MediaWell

Article

Extreme Speech in Encrypted Messaging: Recommendations for Holistic Policy

September 5, 2025

This report is a summary of the findings of the 2024–2025 research group on encrypted messaging and extreme speech at the Center for Advanced Studies, LMU Munich, written by Anmol Alphonso, Sérgio Barbosa, Cayley Clifford, Kiran Garimella, Elonnai Hickok, Martin Riedl, Erkan Saka, Herman Wasserman and Sahana Udupa. Sahana Udupa is the corresponding author.

Online platforms are increasingly politicized spaces. Questions regarding what speech is permissible, who decides, on what basis, and who assumes responsibility for the harms emerging from online communication are constantly under debate. Platforms have become important means of social and political communication and have democratized information sharing, lowering the barriers to who can speak and be heard. But they have also allowed for the large-scale circulation of "extreme speech" — or what Udupa defines as "speech acts that stretch the boundaries of legitimate speech along the twin axes of truth/falsity and civility/incivility."

Encrypted messaging platforms like WhatsApp or Telegram are particularly significant in extreme speech ecosystems: they are deployed to entrench hierarchies, legitimize false information and conspiracy theories, and weaponize online discourse. These harms can be difficult to balance against the benefits of encrypted platforms, like the opportunities they offer for civic mobilization, journalistic practices, and wide-ranging social interactions.

Recent regulatory directions around the world have sought to break open encryption by creating backdoors. The European Commission's ProtectEU initiative aims to give law enforcement legal access to encrypted online data. The digital rights community has resisted the proposal, arguing that weakening end-to-end encryption will ultimately undermine cybersecurity. On June 24, 2025, the European Commission presented a Roadmap outlining a plan to ensure that law enforcement can access necessary data. It further commits to developing an encryption-specific Technology Roadmap by 2026, which will identify and assess solutions enabling lawful access to encrypted data by law

enforcement, while protecting cybersecurity and fundamental rights.

In 2024, the UK government expanded its surveillance powers under the Investigatory Powers Act to include the ability to demand access to encrypted data. It issued one such "Technical Capacity Notice" to Apple in early 2025 regarding the company's Advanced Data Protection (ADP) feature, which provides end-to-end encryption for iCloud. Rather than create backdoor access for law enforcement, Apple withdrew ADP from the UK market to protect user privacy. This legislation followed 2023's UK Online Safety Act, which requires online messaging platforms to ensure that they can apply state-accredited technologies to identify and remove harmful content on encrypted channels if ordered to do so. Companies and civil society organizations pushed back on both regulations during their draft stages, stating that such requirements would undermine safety and privacy that encryption promises. Indeed, shortly before the passage of the Online Safety Act, the UK government admitted that the "technology needed to securely scan encrypted messages sent on WhatsApp and Signal does not exist." Last month, the UK government walked back its demand, reportedly following pressure from the Trump administration.

Governments have also used legal action against employees to enforce compliance. In Brazil, authorities arrested WhatsApp executives for refusing to provide user data, and the messaging service was banned from 2015 to 2016. Similar strategies have been observed in Uganda and Zambia, where access to online platforms was blocked during elections. Countries around the world are expanding legal tools and actions to limit encryption.

Platforms have responded with legal challenges to governmental measures, while simultaneously curtailing responsible content moderation measures and the resources needed to implement them. On January 7, 2025 – in a statement that sent shockwaves to fact-checkers and civil society groups around the world – Meta announced that it would remove fact-checkers across its services in the United States, replacing them with a crowdsourced system based on user-driven consensus, known as "Community Notes." It also announced its intention to simplify content policies, including the removal of hate speech restrictions on categories related to gender identity and immigration status. While these measures do not directly impact the company's encrypted messaging services, they set a precedent for further reduction in platform oversight. Incidents of violence linked to rumors, disinformation, and conspiracy theories on encrypted messaging apps have stressed the need for urgent policy actions.

In this policy report, we argue that *existing debates around regulation, moderation, and policy need to address the broader political ecosystem of extreme speech and disinformation,* as well as measures that account for contextual realities. We caution against indiscriminately targeting encryption and suggest that such measures can undermine safety

and security of encryption for ordinary users. We also highlight grave issues in platform measures and content moderation practices. We offer several recommendations to make online encrypted messaging platforms safe and secure for users, rooted in international human rights principles and the protection of democratic values.

This report focuses primarily on the case of WhatsApp; as the world's largest encrypted messaging platform, a grounded understanding of the challenges it poses is the first step toward context-sensitive policy and regulation.

Key Challenges

While the technical architecture and governance of online encrypted platforms influence the online space, they by no means determine how encrypted platforms are used. Long-standing structures of power, social habits, and political cultures are intertwined with technological architectures and corporate policies, resulting in what Udupa and Wasserman define as "lived encryptions."

Contradictions

The technological features of WhatsApp promise privacy and secure communication. But the actual use and applications on the ground are often suffused with contradictions: in conflict settings and ordinary law enforcement contexts alike, the safety of a WhatsApp conversation is not a taken-for-granted condition, and privacy safeguards can be swiftly overturned by authoritarian and surveilling governments. For example, as Schumann explores in the case of Cameroon state actors routinely seize phones from suspected Anglophone dissenters to inspect data. Such measures do not require sophisticated encryption-breaking techniques. Incidents of coercion have been reported in India, where local police have been accused of using extrajudicial tactics to pressure people to reveal their private WhatsApp chats.

Family and trust-based networks

WhatsApp's influence in Global South contexts has emerged from the deep inroads the platform has made into local community networks, family groups, and social relations that are seen by their members as trustworthy. Saka observes that WhatsApp is seen as more familial compared to other platforms in Turkey. Political actors have expanded campaign activities to WhatsApp groups to gain "organic" influence. Describing them as "deep extreme speech," Udupa shows that they contain "community-based distribution networks and a distinct context mix, which both build on the charisma of local celebrities, social trust, and everyday habits of exchange." They weave exclusionary messages with good morning greetings, religious hymns, local issues of water supply and other socially vetted and existentially relevant content.

Microtargeting and cross-media manipulation

Microtargeting occurs when WhatsApp messages are aimed at small groups through what Evangelista and Bruno identify as a centralized structure "built to manage and to stimulate members of discussion groups, which [are] treated as segmented audiences." This process creates complex flows of information that are germinated and fertilized across different WhatsApp groups and social media platforms. Garimella and colleagues have tracked how WhatsApp groups are linked to other social media platforms for political propaganda in India.

Fact-checking on WhatsApp

Encryption prevents fact-checkers from being able to find disinformation or extreme speech on the platform themselves. They rely instead on user-reported examples – which is complicated by the fact that WhatsApp users tend to trust the information they receive from friends, family, or colleagues. As a result, they are not always likely to verify or question the information they receive, or send it to fact-checkers for verification. Wasserman and Madrid-Morales show that young users hesitate to correct false information coming from elders on WhatsApp groups because of a sense of respect.

Once information is verified or debunked by fact-checkers, it does not always reach those who saw the original content and may still be unaware of its problematic nature. Even if it does, not everyone will believe fact checks — especially if the false information has a stronger emotional appeal. Users may continue to share false content if they are under the impression that doing so may be helpful to those in their networks and communities. While several fact-checking organizations have set up tiplines and other services for this purpose, practical considerations limit the potential impact of these efforts. "Zombie claims" — false information that will not die, no matter how many times it has been previously debunked — are a major challenge.

AI-generated content

While AI technologies are being explored for fact-checking and the automated dissemination of prosocial narratives, the broader impacts of generative AI on social media – including encrypted messaging platforms – are becoming increasingly evident. As AI becomes more accessible and user-friendly, individuals and groups with limited resources can create high-quality content that rivals that of well-funded organizations. This democratization of content creation could lead to a more diverse range of voices and perspectives on the platform. However, it also raises concerns about the spread of disinformation and extreme speech, as malicious actors may exploit the technology for their own agenda.

Recommendations

Platform governance

Government measures to enforce platform accountability should adhere to international human rights standards, including the principles of necessity, legality, and proportionality – avoiding the use of spyware or other norm-violating surveillance practices. Governments can commit to principles such as the Necessary and Proportionate Principles and the Freedom Online Coalition Principles on Government Use of Surveillance Technologies. Existing legal frameworks that provide remedies should also be strengthened to ensure due process for content removal and other moderation actions.

In line with the UN Guiding Principles on Business and Human Rights, platforms should conduct ongoing due diligence of their services across the markets they operate in to understand and address emerging risks to human rights in different contexts. Encrypted messaging platforms should participate in applying a contextually responsive, industry-wide code of conduct grounded in international human rights principles. Trust, safety, and human rights play important roles in developing and enforcing Terms of Service and content policies on platforms. Platforms should ensure they have robust teams in place that are funded and supported.

Metadata analysis and user reporting

Rather than requiring content moderation that would undermine encryption, governments and platforms should explore alternate interventions. Metadata analysis, for example, provides information on who sends or receives a piece of content, the type and size of files shared, etc., and can be done with the use of machine learning.We also recommend the development of stronger user reporting mechanisms to identify and address online harms.

Digital influence operations

While user reporting infrastructures should be improved, organized disinformation campaigns that purposefully misuse reporting as a way to overwhelm platform systems are not uncommon. The political weaponization of WhatsApp Channels and groups, microtargeting and segmentation, coordinated manipulation, and gender-based violence are constantly evolving on encrypted messaging platforms. Riedl and colleagues have shown that women and queer journalists experience “infrastructural platform violence on WhatsApp” in Lebanon.

Multiple stakeholders need to collaborate to address the vast networks of extreme speech and disinformation that commercial political consultants, political parties, and state actors

have created on encrypted messaging platforms – including WhatsApp – through the use of grey networks, clickbait operators, and digital influencers.

Platforms should conduct an assessment of systematic risks that arise from manipulative digital influence operations and take the appropriate steps for risk mitigation. Other measures include ensuring transparency in election expenditure, regulating campaign finance, and setting professional codes of conduct and co-regulatory models for digital influence operations.

Research

Governments should develop legal frameworks to promote researcher access to data, including data donation initiatives. Platforms should also provide researchers access to viral content, or content that has surpassed a predefined exposure threshold (like messages labelled as "forwarded many times.") This access could be facilitated through a public platform, empowering researchers and journalists to analyze and understand the dissemination of content on WhatsApp.

Increased support for fact-checking

Online encrypted platforms and the donor community should support fact-checkers' work through continued and strengthened collaboration. Platforms should develop dedicated fact-checking channels, or provide civil society organizations with the means and access to do so themselves. Such channels could share fact-checks, media and information literacy materials, and credible updates during critical events like elections.

Responsible AI use

Platforms and companies can also support fact-checkers by helping them to leverage AI to develop and share accessible, easily understandable fact-checked material, including through funding and technical expertise. Companies should invest in developing AI models that can work in multiple languages – especially minoritized languages – and provide free access to community moderators and fact-checkers. An AI-enabled reporting mechanism can be integrated into platforms for flagging harmful content in multiple languages.

Conclusion

At a time when platforms are rolling back trust and safety protocols, we call for stronger platform governance and content moderation – while *also* cautioning that removing encryption is not a solution to address extreme speech and disinformation. Instead, we recommend a contextualized approach to the governance of online encrypted messaging services, addressing different stakeholders, challenges, and opportunities.

Future interventions should focus on finding whole-of-society solutions to online harms and challenges. This work will require the support of UN entities and other multilateral agencies, as well as consulting with relevant expert groups, civil society, and the technical community. Collaboration among multiple stakeholders will be necessary to develop and implement technical and nontechnical solutions that are lawful, necessary, proportionate, and informed by expert opinion.

A full version of this policy report is available here; the book, WhatsApp in the World: Disinformation, Encryption and Extreme Speech *(New York University Press, 2025), can be accessed here.*

The research group on encrypted messaging and extreme speech (2024-2025) was supported by the Center for Advanced Studies, LMU Munich.