the spread of content that's been deemed false by third-party fact-checkers. But in countries with fragile information ecosystems, false news can have more serious consequences, including violence. That's why last year we updated our global violence and incitement policy such that we now remove misinformation that has the potential to contribute to imminent violence or physical harm. To enforce this policy, we partner with civil society organizations who can help us confirm whether content is false and has the potential to incite violence or harm.



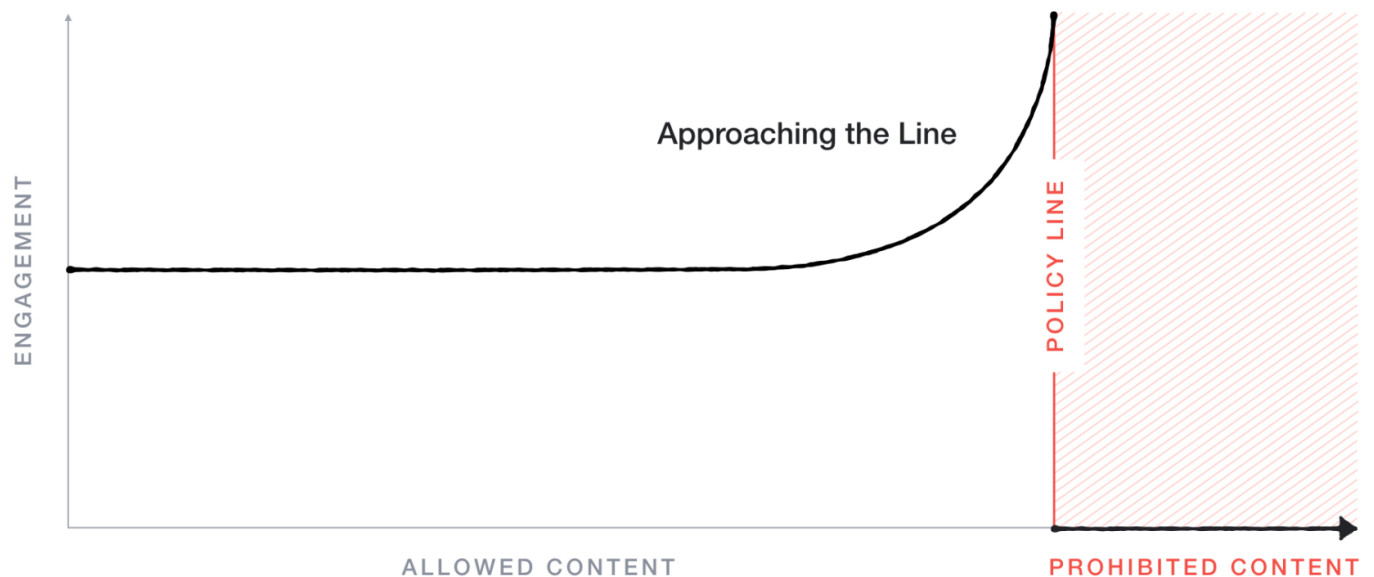
Reducing Misinformation and Borderline Content

We're also making fundamental changes to our products to address virality and reduce the spread of content that can amplify and exacerbate violence and conflict. In Sri Lanka, we have explored adding friction to message forwarding so that people can only share a message with a certain number of chat threads on Facebook Messenger. This is similar to a change we made to WhatsApp earlier this year to reduce forwarded messages around the world. It also delivers on user feedback that most people don't want to receive chain messages.

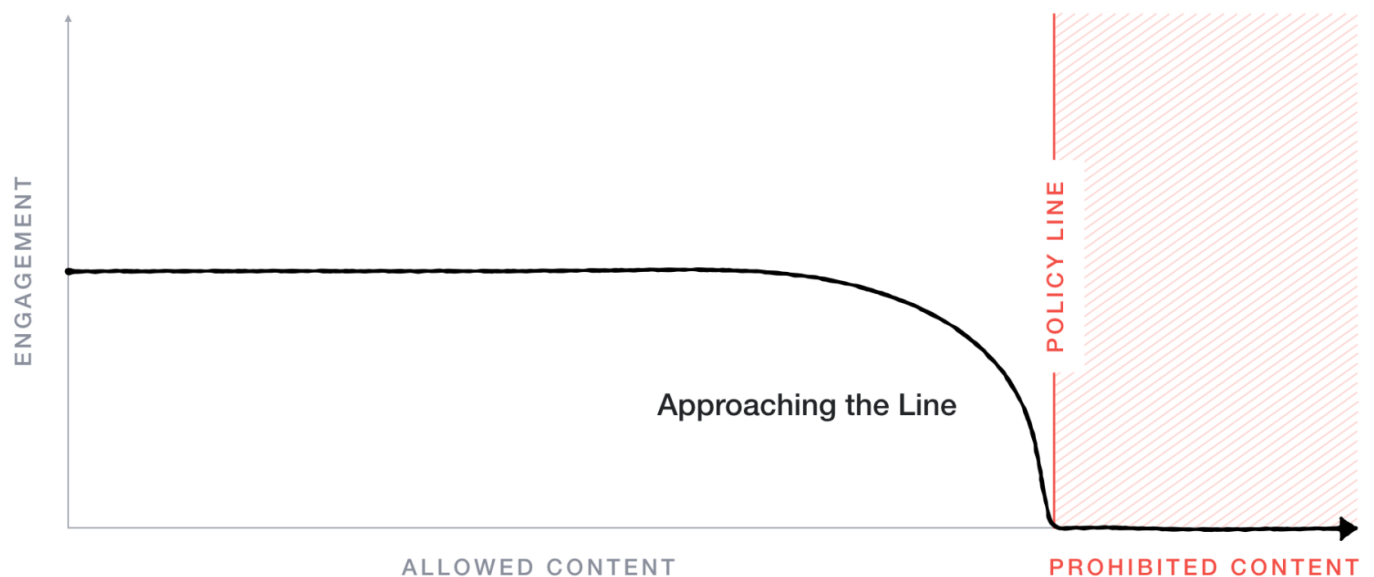


And, as our CEO Mark Zuckerberg detailed last year, we have started to explore how best to discourage borderline content, or content that toes the permissible line without crossing it. This is especially true in countries experiencing conflict because borderline content, much of which is sensationalist and provocative, has the potential for more serious consequences in these countries.

Natural Engagement Pattern



Adjusted to Discourage Borderline Content



We are, for example, taking a more aggressive approach against people and groups who regularly violate our policies. In Myanmar, we have started to reduce the distribution of all content shared by people who have demonstrated a pattern of posting content that violates our Community Standards, an approach that we may roll out in other countries if it proves successful in mitigating harm. In cases where individuals or organizations more directly promote or engage violence, we will ban them under our policy against dangerous individuals and organizations. Reducing distribution of content is, however, another lever we can pull to combat the spread of hateful content and activity.

We have also extended the use of artificial intelligence to recognize posts that may contain graphic violence and comments that are potentially violent or dehumanizing, so we can reduce their distribution while they undergo review by our Community Operations team. If this content violates our policies, we will remove it. By limiting visibility in this way, we hope to mitigate against the risk of offline harm and violence.



Giving People Additional Tools and Information

Perhaps most importantly, we continue to meet with and learn from civil society who are intimately familiar with trends and tensions on the ground and are often on the front lines of complex crises. To improve communication and better identify potentially harmful posts, we have built a new tool for our partners to flag content to us directly. We appreciate the burden and risk that this places on civil society organizations, which is why we've worked hard to streamline the reporting process and make it secure and safe.

Our partnerships have also been instrumental in promoting digital literacy in countries where many people are new to the internet. Last week, we announced a new program with GSMA called Internet One-on-One (1O1). The program, which we first launched in Myanmar with the goal of reaching 500,000 people in three months, offers one-on-one training sessions that includes a short video on the benefits of the internet and how to stay safe online. We plan to partner with other telecom companies and introduce similar programs in other countries. In Nigeria, we introduced a 12-week digital literacy program for secondary school students called Safe Online with Facebook. Developed in partnership with Re:Learn and Junior Achievement Nigeria, the program has worked with students at over 160 schools and covers a mix of online safety, news literacy, wellness tips and more, all facilitated by a team of trainers across Nigeria.

What's Next

We know there's more to do to better understand the role of social media in countries of conflict. We want to be part of the solution so that as we mitigate abuse and harmful content, people can continue using our services to communicate. In the

wake of the horrific terrorist attacks in Sri Lanka, more than a quarter million people used Facebook’s Safety Check to mark themselves as safe and reassure loved ones. In the same vein, thousands of people in Sri Lanka used our crisis response tools to make offers and requests for help. These use cases — the good, the meaningful, the consequential — are ones that we want to preserve.

This is some of the most important work being done at Facebook and we fully recognize the gravity of these challenges. By tackling hate speech and misinformation, investing in AI and changes to our products, and strengthening our partnerships, we can continue to make progress on these issues around the world.

Categories:

Integrity and Security, Meta, Public Policy,
Safety and Expression



Tags:

Community Standards and Enforcement,
Digital Literacy, False News

RELATED NEWS

Meta

An Update on Our Longstanding Work to Protect People in Ethiopia

For years we’ve invested in safety measures in Ethiopia, and recently we’ve taken additional steps to address violating content across our apps and help keep people safe.

November 9, 2021

Topics

Company News

Technology and Innovation

Data and Privacy

Safety and Expression

Combating Misinformation

Economic Opportunity
Election Integrity
Strengthening Communities
Diversity and Inclusion

Featured News

Meta

Investing in Canadian Talent to Help Build for the Metaverse

March 29, 2022

Meta

Sharing the Spirit of Ramadan Through #MonthofGood

March 28, 2022



Follow Us



Who We Are



What We Build



Our Actions



Our Community

▼

Resources

▼

Help Center

▼

Community Standards

|

Data Policy

Terms

|

Cookie Policy

United States (English) ▼