ENCRYPTED MESSAGING AND EXTREME SPEECH: POLICY DIRECTIONS

Anmol Alphonso, Sérgio Barbosa, Cayley Clifford, Kiran Garimella, Elonnai Hickok, Martin Riedl, Erkan Saka, Herman Wasserman & Sahana Udupa*

*Corresponding author: sahana.udupa@lmu.de

July 2025
DOI: https://doi.org/10.5282/ubm/epub.127250

Supported by

European Research Council
Established by the European Commission

Center for DIGITAL DIGNITY

CAS LMU

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CONTENT

Executive Summary ... **3**
Introduction and Context ... **6**
Mapping Online Encrypted Messaging Services ... **14**
Overview and Architecture of WhatsApp ... **15**
Current State of Regulation ... **18**
Challenges on Online Encrypted Platforms ... **26**
AI and Online Encrypted Platforms ... **34**
Case Studies ... **35**
India ... 35
South Africa ... 37
Brazil ... 40
Recommendations ... **44**
Data Access Tools and List of Resources ... **54**

EXECUTIVE SUMMARY

Online platforms, including encrypted messaging platforms, are increasingly politicized spaces. Questions regarding what speech is permissible, who decides this, on what basis, and who assumes responsibility for harms emerging from online communication are constantly debated.

These 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. However, they have also allowed for the large-scale circulation of extreme speech and disinformation.

In this report, we respond to this challenge by focusing on encrypted messaging platforms, using WhatsApp as a case study. WhatsApp is the largest messaging platform in the world and a critical communication infrastructure for many, particularly in the Global South. We approach WhatsApp as a platform located within structures of power, social habits and political cultures, and intertwined with technological architectures and corporate policies. The case studies of Brazil, India, and South Africa demonstrate how encrypted messaging platforms are deployed to entrench hierarchies, legitimize false information, and weaponize online discourse, while also offering opportunities for civic mobilizations, journalistic practices, and wide-ranging social interactions.

Building on a rich body of recent scholarship and policy debates, this report highlights the distinctive features of encrypted messaging platforms and the regulatory challenges they have posed. It begins with a brief outline of different encrypted messaging platforms, highlighting their distinctive encryption features, reach, and moderation structures. Subsequently, it situates WhatsApp in this "market place" of encrypted messaging platforms, outlining its unique architectural elements and the social characteristics of communication enabled by them. This is followed by an overview of the current state of regulation and moderation in the field of encrypted messaging.

The report also considers challenges concerning WhatsApp, emphasizing how 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. The report concludes with several recommendations to make online encrypted messaging platforms safe and secure for users as well as grounded in international human rights principles and the protection of democratic values. Throughout the report, we highlight

the broad human rights principles as enunciated by the United Nations (UN),[1] while noting that universalist normative language and regulatory discourses such as "hate speech" can be perceived as "foreign" values within various regional contexts or weaponized for repressive agendas domestically and for geopolitical domination.[2] We suggest that interventions should emerge from emic categories and lived language, in ways to uphold democratic conditions of belonging, safety, equity, and dignity.[3] Similarly, we note the instrumentalization of "free speech" and "freedom of speech" discourses for right-wing political agendas globally, paricularly recent developments in North America which may contribute to self-censorship and the withdrawal of funding for disinformation studies.[4]

The report highlights interventions in six main categories, namely:

- Platform governance
- Mitigating digital influence operations
- Supporting research
- Strengthening fact-checking
- Awareness raising and capacity building
- Leveraging artificial intelligence responsibly

The recommendations highlight the need for developing approaches that are grounded in lived realities of specific contexts and international human rights standards. They call for close knowledge of diverse and dynamic political and social practices that have emerged around encrypted messaging platforms, which often contradict promises of privacy and secure communication signaled by encryption technology as well as undermine regulatory efforts with the clever use of campaign tactics. These developments have been prominent in the Global South, as this report highlights.

1 https://www.un.org/en/global-issues/human-rights (accessed 10 February 2025).

2 Critical scholarship has pointed out that a distinctive liberal discourse of human rights in the West, for instance, has sought to depoliticize issues of inequality and marginalization [see Whyte J (2019) *The Morals of the Market: Human Rights and the Rise of Neoliberalism.* London/New York: Verso]; also, Udupa S and Pohjonen M (2019) Extreme speech and global digital cultures. *International Journal of Communication* 13: 3049–3067.

3 Udupa S. Digital technology and extreme speech: Approaches to counter online hate. United Nations Peacekeeping. p. 10. Available at: https://peacekeeping.un.org/sites/default/files/digital_technology_and_extreme_speech_udupa_17_sept_2021.pdf (accessed 31 October 2021).

4 https://www.whitehouse.gov/presidential-actions/2025/01/restoring-freedom-of-speech-and-ending-federal-censorship/ (accessed 10 March 2025).

At a time when platforms are rolling back trust and safety protocols, this report serves as yet another call to take platform governance and content moderation seriously while also cautioning that removing encryption is not a solution to address extreme speech and disinformation. We call for a contextualized approach to the governance of online encrypted messaging services, addressing different stakeholders, challenges, and opportunities. Multiple stakeholders, with the support of UN entities and other multilateral agencies, should focus on finding whole-of-society solutions to online harms and challenges. This means working with relevant expert groups, civil society, and the technical community to develop and implement technical and nontechnical solutions which are lawful, necessary, proportionate, and informed by expert opinion.

For a global analysis of the interplay between encrypted messaging and extreme speech, see the open-access book, "WhatsApp in the World: Disinformation, Encryption and Extreme Speech", New York University Press, 2025, edited by Sahana Udupa and Herman Wasserman.

INTRODUCTION AND CONTEXT

On 7 January 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 (US), replacing them with a crowdsourced system based on user-driven consensus, known as "Community Notes".[5] It also announced its intention to simplify content policies, including removing restrictions on categories such as gender and immigration, and move "Trust & Safety" teams from the Democrat-led state of California in the US to the Republican stronghold of Texas.

Meta owner Mark Zuckerberg framed the shift as a way to counter online censorship, protect freedom of expression, and enable innovation.[6] As part of this shift, Meta introduced significant changes to its "policy rationale". A sentence on hate speech which used to read: "That is why we don't allow hateful speech on Facebook, Instagram, or Threads. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence" was shortened to: "That is why we don't allow hateful conduct on Facebook, Instagram, or Threads."[7] Experts have criticized the move, cautioning that these spaces will become less safe with instances of disinformation and hateful speech circulating on the platform.[8] If these changes are extended beyond the US, the impact is likely to be particularly strong in the Global South, where a "dangerous void in frontline defences against disinformation" will leave "vulnerable communities at the mercy of unchecked narratives".[9] Prior to Meta's decision, X made similar changes to its platform, replacing fact-checkers with "Community Notes".[10]

Simultaneously, state actors have been active in seeking control over social media companies and encrypted messaging platforms. The European Commission's ProtectEU initiative aims to give law enforcement legal access to encrypted online data, including through the use of encryption backdoors. On June 24, 2025, the European Commission presented a Roadmap outlining a plan to ensure law enforcement can access necessary data. Among other things, the Roadmap commits to developing, by 2026, a Technology Roadmap on encryption that will

5 https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ (accessed 15 January 2025).

6 Kleinman Z (2025) Meta to replace 'biased' fact-checkers with moderation by users. BBC. 7 January. Available at: https://www.bbc.com/news/articles/cly74mpy8klo

7 Meta "Hateful Conduct" available at: https://transparency.meta.com/zh-tw/policies/community-standards/hate-speech/; "Current version" in comparison with 7 January 2025 version (accessed 16 January 2025).

8 Fraser G (2025) Huge problems with axing fact-checkers, Meta oversight board says. BBC. 8 January. Available at: https://www.bbc.com/news/articles/cjwlwlqpwx7o. Kayyali D (2025) Meta's content moderation changes are going to have a real world impact. It's not going to be good. Tech Policy.Press. Available at: https://www.techpolicy.press/metas-content-moderation-changes-are-going-to-have-a-real-world-impact-its-not-going-to-be-good/ (accessed 10 January 2025).

9 Divon T and Ong JC (2025) Tech bro power play: Zuckerberg vs. global tech justice. TechPolicy.Press. Available at: https://techpolicy.press/tech-bro-power-play-zuckerberg-vs-global-tech-justice (accessed 14 January 2025).

10 Schleifer T, Isaac M, Conger K, Kang C, Grant N, Satariano A and Kaye D (2025) Meta to end fact-checking on Facebook, Instagram ahead of Trump term: Live updates. The New York Times. Available at: https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking.

identify and assess solutions enabling lawful access to encrypted data by law enforcement, while protecting cybersecurity and fundamental rights.[11] Countries around the world are expanding legal tools and actions to limit encryption. [12] In January 2025, South Sudan banned social media platforms, including Facebook and Tiktok, on grounds of preventing violence in the country.[13] In November 2024, Pakistan blocked WhatsApp, X, Instagram, and Facebook in response to anti-government protests.[14] In 2015, two journalists from *Vice News* in Turkey and their translator were arrested in the southwestern region while covering conflict near the Syrian border.[15] Turkish authorities charged them with aiding a terrorist organization, citing the possession of encryption software as evidence.

The move by Meta and other platforms, as well as surveillance and coercive actions by state authorities, demonstrate how online platforms, including encrypted messaging platforms, are increasingly politicized spaces. This has raised questions about what speech is permissible, who makes this determination, and on what basis, as well as the availability of resources and infrastructures for oversight and moderation.

The fluctuating positions and back-and-forth exchanges among regulators, companies, and users over permissible speech are not new but part of a recurring debate about the architecture, design, and governance of social media platforms and questions about who assumes responsibility for harms that emerge through the use of such platforms.

Online platforms have become important means of social and political communication, enabling the sharing of information across medium types, at scale, and in real-time. This has democratized information sharing, lowering the barriers of who can speak and be heard, but has also allowed for large-scale circulation of extreme speech and disinformation. A 2023 report by the Brazilian federal government, for example, states that hate speech on the internet increases particularly during the electoral periods and among students at school level.[16] In the UN peacekeeping contexts in the Central African Republic, research findings show that "strategic influence operations" on social media that use propaganda, rumors, and facts "are multi-directional and participatory, involving interactions among state actors, influencers

11 https://home-affairs.ec.europa.eu/news/commission-presents-roadmap-effective-and-lawful-access-data-law-enforcement-2025-06-24_en

12 https://www.gp-digital.org/world-map-of-encryption/

13 https://www.theeastafrican.co.ke/tea/news/east-africa/south-sudan-shuts-down-social-media-for-three-months-4898822#google_vignette

14 https://www.techradar.com/pro/vpn/whatsapp-becomes-the-latest-social-media-app-blocked-in-pakistan

15 https://www.theregister.com/2015/09/02/turkey_terror_arrests/ (accessed 22 January 2025).

16 https://www.gov.br/mec/pt-br/acesso-a-informacao/participacao-social/grupos-de-trabalho/prevencao-e-enfrentamento-da-violencia-nas-escolas/resultados (accessed 23 January 2025).

and ordinary individuals."[17]

Specific incidents of physical violence have revealed the role of social media posts in heightening the tensions. The United Kingdom (UK) regulator Ofcom found a clear connection between online social media posts and increased violence during riots in Southport in England in 2024.[18] In a case from Kolhapur, in India's western state of Maharashtra, violence broke out after WhatsApp statuses were used by right-wing groups to rally forces and organize a protest in June 2023, targeting the Muslim community in the city.[19] The wide-ranging consequences of social media communication in different parts of the world highlight the challenges in the current information environment. Incidents of harm and violence that are linked to social media discourses, as well as the instability and unevenness of platform policies, signal that the regulatory question has become more complex.

In this report, we respond to this challenge by focusing on encrypted messaging, a unique node in the fraught environments of political speech and information. Among social media platforms, a broad distinction can be drawn between social media platforms, which have emerged as public-facing spaces that allow individuals to have unique profiles and send bulk messaging to a broad audience, and encrypted messaging platforms, which have traditionally focused on facilitating private communication between individuals, protected through technologies like encryption. End-to-end encryption ensures that the conversation between the sender and the receiver is not tracked or accessed by platforms or other third parties, although the extent of encryption varies across different messaging platforms.

Encryption has been a point of tension, both enabling privacy and freedom of expression and making it more challenging for third parties, including regulators, law enforcement, and academics, to access content and understand how messaging platforms are used. Civil society and government interests are often misaligned on the question of encryption and online messaging platforms: governments argue for more access and so-called "backdoors" that would allow them to access content,[20] whereas online privacy advocates and civil society groups argue that any backdoor would fundamentally alter and ultimately eliminate the protections of end-to-end encrypted communications.[21] These so-called "crypto wars" are not a new phenomenon; for as long as different forms of encryption have been

17 Miyashita N *et al.* (2024) How strategic information operations affect peacekeeping: Two case studies from the Central African Republic. *International Peacekeeping,* 0(0), pp. 1–43. Available at: https://doi.org/10.1080/13533312.2025.2470342

18 Desmarais A (2024) Clear connection between social media posts and violence during Southport riots, UK regulator finds. Euro News. October 23. Available at: https://www.euronews.com/next/2024/10/23/clear-connection-between-social-media-posts-and-violence-during-southport-riots-uk-regulat

19 Niranjankumar N (2023) Kolhapur violence: An Instagram story and a flurry of WhatsApp posts. BOOM. 12 June. Available at: https://www.boomlive.in/decode/kolhapur-violence-aurangzeb-maharashtra-bjp-shivaji-whatsapp-instagram-communal-22221

20 Heemsbergen L and Maddox A (2022) Distributing journalism: Digital disclosure, secrecy, and crypto-cultures. In Filimowicz M (ed) Privacy: *Algorithms and Society.* London: Routledge. pp. 1-29.

21 Pfefferkorn R (2020) Client-side scanning and Winnie-the-Pooh redux (plus some thoughts on Zoom). The Center for Internet and Society at Stanford University. Available at: https://cyberlaw.stanford.edu/blog/2020/05/client-side-scanning-and-winnie-poohredux-plus-some-thoughts-zoom

around, intelligence services have tried to find ways in which they could still access information.[22] Regulatory debates about discontinuing end-to-end encryption are often framed around critical platform governance issues, such as the moderation of child sexual abuse material (CSAM) or terrorist content.[23] On the other side, sections of civil society argue that end-to-end encryption is one of the last remnants of user privacy on the internet and key to protecting freedom of speech. They highlight the importance of encrypted spaces for enabling users to access and share critical information anonymously and protecting dissenters, victims, journalists, civil society, and minority groups and populations. They advocate for solutions that focus on addressing the societal challenge of harms like CSAM and measures that are based on principles of necessity, proportionality, and legality.[24]

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 is defined as "lived encryptions".[25]

> "Lived encryption stresses that encryption as a technological feature cannot be taken at face value…rather, it embeds different, often contradictory, social and political formations and interactions.[26]"
>
> – Udupa and Wasserman, 2025

While encryption promises confidentiality, there have been incidents where state actors and law and order officials have seized cell phones to inspect conversations.[27] Similarly, political

22 Rider K (2018) The privacy paradox: How market privacy facilitates government surveillance. *Information, Communication & Society* 21(10) 1369-1385. Also https://www.economist.com/international/2024/09/05 how-encrypted-messaging-apps-conquered-the-world

23 Duan C and Grimmelmann J (2024) Content moderation on end-to-end encrypted systems: A legal analysis. 8(1) *Georgetown Law Technology Review* 1–92; Chousou S, Magaud J, Pavoni L and Williams M (2023) Is encryption a fundamental right? A case study on CSAM regulation in the EU. Sciences Po. Available at: https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2023/07/Encryption.pdf; Wasserman S (2024) Why end-to-end encryption is the next battlefield for tech justice. The Hill. 22 February. Available at: https://thehill.com/opinion/technology/4482935-why-end-to-end-encryption-is-the-next-battlefield-for-tech-justice/; Hartel P and van Wegberg R (2023) Going dark? Analysing the impact of end-to-end encryption on the outcome of Dutch criminal court cases. *Crime Science* 12(5).

24 Nojeim G and Knodel M (2021) CDT welcomes rollout of encryption-by-default for Facebook Messenger. Center for Democracy and Technology. Available at: https://cdt.org/insights/cdt-welcomes-rollout-of-encryption-by-default-for-facebook-messenger/#:~:text=The%20extension%20of%20E2EE%20to,the%20contents%20of%20a%20message; Mullin J (2024). Now the EU Council should finally understand: No one wants "chat control". *Electronic Frontier Foundation*. Available at: https://www.eff.org/deeplinks/2024/06/now-eu-council-should-finally-understand-no-one-wants-chat-control

25 Udupa S and Wasserman H (2025) *WhatsApp in the World: Disinformation, Encryption and Extreme Speech.* New York: New York University Press.

26 Udupa and Wasserman (2025) *WhatsApp in the World*, p. 6.

27 Schumann K (2025) Delete this message: Media practices of Anglophone Cameroonian WhatsApp users in the face of counterterrorism. In Udupa and Wasserman (2025) *WhatsApp in the World*, pp 111-125.

groups have utilized closed-group communication features to broadcast top–down messages.[28] Such uses are facilitated by the platform's group chat functionality and the introduction of "Channels" for broadcasting messaging to large audiences. Importantly, encryption features do not pave the way for unhindered safety and security. Especially within conflict and authoritarian environments, users navigate tense situations of safe communication and intrusive surveillance.[29] The case studies of Brazil, India, and South Africa demonstrate how encrypted messaging platforms have raised the tension between safe and unsafe online spaces, especially how they are deployed to entrench hierarchies, legitimize false information, and weaponize online discourse, while also offering opportunities for civic mobilizations, journalistic practices, and wide-ranging social interactions.

The report uses the definition of "extreme speech" to approach different types of politically contentious content and examine their policy implications for encrypted messaging platforms.[30] They include "derogatory extreme speech"[31] —extreme expressions aimed at any group, including those holding power; "exclusionary extreme speech"[32] —expressions that implicitly or explicitly exclude or cause harm to a person or a group on the basis of their group belonging; and "dangerous speech"[33] —expressions that have reasonable chances to trigger or cause harm and violence against target groups (see Table 1, p. 11). In terms of exclusionary extreme speech, the analysis builds on existing definitional standards around hate speech developed by the UN[34] and the distinction between disinformation ("when false information is knowingly shared to cause harm") and malinformation ("when genuine information is shared to cause harm").[35] Extreme speech analysis covers misinformation (spreading false information without the intention to cause harm) "so far as it is part of the social fields where deliberate efforts to spread hate activate a variety of actors and networks that end up spreading hateful language that could cause harm to vulnerable groups".[36] The report grounds its approach in the information integrity framework set out by the UN[37] and the boundaries of permissible speech

28 See the section on "Challenges" in this report.

29 Udupa and Wasserman (2025) *WhatsApp in the World*, p. 6.

30 This report does not cover child sexual abuse and other forms of illegal content.

31 Udupa (2021) Digital technology and extreme speech, p. 10.

32 Udupa (2021) Digital technology and extreme speech, p. 10.

33 Benesch S (2012) Dangerous speech: A proposal to prevent group violence. New York: World Policy Institute.

34 United Nations (2020) United Nations strategy and plan of action on hate speech: A detailed guidance on implementation for United Nations field presences. https://www.un.org/en/genocide-prevention/hate-speech/strategy-plan-action The UN definition identifies hate speech as "Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive".

35 United Nations Department of Peace Operations & Office on Genocide Prevention and the Responsibility to Protect. 2024. A conceptual analysis of the overlaps and differences between hate speech, misinformation, and disinformation. Available at: https://peacekeeping.un.org/sites/default/files/report_-_a_conceptual_analysis_of_the_overlaps_and_differences_between_hate_speech_misinformation_and_disinformation_june_2024_qrupdate.pdf (accessed 11 March 2024).

36 Udupa (2021) Digital technology and extreme speech.

37 https://www.un.org/en/information-integrity/global-principles (accessed 11 March 2024).

A FRAMEWORK TO DEFINE CONTENTIOUS CONTENT

Type of extreme speech	Definition	Recommended actions
Derogatory extreme speech	Expressions that do not conform to accepted norms of civility within specific regional/local/national contexts and target persons/groups based on racialized categories or protected characteristics (caste, ethnicity, gender, language group, national origin, religious affiliation, sexual orientation) as well as other groups holding power (state, media, politicians).[38] It includes derogatory expressions not only about people but also about abstract categories or institutions that they identify targeted groups with. It includes varieties of expressions that are considered within specific social-cultural-political contexts as "the irritating, the contentious, the eccentric, the heretical, the unwelcome, and the provocative, as long as such speech…[does]…not tend to provoke violence".[39]	Closer inspection and downranking, counter speech, monitoring, redirection, and awareness raising but not necessarily removal of content.
Exclusionary extreme speech	Expressions that call for or imply exclusion of historically disadvantaged and vulnerable people/groups from the "in-group" based on caste, ethnicity, gender, language group, national origin, racialized categories, religious affiliation, and/or sexual orientation. These expressions incite discrimination, abhorrence, and delegitimization of targeted groups. The label does not apply to abstract ideas, ideologies, or institutions, except when there are reasonable grounds to believe that attacks against ideas/ideologies/institutions amount to a call for or imply exclusion of vulnerable groups associated with these categories. For example, if attacking a particular religion in a specific context has a reasonable chance to incite hatred and exclusion of people who practice this religion, such expressions would fall under 'exclusionary extreme speech'. In terms of exclusionary extreme speech, the analysis builds on existing definitional standards around hate speech set up by the United Nations.[40]	Closer inspection and possible removal.
Dangerous speech	Dangerous speech refers to expressions that have a reasonable chance to trigger/catalyze harm and violence against target groups (including ostracism, segregation, deportation, and genocide).[41]	Immediate removal.

Table 1: *Definitions of types of extreme speech and recommended moderation actions*

Table reproduced from: Udupa S, Maronikolakis A and Wisiorek A (2023) Ethical scaling: Extreme speech and the (in)significance of artificial intelligence. Big Data & Society 1–15, DOI: 10.1177/20539517231172424

38 Udupa S (2018) Gaali cultures: The politics of abusive exchange on social media. New Media & Society 20(4): 1506–1522.

39 Redmond Bate vs Director of Public Prosecutions before the Lord Justice Sedley and Justice Collins on July 23, 1999; The Times, July 28, 1999.

40 United Nations (2020) United Nations strategy and plan of action on hate speech. Available at: www.digitallibrary.un.org (accessed 10 August 2021).

41 Benesch S (2012) Dangerous speech: A proposal to prevent group violence. New York: World Policy Institute.2

guided by international law, which clarify that only the most serious forms of hate speech amounting to incitement are prohibited.[42] Considering its vast societal implications, regular, systematic inquiries into encrypted messaging are necessary. However, accessing and analyzing the data on messaging platforms to inform evidence-based policies remains a challenge. While developments in the European Union (EU) such as the Digital Services Act (DSA) now require so-called "Very Large Online Platforms and Search Engines" to provide researchers access, this is limited to research focused on systemic risks within the boundaries of the EU, and the regulatory structure has yet to be implemented, raising questions of how effective it will be in practice.[43] Data sharing programs implemented by companies or mediated by third parties have also had their challenges with completeness, reliability, and interoperability of data.[44]

Building on a rich body of recent scholarship and policy debates[45] as well as the diverse research fields and practitioner experiences of the authors, this report highlights the distinctive features of encrypted messaging platforms and the regulatory challenges they pose to suggest how different entities, such as the UN and multilateral organizations, fact-checkers, civil society and nongovernmental organizations, platforms, and state regulators, can respond. Our specific focus is on WhatsApp—the most widely used messaging platform. The report examines the role of WhatsApp in facilitating extreme speech and disinformation, including its instrumentalization in organized hate and disinformation campaigns by political actors, while also documenting how the platform has been used by fact-checkers and journalists for prosocial and progressive discourses.

The report begins with a brief outline of the different encrypted messaging platforms, highlighting their distinctive encryption features and moderation structures. Subsequently, it situates WhatsApp in this "marketplace" of encrypted messaging platforms, outlining its unique architectural elements and the social characteristics of communication enabled by them. This is followed by an overview of the current state of regulation in the field of encrypted messaging, covering different regulatory measures that have been implemented across different jurisdictions, including outright bans on encryption, limitations on the strength of encryption, weakening of encrypted technologies, requirements for traceability of users, requirements for backdoors to be built into services and products to enable government access to information, and mandates for proactive monitoring of encrypted content.

42 Prohibited speech includes "direct and public incitement to commit genocide" and "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence". See the Rabat six-part test, https://www.ohchr.org/en/freedom-of-expression (accessed 11 March 2024).

43 https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (accessed 15 January 2024).

44 Van Drunen MZ and Noroozian A (2024) How to design data access for researchers: A legal and software development perspective. Computer Law & Security Report. 52. doi: https://doi.org/10.1016/j.clsr.2024.105946

45 For paucity of space, we are not able to cover all relevant studies but a select few to foreground specific points.

The next section discusses prominent features of the social use and political deployment of WhatsApp—primarily based on examples from Brazil, India, and South Africa—to highlight the challenges of addressing harms on encrypted messaging platforms.

A closer exploration of WhatsApp use in these three countries illustrates the ground realities of extreme speech ecosystems and diverse social opportunities of WhatsApp affordances and practices. Recent developments around artificial intelligence (AI) generated content and bot activities on WhatsApp signal new regulatory challenges.

Perusing the opportunities and risks of various regulatory measures against the vast social complexity and political misuse of WhatsApp, the report offers policy recommendations addressing platform governance, digital influence operations, research, fact-checking, capacity building, and AI. The report also provides an indicative, non-exhaustive list of resources for future research and policymaking. The "toolkit" includes a description of an open-source tool to access and analyze WhatsApp data in a privacy-preserving manner and URL links to various initiatives that have offered technical tools for fact-checking, support for researchers who face harassment, and civil society initiatives against disinformation and hateful communication on encrypted messaging. It also includes a framework to define and assess problematic speech, as often one of the vexing issues is defining what is and isn't problematic speech (see Table 1, p. 11).

MAPPING ONLINE ENCRYPTED MESSAGING SERVICES

There is a growing ecosystem of encrypted messaging platforms available to users, with a variety of services offered across different jurisdictions. The type of encryption used—whether all services are end-to-end encrypted by default, whether the platform backs up data in the cloud automatically or when a user opts-in, and the type of metadata the platform collects—impact the level of security provided to the user and the extent to which companies may comply with government orders to provide access to content or user information.[46]

Platforms also have varying terms of services, acceptable use and privacy policies, and approaches to enforcement. Civil society organizations, such as Ranking Digital Rights, support user choice and other advocacy efforts by ranking the transparency of factors related to privacy and freedom of expression, including platform encryption practices.[47] Civil society and academia have also worked to compare and contrast online encrypted platforms, highlighting differences and similarities between platform architecture and use. Whereas WhatsApp is used for a range of social, cultural, and political activities in the Global South based on its closed chat and group functionalities, Telegram's design prioritizes content availability on different devices, and its communication channels are not end-to-end encrypted by default.[48] Telegram has also emerged as a haven for extremist groups deplatformed on other social media platforms, individuals on the far right, and "hate influencers".[49] Yet, Telegram is often discussed in the same context as WhatsApp and Signal as being a privacy-preserving platform. As another example, WeChat focuses on interaction with service and official accounts and automated monitoring is built into the platform architecture, aligning with Chinese regulations.[50] Among privacy enthusiasts, Signal distinguishes itself as the standard purveyor of secure communications.[51]

Several internal and external forces, actors, and processes shape content, communication, and user experience on a platform and influence how companies design and govern their services. Internal factors include resource allocation, company values, company policy

46 Maheshwari N (2024) Encryption and human rights: what you should know. Access Now. Available at: https://www.accessnow.org/encryption-faq/#four (accessed 9 January 2025).

47 Ranking Digital Rights (2020) 2020 Ranking digital rights corporate accountability index. Available at: https://rankingdigitalrights.org/index2020/indicators/P16 (accessed 9 January 2025).

48 Puyosa I (2023) Protecting point-to-point messaging apps: Understanding Telegram, WeChat, and WhatsApp in the United States. Atlantic Council. Available at: https://www.atlanticcouncil.org/in-depth-research-reports/report/point-to-point-messaging-apps/

49 Urman A and Katz S (2022) What they do in the shadows: Examining the far-right networks on Telegram. *Information, Communication & Society* 25(7): 904–923; Buehling K and Heft A (2023) Pandemic protesters on Telegram: How platform affordances and information ecosystems shape digital counterpublics. *Social Media + Society* 9(3); Stewart NK, Al-Rawi A, Celestini C and Worku N (2023) Hate influencers' mediation of hate on Telegram: "We declare war against the anti-white system". *Social Media + Society* 9(2).

50 Puyosa (2023) Protecting point-to-point messaging apps.

51 Glover K, Dila M, Pate N, Little K, Trauthig I and Woolley S (2023) Encrypted messaging applications and political messaging: How they work and why understanding them is important for combating global disinformation. Center for Media Engagement at the University of Texas at Austin. Available at: https://mediaengagement.org/research/encrypted-messaging-applications-and-political-messaging

and enforcement practices, user reporting structures, moderation structures, encryption structures, and communication channels. External factors include legal and regulatory environments, market values, politics, sociocultural norms, and user bases.

It is important to recognize that experiences on encrypted messaging platforms are not equal. Research highlights unequal enforcement of policies, with multiple crises being inadequately addressed, unequal distribution of trust and safety resources often favoring Global North contexts, and technological solutions that are discriminatory in practice—as seen in automated content moderation tools.[52]

OVERVIEW AND ARCHITECTURE OF WHATSAPP

WhatsApp is the largest messaging platform in the world and a critical communication infrastructure for many, particularly, though not exclusively, in the Global South.[53] The platform is popular for a variety of reasons, including its ease of use and convenience, zero-rate phone plans that allow people to text for free via the app, interoperability between different mobile operating systems, and the possibility to send and receive audiovisual material and carry out phone and video calls.[54]

The rise of WhatsApp coincides with a larger shift toward more private forms of engagement in closed groups and among smaller publics[55] as well as changes in several platforms' apps toward providing avenues for more content ephemerality, including the disappearing messages feature.[56] WhatsApp's unique appeal emanates from offering features that harness the closeness and proximity of small and one-on-one conversations with family and friends,[57] which leads

52 Shahid F (2024) Colonialism in content moderation research: The struggles of scholars in the majority world. Center for Democracy and Technology. Available at: https://cdt.org/insights/colonialism-in-content-moderation-research-the-struggles-of-scholars-in-the-majority-world

53 Matassi M, Boczkowski PJ and Mitchelstein E (2019) Domesticating WhatsApp: Family, friends, work, and study in everyday communication. *New Media & Society* 21(10): 2183-2200; Cruz EG and Harindranath R (2020) WhatsApp as 'technology of life': Reframing research agendas. *First Monday* 25(12); Shahid F, Agarwal D and Vashistha A (2024) 'One style does not regulate all': Moderation practices in public and private WhatsApp groups. arXiv preprint arXiv:2401.08091; Herrada Hidalgo N, Santos M and Barbosa S (2024) Affordances-driven ethics for research on mobile instant messaging: Notes from the Global South. *Mobile Media & Communication* 12(3): 475-498.

54 That is, Baulch E, Matamoros-Fernández A and Johns A (2020) Introduction: Ten years of WhatsApp: The role of chat apps in the formation and mobilization of online publics. First Monday 25(1); Cruz and Harindranath (2020) WhatsApp as 'technology of life': Reframing research agendas; Manjoo F (2014) The other big winner in the WhatsApp deal: Your wallet. The New York Times. 20 February. Available at: https://archive.nytimes.com/bits.blogs.nytimes.com/2014/02/20/e-other-big-winner-in-the-whatsapp-deal-your-wallet; Martin ZC, Riedl MJ and Woolley SC (2023) How pro-and anti-abortion activists use encrypted messaging apps in post-Roe America. Big Data & Society 10(2).

55 Treré E (2020) The banality of WhatsApp: On the everyday politics of backstage activism in Mexico and Spain. First Monday 25(12).

56 Martin, Riedl and Woolley (2023) How pro-and anti-abortion activists use encrypted messaging apps in post-Roe America.

57 Gursky J, Riedl MJ, Joseff K and Woolley S (2022) Chat apps and cascade logic: A multi-platform perspective on India, Mexico, and the United States. Social Media + Society 8(2).

to a certain kind of platform 'stickiness',[58] as well as features meant to share information with large groups.[59] Such interstitial, in-between spaces have been referred to as meso-news spaces,[60] which acknowledges their importance for the sharing of news and information among trusted circles.

Communication on WhatsApp is end-to-end encrypted by default, which means that only the sender and receiver can decode messages. It also means that once a message has been sent, it is deleted from the server. While WhatsApp has been end-to-end encrypted by default since 2016,[61] in 2020, Meta owner Mark Zuckerberg announced a move toward more end-to-end encrypted communications across other Meta platforms,[62] with the rollout to Messenger taking place over the course of 2024. WhatsApp runs on the so-called Signal Protocol, a cryptographic protocol created by the Signal Foundation, which is also the namesake of the end-to-end encrypted messaging app Signal and the "de-facto standardization" in end-to-end encryption.[63]

WhatsApp is governed by its Terms of Service, and "Channels" are governed by Channel Guidelines. While WhatsApp discloses that it complies with legal government requests for access to basic subscriber information,[64] with regard to the spread of harmful content on WhatsApp, the platforms' mechanisms for intervening are limited to content-neutral interventions, since content including voice, text, and video is protected by end-to-end encryption.

58 Johns A, Matamoros-Fernandez A and Baulch E (2024) *WhatsApp: From a One-To-One Messaging App to a Global Communication Platform.* Cambridge: Polity Press.

59 Resende G, Melo P, Sousa H, Messias J, Vasconcelos M, Almeida J and Benevenuto F (2019) (Mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. *The World Wide Web Conference:* 818-828.

60 Tenenboim O and Kligler-Vilenchik N (2020) The Meso news-space: Engaging with the news between the public and private domains. *Digital Journalism* 8(5): 576-585.

61 Santos M and Faure A (2018) Affordance is power: Contradictions between communicational and technical dimensions of WhatsApp's end-to-end encryption. *Social Media + Society* 4(3).

62 Greenberg A (2020) Facebook says encrypting messenger by default will take years. Wired. 10 January. Available at: https://www.wired.com/story/facebook-messenger-end-to-end-encryption-default/

63 Ermoshina K and Musiani F (2019) "Standardising by running code": The Signal protocol and defacto standardisation in end-to-end encrypted messaging. Internet Histories 3(3–4): 343–363.

64 https://faq.whatsapp.com/444002211197967

Due to the nature of end-to-end encryption, content moderation undertaken is not based on the actual content of messages but instead on metadata that platforms collect, account-based information, user-based reporting, or in the case of WhatsApp, profile images, which are not part of the platform's end-to-end encryption.

In practice, the majority of moderation that happens on WhatsApp is through user reporting, group admins, and community moderators,[65] resulting in different styles and approaches to moderation depending on the individual and the context.[66]

WhatsApp has also built different types of friction into the system to slow the spread and consumption of content. This has included limiting the number of forwards possible, labelling forwarded messages, limiting the size of files that can be shared, blocking high volumes of unknown messages, and allowing users to block other users and messages. Group members can share messages with admins for review, and admins can remove messages and members from a group. Users can report other users, and WhatsApp may ban accounts.[67]

Several actors within the WhatsApp ecosystem play different roles in determining how the platform is architected, governed, and ultimately used. They include companies, state actors, fact-checkers, community moderators and admins, researchers and civil society, and end-users (See Figure 1).

65 WhatsApp (2022) Enforcing community rules and managing members. Available at: https://www.whatsapp.com/communities/learning/enforcingrules (accessed 21 January 2025).

66 Shahid, Agarwal and Vashistha (2024) 'One Style Does Not Regulate All'.

67 WhatsApp (2018) *WhatsApp FAQ.* Available at: https://faq.whatsapp.com/

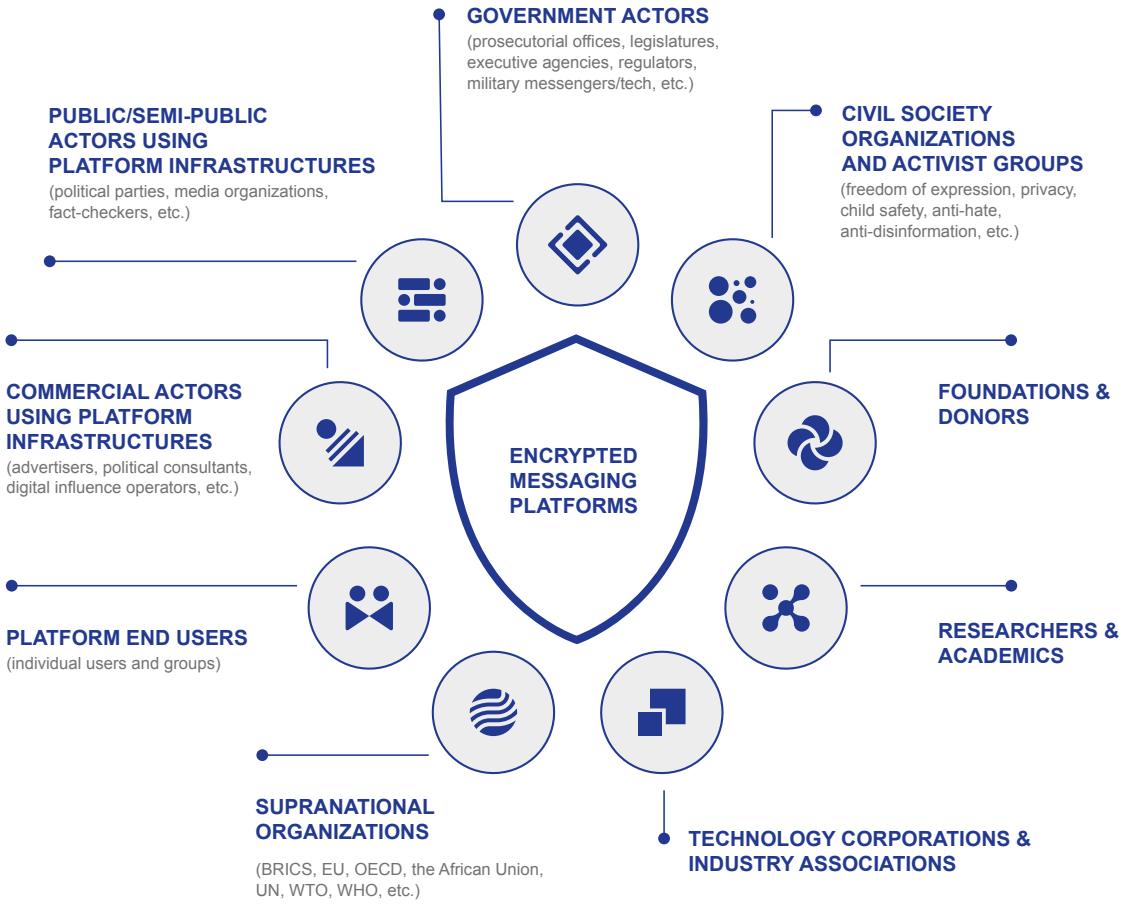

Figure 1: *Different stakeholders in the use and regulation of WhatsApp*

CURRENT STATE OF REGULATION

While it has been argued that encryption is part of international human rights standards,[68] governments across jurisdictions increasingly highlight the challenges that encrypted messaging platforms pose: serving as "honeypots" for a range of online harms. Attempts to regulate encrypted messaging platforms are part of a larger trend of governments seeking to hold online platforms accountable, as well as efforts to address specific harms such as disinformation, CSAM, and terrorist content. Regulatory efforts tend either to focus on expanding governmental powers—through decryption, backdoors, traceability, and moderation requirements—or to take a more systems approach, requiring companies to undertake risk assessments and due diligence. In extreme examples, governments also attempt to restrict access to encrypted messaging platforms through blocking orders or limitations on specific aspects of a service such as video. Examples of such measures have been

68 Kaye D (2015) Human Rights Council Twenty-ninth session Agenda item 3, Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Available at: https://documents.un.org/doc/undoc/gen/g15/095/85/pdf/g1509585.pdf. and https://cdt.org/insights/the-european-court-of-human-rights-concludes-encryption-backdoor-mandates-violate-the-right-to-private-life-of-all-users-online/#:~:text=The%20case%20concerned%20the%20statutory,or%20security%20services%20together%20with

particularly connected to elections, political content, times of social unrest, and attempts by governments to compel companies to decrypt content. Civil society groups have raised concerns that many regulatory efforts around encrypted messaging platforms, particularly those focused on expanding governmental powers, are disproportionate and incentivize platforms to implement measures that undermine the privacy and security of encrypted services, facilitate surveillance, and negatively impact freedom of expression and privacy.[69] Key trends include:

- **Traceability:** Requirements compelling companies to place an identifier on messages so the originator of a message can be shared if requested by law enforcement. An example of traceability requirements can be seen in the 2021 Intermediary Guidelines and Digital Media Ethics Code passed in India. These require online platforms to ensure that the original sender of an online message can be identified and disclosed to law enforcement when required through a court order.[70] In 2021, WhatsApp and Facebook challenged these requirements through a petition in the Delhi High Court, arguing that it violates the right to privacy enshrined in the Constitution and goes against principles of proportionality, necessity, and minimization.[71] Several civil society organizations have also pushed back against traceability requirements, noting that they would undermine encryption.[72]
- **Hash databases:** While some efforts to regulate encrypted messaging platforms seek to address harmful content online more broadly, many are specifically in the context of addressing online terrorist content and CSAM. Hash databases involve the regulator or a third party maintaining identified CSAM or terrorist content and requiring or encouraging companies to report detected violating content and/or scan their services using the hashes stored in the database. For example, in the US, Electronic Service Providers must report detected CSAM to the National Center for Missing and Exploited Children cybertip line.[73] In the proposed EU Child Sexual Abuse Regulation, already identified CSAM would be hashed and stored in a database maintained by the EU Centre to Prevent and Combat

69 Center for Democracy and Technology (2023) Encryption and government hacking archives. Available at: https://cdt.org/area-of-focus/government-surveillance/encryption-and-government-hacking/

70 MEITY (2021) The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. www.meity.gov.in. Available at: https://www.meity.gov.in/writereaddata/files/Information%20Technology%20%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20%28updated%2006.04.2023%29-.pdf

71 SFLC.in (2021) Legal challenges to the traceability provision – What is happening in India? Available at: https://sflc.in/legal-challenges-traceability-provision-what-happening-india/

72 Internet Society (2024) Traceability in end-to-end encrypted environments - Internet Society. Internet Society. Available at: https://www.internetsociety.org/resources/doc/2024/traceability-in-end-to-end-encrypted-environments/

73 National Centre for Missing and Exploited Children (2019) CyberTip Report. Cybertip.org. Available at: https://report.cybertip.org/

Child Sexual Abuse[74] and companies would be required to remove matching content from their platform.[75]

- **Blocking and feature restrictions:** In some contexts, governments have blocked encrypted messaging platforms or restricted certain features. These measures have often been taken during elections and times of political unrest, as seen in recent years in Brazil, India, Uganda, Zambia, and other countries. In May 2023, based on an order from the Indian government, 14 online messaging apps were blocked.[76] In 2015 and 2016 in Brazil, the government blocked WhatsApp for refusing to hand over user data and intercept messages on its platform.[77] In 2021, the Uganda Communications Commission blocked online messaging platforms Twitter (now X), WhatsApp, Signal, and Viber before the presidential elections.[78] During the 2021 elections, in Zambia the ICT regulator blocked several social media platforms including WhatsApp, Facebook, Twitter, and Messenger.[79] In some countries, governments have put in place a partial ban on messaging services like WhatsApp, where voice and video do not work but messaging does.[80]
- **Detection technologies:** Governments have explored requirements for companies to use detection and accredited technologies to identify violating content.[81] For example, the proposed European Union Child Sexual Abuse Regulation would require service providers to comply with detection orders and use technologies to identify and remove violating content.[82] The UK Online Safety Act requires online messaging platforms to ensure that they can apply accredited technologies on encrypted channels if ordered to do so.[83] Companies and multiple civil society organizations pushed back on both regulations during their draft stages,[84] stating that such requirements would undermine encryption. In 2023, the UK government admitted

74 European Commission (n.d.) EU centre to prevent and combat child sexual abuse – European Commission. Available at: https://home-affairs.ec.europa.eu/whats-new/communication-campaigns/euvschildsexual-abuse-campaign-prevent-and-combat-child-sexual-abuse/eu-centre-prevent-and-combat-child-sexual-abuse_en

75 European Commission (2022) Proposal for a regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0209

76 https://www.indiatoday.in/india/story/centre-blocks-14-mobile-apps-used-by-terrorists-in-pak-to-send-coded-texts-in-j-k-2366821-2023-05-01 (accessed 31 January 2025).

77 InternetLab (2024) WhatsApp Case IV: Non-compliance with judicial requests for user data. Available at: https://bloqueios.info/en/casos/block-for-non-compliance-with-judicial-requests-for-user-data/ (accessed 10 January 2025).

78 Reporters without borders (2021) Uganda blocks social media and messaging apps, isolating election. Available at: https://rsf.org/en/uganda-blocks-social-media-and-messaging-apps-isolating-election (accessed 10 January 2025).

79 Freedom House (2021) Zambia: Freedom on the Net 2021 Country Report. Freedom House. Available at: https://freedomhouse.org/country/zambia/freedom-net/2021

80 BBC (2024) Tens of millions secretly use WhatsApp despite bans. Available at: https://www.bbc.com/news/articles/ckke9x0e50x0

81 https://www.internetsociety.org/resources/doc/2020/fact-sheet-client-side-scanning

82 Article 7. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0209

83 UK Government (2023) Online Safety Act 2023. Available at: https://www.legislation.gov.uk/ukpga/2023/50/contents.

84 https://signal.org/blog/pdfs/upload-moderation.pdf and https://edri.org/wp-content/uploads/2024/07/Statement_-The-future-of-the-CSA-Regulation.pdf ; https://techcrunch.com/2023/09/21/meredith-whittaker-reaffirms-that-signal-would-leave-u-k-if-forced-by-privacy-bill/#:~:text=Onstage%20at%20TechCrunch%20Disrupt%202023,end%2Dto%2Dend%20encryption and https://www.internetsociety.org/resources/internet-fragmentation/uk-online-safety-act

that the "technology needed to securely scan encrypted messages sent on WhatsApp and Signal does not exist".[85]

- **Risk assessments:** Governments are increasingly requiring companies—such as encrypted messaging platforms—to undertake measures including risk assessments. For example, as per the UK Online Safety Act, all covered platforms need to undertake a risk assessment of illegal content on their platform, including the level of risk of users encountering illegal content and the risk of their platform being used for the commission of a priority offense.[86] The DSA also takes a risk-based approach, requiring large online platforms and search engines to undertake risk assessments of their services to the EU market, audit the same, and comply with heightened transparency requirements.[87]
- **Registration requirements:** Some governments have explored the possibility of requiring over the top services (OTTs), which would include platforms such as WhatsApp, Telegram, and Signal, to obtain a license to operate in the country, similar to a telecommunication company. This would subject them to licensing agreements which typically contain extensive surveillance and security requirements. For example, in May 2024, the Zambian government announced that it requires social media companies to acquire a license to operate in the country.[88]
- **Expanding regulatory ecosystems:** Whether or not specifically tailored to encrypted services, regulatory ecosystems around online platforms are evolving, with regulators creating new types of positions and entities to be involved in different aspects of ensuring platform accountability. For example, the UK Online Safety Act has created the role of a "skilled person" who must be consulted before Ofcom can issue an order for the use of "accredited technologies".[89] The DSA has created the role of autonomous public regulators both within the member states and at the EU level to implement regulations for social media networks.[90] Canada's Online Harms Act establishes the "Digital Safety Office" which has the power to conduct audits, issue compliance orders, and issue penalties on social media services. It can also establish online safety standards, engage in research, and develop resources for the public.[91]
- **Arrests:** Governments are taking more draconian measures by arresting company employees and WhatsApp administrators. In 2023, in India, the Uttar Pradesh Police arrested a WhatsApp

85 https://www.wired.com/story/britain-admits-defeat-online-safety-bill-encryption/ (accessed 23 January 2025).

86 UK Government (2023) Online Safety Act 2023. Available at: https://www.legislation.gov.uk/ukpga/2023/50/contents.

87 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065

88 Short K (2024) Uproar over Zambia's plan to regulate online broadcasting. dw.com. Available at: https://www.dw.com/en/uproar-over-zambias-plan-to-regulate-online-broadcasting/a-69163034 (accessed 10 January 2025).

89 UK Government (2023) Online Safety Act 2023.

90 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065 (accessed 23 January 2025).

91 Parliament of Canada (2024) Government Bill (House of Commons) C-63 (44-1) - First Reading - Online Harms Act - Parliament of Canada. Available at: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading

group administrator for failing to moderate derogatory messages against the Chief Minister.[92] In 2016, a Facebook executive was arrested in Brazil as courts demanded the company provide WhatsApp data to support law enforcement in a drug-trafficking case.[93] In 2017, the Kenyan government arrested two WhatsApp group administrators for sharing hate messages that purportedly threatened national security.[94]

- **Access through surveillance powers, weakening encryption, and spyware:** Many governments have in place powers that allow intelligence agencies to require the decryption of content. In the UK, the 2022 Interception of Communications Code of Practice under the 2016 Investigatory Powers Act allows covered intelligence agencies with the relevant powers to remove encryption, if "reasonably practical", from messages and content after receiving a legal "technical capability" notice.[95] Governments have also sought to place limits on the strength of encryption permitted. Research by civil society on encryption laws globally shows that such restrictions exist in multiple countries, including China, Egypt, India, Iran, Pakistan, Russia, and more.[96] Governments have also accessed encrypted communications through the use of spyware and hacking techniques such as FinFisher and Pegasus.[97]
- **Controlling content:** Governments have taken different approaches to control content on social media, including encrypted messaging platforms. In 2014, the Ugandan government imposed a daily tax on the use of OTTs, including Facebook, WhatsApp, and Viber, in order to slow the spread of false information on those platforms.[98]
- **Political pressure:** In addition to regulation, governments have placed political pressure on companies to not implement encryption across services. These are often in the context of child safety and countering online terrorism. For example, according to the UK's 2022 and 2023 Freedom on the Net report, in 2019, the UK, the US, and Australia spoke out against Meta's plans to implement end-to-end encryption across messaging platforms.[99] In 2022,

92 Scroll.in. (2023) UP: WhatsApp admin arrested after group member makes 'derogatory' remarks against CM Adityanath. Available at: https://scroll.in/latest/1053894/up-whatsapp-admin-arrested-after-group-member-makes-derogatory-remarks-against-cm-adityanath (accessed 21 January 2025).

93 Haynes B (2016) Facebook exec jailed in Brazil as court seeks WhatsApp data. Reuters. 1 March. Available at: https://www.reuters.com/article/technology/facebook-exec-jailed-in-brazil-as-court-seeks-whatsapp-data-idUSKCN0W34WA/

94 Muendo M (2017) Kenya targets WhatsApp administrators in its fight against hate speech. The Conversation. Available at: https://theconversation.com/kenya-targets-whatsapp-administrators-in-its-fight-against-hate-speech-82767 (accessed 10 January 2025).

95 UK Home Office (2022) Interception of communications code of practice 2022 (accessible). Available at: https://www.gov.uk/government/publications/interception-of-communications-code-of-practice-2022/interception-of-communications-code-of-practice-2022-accessible

96 Global Partners Digital (2009) World map of encryption laws and policies. Available at: https://www.gp-digital.org/world-map-of-encryption/

97 Basu S (2024) Rebekah Brown discusses the global abuse of commercial spyware on TaiwanPlus. The Citizen Lab. Available at: https://citizenlab.ca/2024/12/rebekah-brown-discusses-the-global-abuse-of-commercial-spyware-on-taiwanplus/ (accessed 10 January 2025).

98 Altman-Lupu M (2020) Uganda's tax on social media: Financial burdens as a means of suppressing dissent. *Columbia Human Rights Law Review.* Available at: https://hrlr.law.columbia.edu/files/2020/02/51.2.6-Altman-Lupu-1.pdf

99 Freedom House (2022) United Kingdom: Freedom on the Net 2022 Country Report. Freedom House. Available at: https://freedomhouse.org/country/united-kingdom/freedom-net/2022; https://freedomhouse.org/country/united-kingdom/freedom-net/2023 (accessed 8 February 2025).

the No Place to Hide Campaign, supported by the UK Government, was launched to raise awareness about the danger of encrypted messaging to children and to prevent Meta from expanding the use of end-to-end encryption.[100] The Five Country Ministerial—an annual meeting between the ministries of home affairs, public safety, interior, security, border and immigration from Australia, Canada, New Zealand, the UK, and the US—has also criticized companies providing encrypted products, arguing that they limit law enforcement access to content.[101]

The overall expanding regulatory ecosystem on platform accountability underscores the deep tension that exists between online platforms and governments. Actions taken by some governments, such as blocking, the use of spyware, and surveillance, exemplifies how governments can undermine human rights and make spaces that are meant to be secure insecure. Arrests of company personnel and group admins demonstrate the precariousness of these roles and the struggle of governments to appoint responsibility and liability for content they deem harmful. Legislative requirements for traceability make it clear that approaches to regulating online encrypted platforms are still blunt. The political pressure placed on companies by governments to not implement encryption further underscores this challenge.

100 Mullin J (2022) The U.K. paid $724,000 for a creepy campaign to convince people that encryption is bad. It won't work. Electronic Frontier Foundation. Available at: https://www.eff.org/deeplinks/2022/01/uk-paid-724000-creepy-campaign-convince-people-encryption-bad-it-wont-work (accessed 12 January 2025).

101 UK Home Office (2019) Joint meeting of Five Country Ministerial and quintet of Attorneys-General: communiqué, London 2019 (accessible version). Available at: https://www.gov.uk/government/publications/five-country-ministerial-communique/b9adc2fe-d82c-4615-9c0b-6427d09733af (accessed 12 January 2025).

Content Moderation and Online Encrypted Platform

Presently, the main form of content moderation that takes place on encrypted messaging platforms is through user reporting. While some governments are calling for platforms to moderate content, either before or after it is shared, doing so would require undermining encryption to different degrees. Techniques for content moderation on encrypted messaging platforms that are being explored include:

- **Traditional backdoors:** These involve key escrow, where a service provider is required to share encryption keys with a third party that the government has approved. In turn, the government can access the keys, and decrypt content, for established purposes such as during an investigation.[102] It could also involve creating a middle server where content is decrypted, scanned, and then re-encrypted. While favored by governments and intelligence services, such backdoors undermine the core premises of end-to-end encryption.
- **Client-side scanning[103]:** This involves searching for and flagging matches of violating content, such as a wordlist-based profanity filter or hashing techniques such as the PhotoDNA database, for detecting CSAM before a message is sent to a recipient or after it is received. Client-side scanning happens on the device of the sender or receiver. Its proponents argue that it preserves end-to-end encryption, while its critics attest that scanning content before it is sent through an e2ee (end-to-end encrypted) messenger defeats the purpose of end-to-end encryption.
- **Message franking[104]:** Confirmation and moderation of user-reported messages, including verifying the user without compromising anonymity. Through message franking, a user who reported problematic content in an end-to-end encrypted chat creates evidence that the platform can access, limiting the deniability of a sender.
- **Automated scanning using homomorphic encryption[105]:** Homomorphic encryption enables computations to be performed directly on encrypted data, ensuring data privacy and security throughout the process. The use of automated tools to scan content using homomorphic encryption allows for the detection of harmful content in a message and the

102 Duan and Grimmelmann (2024) Content moderation on end-to-end encrypted systems: A legal analysis.

103 Abelson H, Anderson R, Bellovin SM, Benaloh J, Blaze M, Callas J, Diffie W, Landau S, Neumann PG, Rivest RL and Schiller JI (2024) Bugs in our pockets: The risks of client-side scanning. *Journal of Cybersecurity* 10(1). DOI: 10.1093/cybsec/tyad020; Geierhaas L, Otto F, Häring M and Smith M (2023) Attitudes towards client-side scanning for CSAM, terrorism, drug trafficking, drug use and tax evasion in Germany. *2023 IEEE Symposium on Security and Privacy (SP)*: 217-233. DOI: 10.1109/SP46215.2023.10179417; Jain S, Creţu AM, Cully A and de Montjoye YA (2023) Deep perceptual hashing algorithms with hidden dual purpose: when client-side scanning does facial recognition. *2023 IEEE Symposium on Security and Privacy (SP)*: 234-252. DOI: 10.1109/SP46215.2023.10179310

104 Mayer J (2019) Content moderation for end-to-end encrypted messaging. Princeton University. Available at: https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf; Rahalkar C and Virgaonkar A (2022) SoK: Content moderation schemes in end-to-end encrypted systems. arXiv preprint arXiv:2208.11147. DOI: 10.48550/arXiv.2208.11147

105 Knodel M, Fábrega A, Ferrari D, Leiken J, Hou BL, Yen D, de Alfaro S, Cho K and Park S (2024) How to think about end-to-end encryption and AI: Training, processing, disclosure, and consent. arXiv preprint arXiv:2412.20231. DOI: 10.48550/arXiv.2412.20231

sharing of the result with the recipient without the platform itself knowing information about the content or the result. While an active area of research and development, homomorphic encryption is resource-intensive and currently not viable at a large scale. Free expression and privacy activists have criticized several of the techniques being explored for content moderation. They have instead emphasized approaches that do not undermine encryption, such as metadata analysis and user reporting.

- **Metadata analysis:** This involves analyzing data such as file size, type, date or time, and sender or receiver to detect potentially harmful messages. Machine learning techniques can enhance the type and scale of metadata analysis that can be undertaken.[106]
- **User reporting and community moderation:** These rely on engaged users who intervene when problematic content is shared and report it to platforms or address its dissemination directly in chats by confronting its senders. After receiving reported content, platforms can review the content through human moderators and/or automated tools against applicable policies. Depending on the platform, the actions taken on moderated content and accounts vary. For example, according to WhatsApp, actions the company may take include issuing a warning, suspending an account, preventing further activity in a group, removing a reported profile or account information, revoking or blocking invite links, and reporting violating content to competent authorities.[107] While social interventions through group moderation,[108] such as group admins, community moderators, and user reporting, can preserve privacy and incorporate context into moderation decisions, platforms need to make significant inroads into providing the relevant infrastructure and support to users who take on these roles. We discuss other significant challenges under recommendations.

106 Jones A (2017) Practical data privacy: The emergence of homomorphic encryption. Enigma 2017 conference program. Available at: https://www.usenix.org/conference/enigma2017/conference-program/presentation/jones (accessed 22 January 2025); Kamara S, Knodel M, Llansó E, Nojeim G, Qin L, Thakur D and Vogus C (2022) Outside looking in: Approaches to content moderation in end-to-end encrypted systems. Center for Democracy and Technology. Available at: https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/

107 WhatsApp (2022) WhatsApp messaging guidelines. Available at: https://www.whatsapp.com/legal/messaging-guidelines

108 Scheffler S and Mayer J (2024) Group moderation under end-to-end encryption. *Proceedings of the Symposium on Computer Science and Law:* 36-47. DOI: 10.1145/3614407.3643704

CHALLENGES ON ONLINE ENCRYPTED PLATFORMS

This section looks closely at the challenges found on WhatsApp. Our analysis exposes distinct affordances, cultures of use, and political deployments surrounding encrypted messaging platforms, and the ways they enable and disrupt different forms of exclusionary political discourses and strategies. It also reveals the rapid evolution of disinformation and extreme speech as users find new ways and tactics to communicate them and circumvent measures to curb the misuse of the platform.

Contradictions

Although the technological features of WhatsApp promise privacy and secure communication, the actual use and applications on the ground are suffused with contradictions. "Lived encryptions" intimates that the promised privacy of the encrypted service is swiftly overturned by authoritarian and surveilling governments when they intend to, by, for example, seizing phones from suspected dissenters and other brutal measures which do not necessarily require sophisticated techniques of breaking open encryption.[109] In conflict situations as well as ordinary contexts of law and order enforcement, the safety of a WhatsApp conversation is not a taken-for-granted condition because of extra-legal pressure tactics. Incidents of coercion have been reported in India, where local police have been accused of using extrajudicial tactics to coerce people to reveal their private WhatsApp chats. In October 2021, in Hyderabad, Telangana, city police were seen in a video asking for people's phones and checking them for suspected drug trafficking.[110] In Ghana, the military-grade cyber-surveillance system Pegasus has reportedly been used to spy on the private communications of political opponents and dissidents, including on WhatsApp.[111] In September 2016, in the city of São Paulo in Brazil, an infiltrating military operation was executed to target WhatsApp progressive activism groups. Military intelligence officer Willian Botelho appeared on WhatsApp groups as "Balta Nunes" to track the daily routines of activists and bring them to the police.[112]

In authoritarian or non-democratic regimes, the encryption offered by WhatsApp is also not a guarantee that political critics, journalists, and fact-checkers will not be harassed, threatened, or forced to divulge data or the identity of sources. Yet, journalists and activists also see WhatsApp as a form of secure communication. Similarly, digital influence actors circumvent and reshape closed chat architecture with new forms of top–down political messaging.

109 Udupa and Wasserman (2025) *WhatsApp in the World,* p. 6.

110 Oommen P (2021) Hyderabad cops are stopping people on the road, checking WhatsApp chats for 'drugs'. The News Minute. Available at: https://www.thenewsminute.com/telangana hyderabad-cops-are-illegally-checking-phones-whatsapp-citizens-part-drug-crackdown-156997

111 Kabir A and Adebajo K. How digital surveillance threatens press freedom in West Africa. HumAngle. Available at: https://humanglemedia.com/how-digital-surveillance-threatens-press-freedom-in-west-africa/ (accessed 19 January 2025).

112 https://brasil.elpais.com/brasil/2018/06/29/politica/1530293956_036191.html (accessed 19 January 2025).

The broadcasting abilities of WhatsApp were boosted when the platform introduced the "Channels" function. Here, only moderators can post content, and members can only "like" or react with an emoji.[113]

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 seen as trustworthy. Political actors have expanded campaign activities to WhatsApp groups to gain "organic" influence. In countries like India and Brazil, representatives of political parties or hired influencers have penetrated existing WhatsApp groups or created new groups by enlisting local community leaders, neighborhood association members, local service providers, as well as members of extended families. Content that flows through such groups taps into existing social ties and trustful relations, thereby creating channels for extreme speech that are hard to dismantle. In such groups, exclusionary messages come mingled with pleasant messages such as "good morning" greetings and religious hymns, creating a "lived rhythm of the social".[114] Defined as "deep extreme speech", 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".[115] When disinformation is shared on WhatsApp, the proximity one has to other individuals like family and friends, combined with social expectations in a group and the attempt to allow others to save face, may serve as a hindrance to corrections and other interventions.[116]

Microtargeting and segmentation

Microtargeting occurs when WhatsApp messages are aimed at small groups through a centralized structure "built to manage and to stimulate members of discussion groups, which [are] treated as segmented audiences".[117] Brazil serves as a case study for how such microtargeting and segmentation is used in political campaigns. During the 2018 general election in Brazil, the pro-Bolsonaro campaign weaponized WhatsApp as a very efficient political weapon, particularly to target economically disadvantaged sections of the Brazilian population. Given the high importance of WhatsApp in the everyday life of Brazilians,[118] the platform facilitates microtargeting and segmentation, as it attracts users without needing to

113 WhatsApp Channels: Here's Everything You Need To Know. Available at: https://about.fb.com/news/2023/09/whatsapp-channels-heres-everything-you-need-to-know/ (accessed 14 January 2025).

114 Udupa S (2025) Deep extreme speech: Intimate networks for inflamed rhetoric on WhatsApp. In: Udupa S and Wasserman H *WhatsApp in the World*, p. 82.

115 Udupa (2025) Deep extreme speech, p. 70.

116 Malhotra P (2024) "What you post in the group stays in the group": Examining the affordances of bounded social media places. *Social Media + Society,* 10(3). https://doi.org/10.1177/20563051241285777

117 Evangelista R and Bruno F (2019) WhatsApp and political instability in Brazil: Targeted messages and political radicalisation *Internet Policy Review* 8 (4). https://doi.org/10.14763/2019.4.1434, p. 3.

118 Spyer J (2017) *Social Media in Emergent Brazil.* London: UCL Press.

access information published on other platforms owned by Meta, such as Facebook and Instagram. It is segmented because these WhatsApp groups disseminate large amounts of text and audio messages, videos, and images to reach different cohorts of Brazilians. These messages can also be used to spread malicious content and disinformation through WhatsApp groups.[119]

WhatsApp groups allow for the formation of a social media pipeline in which small communities on WhatsApp create many ways of sharing information while reinforcing intimate connections. The process is described as "capillarity" for the way it mirrors the capillaries of blood vessels in the human body.[120] Microtargeting and segmentation is also a way to expand presence on multiple channels and platforms and to replicate information as widely as possible, beyond the ecosystem of messaging services. This process creates complex flows of information that are germinated and fertilized across different WhatsApp groups. Overall, it is a strategic and extensive weaponization of WhatsApp groups. WhatsApp's infrastructure has also been strategically used by political actors in India to spread propaganda and hateful discourse.[121]

Influencers on WhatsApp

The role of online influencers in the production and dissemination of extreme speech and disinformation has been well documented, but their use of encrypted messaging platforms, including WhatsApp, has received less attention. While perhaps not an obvious choice for reaching large numbers of people, WhatsApp provides a direct and personal form of communication through which influencers can interact with their followers using Channels, status updates, and group chats.

Influential WhatsApp accounts in India have traditionally been based on creating WhatsApp groups and are often linked to political actors. In September 2023, Meta enabled Channels in India, allowing influencers to send one-way broadcast messages, including text, images, and videos, to their followers. While media outlets and fact-checkers can make use of Channels for reporting purposes, political parties have also weaponized them. For example, in November 2024, ahead of elections in the state of Jharkhand, a political party shared a communally charged advertisement on its Channel, depicting Muslims invading a man's home and

119 Barbosa S and Back, C (2020) The dark side of Brazilian WhatsAppers. In: Sabriego J, Amaral A J and Salles E B C (eds) *Algoritarismos.* São Paulo, BR, Valencia, ES: Tirant lo Blanch.

120 Barbosa S (2021) COMUNIX WhatsAppers: The community school in Portugal and Spain. *Political Studies Review* 19(2) 171-178. https://doi.org/10.1177/1478929920951076

121 Nizaruddin F (2021) Role of public WhatsApp groups within the Hindutva ecosystem of hate and narratives of "CoronaJihad". *International Journal of Communication* 15: 1102–1119.

blaming the state's ruling party.[122] Though the Election Commission ordered its removal, the advertisement remained on the party's verified channel, which had 32,000 followers. Enlisting "hate influencers"[123] on platforms like WhatsApp through what is defined as "shadow politics"[124] is a phenomenon also seen in Brazil. Research shows how influencers weaponized WhatsApp to operate "a distributed strategy of propagating disinformation"[125] that started before Jair Bolsonaro was elected in 2018 but was further aggravated during his mandate (2018–2022). Such disinformation travelled across multiple platforms and made its way back to WhatsApp, turning this strategy into a flywheel of action (see next section on cross-media manipulation). Due to the numerical limitations imposed on group membership, "offspring" private groups were also created from the "original group", including task-specific groups devoted to, for example, distributing materials and posters about a political target. Individuals also connected with each other on other social media platforms, including Facebook, Instagram, YouTube, and X. While some politically aligned influencers pivot their audiences towards particular narratives during election season, others may be commercially driven. Combatting the effects of such influence on WhatsApp is particularly challenging when mechanisms for monitoring are limited.

Cross-media manipulation

The cross-media effect demonstrates how the origins and sources of messages are not easily traceable on WhatsApp, particularly on private groups. Information on WhatsApp easily flows onto the social media platforms and then moves to other WhatsApp closed groups as well.[126] Cross-media manipulation occurs when group members set up a coordinated manipulation strategy in which producers co-create and edit messages and transmit them to group members, who then disseminate messages widely to offspring groups.

Such patterns of coordinated manipulation were observed in India's 2019 general election. A research study documented sophisticated cross-platform manipulation strategies whereby 600 public WhatsApp groups supporting the ruling party used 75 distinct hashtag manipulation campaigns that successfully engineered Twitter trends through coordinated mass posting.[127] These campaigns utilized centrally controlled but voluntary participation mechanisms,

122 Chowdhury A (2024) BJP Jharkhand's communal poll ad remains online despite EC takedown order. BOOM. 19 November. Available at: https://www.boomlive.in/news/bjp-jharkhand-political-ad-assembly-elections-x-whatsapp-islamophobia-muslims-jmm-27007

123 Stewart, Al-Rawi, Celestini and Worku (2023) Hate influencers' mediation of hate on telegram: *"We declare war against the anti-white system".*

124 Udupa S (2024) Shadow politics: Commercial digital influencers, "data," and disinformation in India. *Social Media + Society* 10(1). https://doi.org/10.1177/20563051231224719

125 Ozawa JVS, Woolley SC, Straubhaar J, Riedl MJ, Joseff K and Gursky J (2023) How disinformation on WhatsApp went from campaign weapon to governmental propaganda in Brazil. Social Media + Society 9(1). https://doi.org/10.1177/20563051231160632, p.2.

126 Gursky, Riedl, Joseff and Woolley (2022) Chat apps and cascade logic: A multi-platform perspective on India, Mexico, and the United States.

127 Jakesch M et al (2021) Trend alert: A cross-platform organization manipulated Twitter trends in the Indian general election. P*roceedings of the ACM on Human-Computer Interaction* 5(CSCW2), pp. 1–19. doi:10.1145/3479523

demonstrating how digital tools can be leveraged for large-scale narrative manipulation while maintaining the appearance of grassroots participation. From there, messages travel in a cascade not only on WhatsApp but also to social media platforms.[128] This cross-platform amplification strategy, observed in both the Brazilian and Indian contexts, highlights how political organizations have evolved to exploit the interconnected nature of modern social media ecosystems, creating sophisticated networks that blur the lines between organic political participation and orchestrated manipulation campaigns.

Gender-based harassment on WhatsApp

WhatsApp's infrastructural dominance and popularity across contexts have resulted in the platform being used as a popular vector for gendered harassment and abuse. While such harassment is often targeted at women, it can be further compounded if individuals are also members of specific religious, racial, or ethnic groups or when they are experiencing harassment based on their sexual orientation.[129] Motivated by and fueled through networked misogyny (how platforms are used to promote violence against women),[130] gendered harassment on WhatsApp relates to the larger domain of technology-facilitated gender-based violence,[131] in which platforms' mediations and affordances are utilized to maximize harm. On WhatsApp (and other end-to-end encrypted platforms), instances of gender-based harassment and violence include the spread of non-consensual intimate imagery which perpetrators possess and then further distribute.[132] It can also include the sending of "dick pics" and other unsolicited sexually explicit content to unwitting recipients,[133] a practice understood as "gendered and sexualized power play".[134] In Lebanon, research has documented the important infrastructural role that WhatsApp assumes in violence targeting women and queer individuals through sexualized doxxing, coercive messages, and rape threats.[135]

128 Hale S A, Belisario A, Mostafa AN and Camargo C (2024) Analyzing misinformation claims during the 2022 Brazilian general election on WhatsApp, Twitter, and Kwai. *International Journal of Public Opinion Research,* 36(3), edae032. Available at: https://doi.org/10.1093/ijpor/edae032

129 Binder L, Ueberwasser S and Stark E (2021) Gendered hate speech in Swiss WhatsApp messages. In: Giusti G and Innàccaro G (eds) *Language, Gender and Hate Speech: A Multidisciplinary Approach.* Fondazione Università Ca' Foscari, pp. 59–74. DOI: 10.30687/978-88-6969-478-3/003

130 Banet-Weiser S and Miltner KM (2016) #MasculinitySoFragile: Culture, structure, and networked misogyny. *Feminist Media Studies* 16(1): 171–174. DOI: 10.1080/14680777.2016.1120490

131 Hinson L, Mueller J, O'Brien-Milne L and Wandera N (2018) Technology-facilitated gender-based violence: What is it, and how do we measure it? *International Center for Research on Women.* Available at: https://www.icrw.org/wp-content/uploads/2018/07/ICRW_TFGBVMarketing_Brief_v8-Web.pdf

132 Semenzin S and Bainotti L (2020) The use of Telegram for non-consensual dissemination of intimate images: Gendered affordances and the construction of masculinities. *Social Media + Society* 6(4). DOI: 10.1177/2056305120984453

133 Lestari SP and Mutmainnah (2023) Choice of action for victims of cyber gender-based violence (sexting) via WhatsApp. *Digital Theory, Culture & Society* 1(1): 61–69. DOI: 10.61126/dtcs.v1i1.10; Paasonen S, Light B, and Jarrett K (2019) The dick pic: Harassment, curation, and desire. *Social Media + Society* 5(2). DOI: 10.1177/2056305119826126

134 Marcotte AS, Gesselman AN, Fisher HE and Garcia JR (2021) Women's and men's reactions to receiving unsolicited genital images from men. *Journal of Sex Research* 58(4): 512–521. DOI: 10.1080/00224499.2020.1779171

135 Riedl MJ, El-Masri A, Trauthig IK and Woolley SC (2024) Infrastructural platform violence: How women and queer journalists and activists in Lebanon experience abuse on WhatsApp. *New Media & Society.* Epub ahead of print 24 April 2024. DOI: 10.1177/14614448241248372

Others have documented how private relationship information was shared on WhatsApp alongside disparaging photos and text,[136] which hearkens to a "masculinist logic of shame"[137] that seeks to intimidate women. Gendered forms of harassment have been shown to affect women journalists who, consequently, sometimes withdraw from platforms like WhatsApp, change their phone numbers, or give up their jobs altogether.[138] Although there are examples of interventions to combat gender-based and other forms of identity-based violence and harassment on WhatsApp,[139] the platform continues to present a harmful environment for individuals who are marginalized and minoritized based on their sexual orientation, gender identity, religious, or racial/ethnic identity.[140]

Fact-checking on WhatsApp

WhatsApp holds the potential to be advantageous for fact-checking. It offers a trusted and secure platform for fact-checkers to engage with the public and receive tip-offs or examples of disinformation and extreme speech circulating on the platform. This can be especially useful during high-stakes events such as elections or natural disasters when fact-checkers can leverage community groups to gauge the virality of disinformation and prioritize the most harmful messages for verification.

136 Dagher J (2018) Online privacy threats to women and LGBTIQ communities in Lebanon. *SMEX*. Available at: https://smex.org/wp-content/uploads/2018/11/OnlinePrivacyThreats_EN.pdf

137 Udupa S (2018) Gaali cultures: The politics of abusive exchange on social media. *New Media & Society* 20(4): 1506–1522. DOI: 10.1177/1461444817698776

138 Koirala S (2020) Female journalists' experience of online harassment: A case study of Nepal. *Media and Communication* 8(1): 47–56. DOI: 10.17645/mac.v8i1.2541; Melki JP and Mallat SE (2016) Block her entry, keep her down and push her out: Gender discrimination and women journalists in the Arab world. *Journalism Studies* 17(1): 57–79. DOI: 10.1080/1461670X.2014.962919; Riedl, El-Masri, Trauthig and Woolley (2024) Infrastructural platform violence: How women and queer journalists and activists in Lebanon experience abuse on WhatsApp.

139 Fotini C, Larreguy H, Muhab N and Parker-Magyar E (2022) Can media campaigns empower women facing gender-based violence amid COVID-19? Toulouse School of Economics White Paper. Available at: https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2022/wp_tse_1294.pdf; Markan M, Dhingra R, Segan M, Dabla V, Sagar M, Neogi S, Dey S and Chakravarty N (2022) Gender-based violence programming in times of COVID-19: Challenges, strategies and recommendations. *Frontiers in Global Women's Health* 3.

140 Riedl, El-Masri, Trauthig and Woolley (2024) Infrastructural platform violence: How women and queer journalists and activists in Lebanon experience abuse on WhatsApp.

However, WhatsApp also presents some challenges for fact-checkers. Due to encryption, fact-checkers cannot find disinformation or extreme speech on the platform themselves unless examples are sent to them by users.[141] Another challenge is that users of WhatsApp tend to trust the information they receive from friends, family, or colleagues and are therefore not always likely to verify or question the information they receive or send it to fact-checkers for verification (see the section on deep extreme speech, p. 27).[142]

A further problem is that 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. Furthermore, not everyone will act on verified information, even if they receive it—especially if the original false information has a stronger emotional appeal. Users may share false information if they are under the impression that doing so may be helpful to those in their networks and communities. Fact-checkers may be able to harness this desire to be helpful by emphasizing the harmful effects of disinformation and extreme speech and encouraging users to correct problematic information circulating on the platform. This will also require fact-checkers to establish trust among users and empower them to check information themselves.[143]

While several fact-checking organizations have set up tiplines and other services for this purpose, practical considerations limit the potential impact of these efforts. For example, in some countries, WhatsApp's diverse user base leads to messages circulating in many languages. Fact-checkers may struggle to verify content in languages they are not fluent in.

The challenge of "zombie claims"[144] —false information that will not die no matter how many times it has been debunked previously—is not specific to WhatsApp. However, when unrelated visuals of old incidents, such as attacks or natural disasters, resurface on WhatsApp in hyperlocal groups, it can be difficult to verify due to limited publicly available information. Furthermore, when misinformation that has already been fact-checked resurfaces on WhatsApp, fact-checkers often struggle to reintroduce their previously written fact-checks and prebunk misinformation before it goes viral again.

AI, including generative AI, is being explored as a potential solution to some of these challenges. For example, some fact-checking organizations are using chatbots to interrupt the flow of

141 Clifford C (2025) Fact-checking on WhatsApp in Africa: Challenges and opportunities. In: Udupa S and Wasserman H (eds) *WhatsApp in the World.*

142 Clifford C (2025) Fact-checking on WhatsApp in Africa: Challenges and opportunities.

143 Clifford C (2025) Fact-checking on WhatsApp in Africa: Challenges and opportunities.

144 Khourie T. Africa Check's guide to zombie claims: how to spot false information that just won't die. Africa Check. Available at: https://africacheck.org/fact-checks/guides/africa-checks-guide-zombie-claims-how-spot-false-information-just-wont-die (accessed 23 January 2025).

disinformation, improve the workflow of fact-checkers, and provide media literacy tools to users. This approach makes it easier to handle large volumes of requests while users benefit from faster verification of claims. The "Chatbot for WhatsApp", developed by independent fact-checking organization Maldita.es, provides a useful model. It can detect, interpret, and respond to user reports of disinformation in all formats.[145] In Brazil, fact-checkers use the system to save time on repetitive technical tasks.[146] This has also allowed fact-checking organizations to create their own database of fact-checks, which are sent directly to the user when the system identifies a claim match.

Another area in which AI is relevant for fact-checking on WhatsApp is automated content generation. While publishing fact-checking articles is the accepted practice, engaging with audiences on WhatsApp calls for a more creative and personal approach. Some fact-checking organizations, such as Lead Stories and Africa Check, have used AI to efficiently create short and engaging video summaries of fact-checks.[147] Research in Brazil has shown the use of WhatsApp chatbots that can reply to messages automatically on public WhatsApp groups.[148]

While there are other ways in which AI could potentially be used to fact-check on WhatsApp, several barriers make it difficult for fact-checking organizations to explore opportunities, including a lack of technical know-how, data capabilities, and funding.

Over 50 International Fact-Checking Network accredited organizations worldwide maintain an active presence on WhatsApp through dedicated communications lines, but many have yet to integrate AI tools into their workflow. WhatsApp is actively developing AI capabilities, such as the WhatsApp AI Studio. However, it is primarily focused on providing users with access to AI chatbots for a range of tasks rather than fact-checking services in particular.[149] The Brazilian government banned Meta AI on WhatsApp in July 2024,[150] citing the need to protect users' data privacy in the face of swiftly evolving misuses of AI. However, after attending to the requests of the National Data Protection Authority, Meta AI is currently testing AI systems on WhatsApp.[151]

145 Maldita.es (2021) Maldita.es' WhatsApp Chatbot to thrive a fact-checking operation on disinformation. European Press Prize. Available at: https://www.europeanpressprize.com/article/maldita-es-whatsapp-chatbot/ (accessed 09 January 2025).

146 https://meedan.com/post/meedan-welcomes-3-new-brazilian-partners (accessed 09 January 2025).

147 Lead Stories WhatsApp Channel: https://whatsapp.com/channel/0029VaKgGTN23n3gpyR5z800; Roger Wilco, AI-generated TikTok videos help Mzansi youth separate fact from fiction ahead of elections. Available at: https://www.businesslive.co.za/redzone/news-insights/2024-05-23-native-ai-generated-tiktok-videos-help-mzansi-youth-separate-fact-from-fiction-ahead-of-elections/ (accessed 23 January 2025).

148 ITS Rio (2018) Computational power: Automated use of WhatsApp in the elections. Available at: https://feed.itsrio.org/computational-power-automated-use-of-whatsapp-inthe-elections-59f62b857033 (accessed 2 January 2025).

149 WhatsApp Help Center. About AI Studio. Available at: https://faq.whatsapp.com/2229193694115919 (accessed 9 January 2025).

150 Han H J (2024) Brazil prevents Meta from using people to power its AI. Available at: https://www.hrw.org/news/2024/07/03/brazil-prevents-meta-using-people-power-its-ai (accessed 9 January 2025).

151 https://fusionchat.ai/news/unlocking-the-power-meta-ai-lands-on-whatsapp-in-brazil (accessed 09 January 2025).

AI AND ONLINE ENCRYPTED PLATFORMS

While AI technologies are being explored for fact-checking and automated dissemination of prosocial narratives, the potential impact of generative AI on social media platforms, including encrypted messaging platforms, is becoming increasingly evident. A recent study analyzing millions of messages in India during the country's biggest election revealed that while the prevalence of generative AI-created content was currently low (<1%), there were troubling issues where it was used, including misleading content, hate speech, and religious propaganda.[152] The potential impact of generative AI on WhatsApp can be found in three key areas: production, distribution, and belief.

- **Production:** Generative AI has the potential to create a level playing field for content production on WhatsApp. As the technology 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 agendas.
- **Distribution:** WhatsApp's decentralized nature and end-to-end encryption make it difficult to moderate content at scale and in multiple languages. As a result, groups with better-established distribution infrastructures may have an advantage in spreading their messages, whether genuine or misleading. The problem of AI-generated content and fakes compound when existing networks of grey operations of clickbait workers, influence operators, and political consultants—defined as "shadow politics"—distribute them to different segments of electoral constituencies (see also the section on microtargeting, p. 27).[153]
- **Belief:** WhatsApp users often place a high level of trust in the content they receive through the platform, making them more susceptible to believing and sharing false information. As it becomes harder to identify AI-generated content, the risk of users falling victim to disinformation and propaganda grows. Overall, while the current prevalence of generative AI-created content on WhatsApp is low, its potential impact on the platform cannot be ignored. As the technology advances, it is crucial to address the challenges posed by AI in content production, distribution, and belief.

152 Garimella K and Chauchard S (2024) How prevalent is AI misinformation? What our studies in India show so far. 5 June. Nature. https://doi.org/10.1038/d41586-024-01588-2

153 Udupa (2024) Shadow politics: Commercial digital influencers, "data," and disinformation in India.

CASE STUDIES

This section explores case studies from India, South Africa, and Brazil to provide a view of ground realities. Analysis of specific cases can shed light on aspects distinct to different geographies and support the development of evidence-based policy and technical solutions by platforms and regulators.

India

WhatsApp is one of the most widely used encrypted messaging platforms in India, with a reported monthly active user base of 530 million.[154] In India, an average WhatsApp user reportedly spends around 21.4 hours per month on WhatsApp.[155] The platform supports a total of 11 languages in India, including several local languages, which broadens its user base and reach.[156]

Over the years, numerous examples of extreme speech and disinformation[157] have circulated on WhatsApp, with only a small percentage being detected or reported, largely due to its end-to-end encryption feature.

Indian fact-checkers have been debunking disinformation on WhatsApp for the past decade. One of the most concerning trends is the spread of communal messages targeting religious minorities, particularly Muslims, which are often driven by disinformation. Among the most persistent narratives is the "Love Jihad" conspiracy theory. "Love Jihad" is a term used to describe the unfounded claim that Muslim men are deliberately luring non-Muslim women into marriage with the intent of converting them to Islam. This conspiracy theory has fueled communal violence and deepened religious divisions in India. One disturbing example of the harms of disinformation and extreme speech on WhatsApp is a case from September 2020.[158] A collage circulated on social media and WhatsApp, with images of an interfaith couple alongside an unrelated image of a dead body of a woman being recovered by the police. The

154 Hariharan S (2024) How Telegram is losing the battle to WhatsApp in India. The Hindu Businessline. 28 August. Available at: https://www.thehindubusinessline.com/how-telegram-losing-battle-whatsapp-india

155 https://www.indiatoday.in/technology/news/story/whatsapp-users-in-india-spent-21-3-hours-per-month-on-an-average-in-2020-report-1759371-2021-01-15 (accessed 31 January 2025).

156 WhatsApp Help Center. About the languages WhatsApp is available in. Available at: https://faq.whatsapp.com/873422324183264 (accessed 16 January 2025).

157 Ponniah K (2019) WhatsApp: The 'black hole' of fake news in India's election. BBC News. 6 April. Available at: https://www.bbc.com/news/world-asia-india-47797151

158 Alphonso A (2020) Photos of inter-faith couple peddled with false murder claim. BOOM. 2 September. Available at: https://www.boomlive.in/fake-news/photos-of-inter-faith-couple-peddled-with-false-murder-claim-9588.

message falsely claimed that Hindu women who marry Muslim men are eventually murdered by their husbands.

In September 2024, a gruesome murder case in Bengaluru, in the southern Indian state of Karnataka, quickly became another example of a criminal case falsely linked to the "Love Jihad" conspiracy theory by Indian media outlets.[159] The victim, a 29-year-old woman, was found dismembered, with her remains stored in a refrigerator at her residence. What started as a tragic crime took on a communal angle when some media reports falsely linked the victim's alleged extramarital relationship with a Muslim man. This claim, based on accusations from her estranged husband, spurred sensationalist news stories framing the incident as part of the "Love Jihad" conspiracy. Posts on social media and encrypted messaging platforms, including WhatsApp, further fueled the communal narrative, with circulating messages labeling the case as an example of "Love Jihad". However, as the investigation unfolded, it became clear that the claims were without merit, and the accusations were false.

These examples highlight how disinformation on WhatsApp has severe implications, targeting minority communities and individuals in India. WhatsApp has implemented measures to reduce the spread of such content, such as the introduction of the "forwarded many times" label[160] and a user can forward messages to up to five people or groups at a time. However, the platform's encrypted nature continues to shield much of this harmful activity from detection and can often—even after detection—spread uninterrupted. These cases also highlight how extreme speech on WhatsApp does not remain on the platform but spreads to other platforms and is misreported by media outlets. Though WhatsApp has restricted mass forwarding in India since 2018, bad actors have found ways to bypass these limitations.[161] For instance, unrelated videos of violent crimes, like stabbings or murders, have been falsely shared on the platform, claiming that the attackers were Muslim. A recent case from September 2024 illustrates this phenomenon. A graphic video showing a young boy stabbing a schoolgirl in Belgharia, in the eastern Indian state of West Bengal, went viral on WhatsApp and other platforms with the false claim that the attacker was Muslim. Both the attacker and the victim were later revealed to come from the Hindu community.[162]

WhatsApp has also been used to incite violence and deepen caste divides. In India's southern

159 Rizwan H (2024) Indian news channels give communal hue to gruesome Bengaluru murder. BOOM. 28 September. Available at: https://www.boomlive.in/explainers/indian-media-reporting-on-the-bengaluru-murder-is-full-of-communal-undertones-26599

160 WhatsApp. About forwarding limits. Available at: https://faq.whatsapp.com/1053543185312573 (accessed 21 January 2025).

161 Hern A (2018) WhatsApp to restrict message forwarding after india mob lynchings. *The Guardian.* 20 July. Available at: https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings

162 Alphonso A (2024) Viral video falsely shared as Muslim youth stabbing Hindu girl in West Bengal. BOOM. 10 September. Available at: https://www.boomlive.in/fact-check/fake-news-video-muslim-man-stabbing-hindu-school-girl-love-jihad-factcheck-26424

state of Tamil Nadu, for instance, caste groups have used WhatsApp audio messages to spread hatred and mobilize support for violence.[163] In September 2015, several audio messages went viral on the platform, igniting a divide between the Paraiyars and Kallars caste groups. Paraiyars in Tamil Nadu are Dalits, and Kallars consider themselves superior to Dalits in the caste system.

Similarly, the platform has also been weaponized to stalk and harass women. In a case from the capital of Delhi in April 2022, a perpetrator was arrested after he was accused of allegedly harassing over 150 women through WhatsApp and fake Instagram accounts.[164] The accused searched for the mobile numbers of women through dating or friendship apps and then reached out to them on WhatsApp. According to the police, after his advances were rejected, he created obscene images of the women and threatened to circulate them on social media.

Indian fact-checking organizations such as BOOM rely on users reporting suspected disinformation to their WhatsApp tiplines or patterns seen across public-facing social media platforms to estimate how widespread a piece of disinformation is. Often, captions that accompany viral visuals on WhatsApp are similar to those found on other platforms, helping fact-checkers identify common trends. Despite WhatsApp's restrictions on forwarded messages and efforts to ban suspicious accounts, the platform remains a hotbed for disinformation, with real-world consequences, especially for minoritized communities.

South Africa

South Africa has a history of racial segregation, the effects of which are still felt today. Among the countries for which data is available, the World Bank ranks South Africa as the most unequal.[165] High levels of poverty, unemployment, and crime persist, leaving many dissatisfied with the quality of life in the country.[166] Public debate about who is to blame for these challenges is rife and rarely guided by factual information. South Africa's media industry is under-resourced, and many credible news organizations have erected paywalls,

163 Ramanathan S (2015) WhatsApp helps TN caste-groups spread hatred, mobilise support for violence. The News Minute. 14 September. Available at: https://www.thenewsminute.com/tamil-nadu/whatsapp-helps-tn-caste-groups-spread-hatred-mobilise-support-violence-34292

164 Haider T (2022) Delhi Police arrests cyberstalker for harassing over 150 women through fake social media accounts. India Today. 7 April. Available at: https://www.indiatoday.in/crime/story/delhi-police-arrests-cyberstalker-for-harassing-over-150-women-through-fake-social-media-accounts-1934500-2022-04-07

165 World Bank (2022) Inequality in Southern Africa: An assessment of the Southern African Customs Union. Report. Washington.

166 Martin G (2022) Poverty and inequality are a national security risk to South Africa. Defence Web. Available at: https://www.defenceweb.co.za/featured/poverty-and-inequality-are-a-national-security-risk-to-south-africa/ (accessed 11 January 2025).

making access to accurate reporting expensive and forcing South Africans to seek out lower-quality information, often on social media.[167] In this context, the ground for disinformation to flourish is fertile.

Disinformation in South Africa exploits existing fault lines in the country. It often emphasizes race or nationality, capitalizing on sensitivities around these topics. For example, in April 2020, the hashtag #putsouthafricansfirst became prominent online.[168] It was started by a network of accounts that shared and engaged with narratives that tapped into South Africans' discontent with crime, unemployment, and poor service delivery. While the content touched on real issues, disproportionate blame was placed on foreign nationals. The hashtag was often used to share videos and images taken out of context. In one instance, an account posted a photo of a crowded hospital with patients sleeping on the floor and claimed that South African patients were suffering because foreign nationals were taking up hospital beds. A reverse image search revealed that the photo was taken at a Nigerian hospital almost 16 months before.[169]

With many South Africans seeking employment or ways to earn an income, financial and job scams are common. These scams directly solicit money or ask for personal information, which can then be used to commit identity fraud. Health disinformation is also widespread, ranging from home remedies for common ailments to conspiracies about shadowing elites seeking to control the population.[170]

Much of this disinformation circulates on WhatsApp. Of the 45.34 million people who use the internet in South Africa, 94 per cent are on WhatsApp.[171] Due to the high cost of mobile data, the availability of "data bundles" specifically for WhatsApp, and the popularity of inexpensive smartphones that use an Android operating system, it remains the dominant messaging platform in the country.[172]

This leaves millions of South Africans vulnerable to hoaxes, fabrications, and conspiracy

167 Finlay A (2019/20) State of the newsroom. Report. Wits Journalism. South Africa.

168 Le Roux J (2021) What's the harm in a hashtag? Spotting disinformation in the wild. Africa Check. Available at: https://africacheck.org/fact-checks/reports/whats-harm-hashtag-spotting-disinformation-wild (accessed 11 January 2025).

169 Le Roux (2021) What's the harm in a hashtag? Spotting disinformation in the wild.

170 Business Tech (2024) WhatsApp hacking warning in South Africa. Available at: https://businesstech.co.za/news/internet/795187/whatsapp-hacking-warning-in-south-africa/ (accessed 11 January 2025); Kirsten C (2023) Exposing health myths: How sneaky science misled the public in 2023. Africa Check. Available at: https://africacheck.org/fact-checks/blog/exposing-health-myths-how-sneaky-science-misled-public-2023 (accessed 11 January 2025).

171 McInnes K (2024) South African Digital & Social Media Statistics 2024. Meltwater. Available at: https://www.meltwater.com/en/blog/social-media-statistics-south-africa (accessed 11 January 2025).

172 Labuschagne H (2024) How WhatsApp became dominant in South Africa. My broadband. Available at: https://mybroadband.co.za/news/software/524910-how-whatsapp-became-dominant-in-south-africa.html (accessed 11 January 2025).

theories across a range of topics, including governance and service delivery, crime and justice, health, and the environment. Disinformation is not the only form of harmful content that circulates on WhatsApp; extreme speech has also proven to be a problem.

South Africa's Prevention and Combating of Hate Crimes and Hate Speech Act makes it an offence for any person to intentionally publish, propagate, advocate, share, or communicate anything that could reasonably be construed to demonstrate a clear intention to be harmful or to incite harm and to promote or propagate hatred based on defined grounds. The law also makes it an offence when speech is intentionally distributed or made available in electronic communication.[173]

In recent years, several high-profile cases appeared to necessitate such a law. In 2018, a South African man, Adam Catzavelos, filmed himself on a beach in Greece, celebrating that no black people were present. He used a derogatory and racist term that the Constitutional Court, two years earlier, said was "the worst insult that can ever be visited upon an African person in South Africa".[174] The video, originally shared with a few of Catzavelos's friends on WhatsApp, was leaked and quickly went viral on social media.

Catzavelos pleaded guilty to and was convicted of a hate crime for the racist rant. The South African Human Rights Commission also lodged a complaint with the Equality Court on the basis of hate speech.[175] Catzavelos agreed to pay a fine of R150,000 (about 8,000 USD) over a period of 30 months and issued a public apology for his comments. However, the now well-known story has done little to deter the spread of extreme speech on WhatsApp, with incidents regularly reported in the media.[176]

While extreme speech on WhatsApp is largely tied to issues of race, xenophobic rhetoric has also been used to mobilize violence and hatred against minority groups. Anti-foreigner sentiment has been a pressing problem in South Africa for many years. In 2022, the UN

173 The Presidency, President Ramaphosa assents to law on the prevention and combating of hate crimes and hate speech. Available at: https://www.thepresidency.gov.za/president-ramaphosa-assents-law-prevention-and-combating-hate-crimes-and-hate-speech (accessed 23 January 2025).

174 South African Revenue Service v Commission for Conciliation, Mediation and Arbitration and Others (2016) Southern African Legal Information Institute, 38.

175 South African Human Rights Commission (2019) SAHRC takes Adam Catzavelos racism case to Equality Court. Available at: https://www.sahrc.org.za/index.php/sahrc-media/news/item/2033-sahrc-takes-adam-catzavelos-racism-case-to-equality-court (accessed 11 January 2025).

176 See https://www.news24.com/news24/investigations/joburg-chief-prosecutor-should-be-axed-for-hate-speech-xenophobia-and-cooking-the-books-inquiry-20220208; https://www.algoafm.co.za/domestic/former-spca-employee-pleads-guilty-to-publishing-hate-speech-on-whatsapp; https://www.sahrc.org.za/index.php/sahrc-media/news/item/3704-anti-gay-whatsapp-group-lands-shopkeeper-in-more-trouble; https://www.iol.co.za/news/south-africa/free-state/nketoana-municipality-director-under-fire-for-kill-the-boer-comment-in-management-whatsapp-group-de1d38dd-536b-44af-83a7-d673693f9fca; https://www.citizen.co.za/news/bolhuis-mum-as-show-cancelled-over-alleged-racial-slurs/

called on the government to act against "escalating violence against foreign nationals".[177] In September 2019, when xenophobic unrest flared up in South Africa's Gauteng province, a number of graphic images and videos, supposedly showing violence against foreign nationals, circulated widely online, including on WhatsApp. While the content was real, it was unrelated to the outbreak and had been used out of context, inflaming tensions.[178] Videos collected from Nigerian, Zimbabwean, and Congolese community WhatsApp groups showed extreme speech against foreign nationals or direct threats warning foreign nationals to leave the country or face attack.[179] More broadly, WhatsApp has become a platform through which discriminatory stories and conspiracies about migrants can be shared. Hashtags such as #PutSouthAfricansFirst and #ZimbabweansMustFall are frequently used to distribute posts that blame migrants for the country's socio-economic ills.[180]

Although WhatsApp has introduced measures, such as forwarding and group size limits, to help combat the spread of disinformation and extreme speech, the platform's infrastructure lends itself to continued issues in this area. In countries like South Africa, where media literacy levels remain low, the above examples demonstrate how real issues are tainted with dangerous agendas and then spread with relative ease on social media, including WhatsApp.

Brazil

Research suggests that 68 per cent of Brazil's population uses WhatsApp as their primary form of communication.[181] The platform's reach is likely larger since the total population is 212 million,[182] and 99 per cent of Brazilians access the internet via mobile phones, while 58 per cent do so via internet offered by telecom companies.[183] WhatsApp, known colloquially

177 United Nations (2022) South Africa: UN experts condemn xenophobic violence and racial discrimination against foreign nationals. Available at: https://www.ohchr.org/en/press-releases/2022/07/south-africa-un-experts-condemn-xenophobic-violence-and-racial (accessed 11 January 2025).

178 Clifford C (2019) Think before you share! Old, misleading videos said to be of xenophobic violence in SA are going viral. Africa Check. Available at: https://africacheck.org/fact-checks/reports/think-you-share-old-misleading-videos-said-be-xenophobic-violence-sa-are-going (accessed 11 January 2025).

179 Fokou G, Yamo A, Kone S, et al (2022) Xenophobic violence in South Africa, online disinformation and offline consequences. *African Identities* 22(4) 943–962. DOI: https://doi.org/10.1080/14725843.2022.2157245

180 Digital Action (2023) A recipe for disaster: Xenophobic hate on social media in South Africa. Global Coalition for Tech Justice. Available at: https://yearofdemocracy.org/case-study/a-recipe-for-disaster-xenophobic-hate-on-social-media-in-south-africa/ (accessed 11 January 2025).

181 Reuters Institute (2024). Reuters Institute Digital News Report 2024. Oxford: University of Oxford. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024 (accessed 20 January 2025).

182 IBGE (2024) Brazil's population reaches 212.6 million. Secretaria de Comunicação Social. Available at: https://www.gov.br/secom/en/latest-news/2024/08/ibge-brazils-population-reaches-212-6-million

183 PNAD (2024) Internet was accessed in 72.5 million Brazilian households in 2023. Agência de Notícias - IBGE. Available at: https://agenciadenoticias.ibge.gov.br/en/agencia-news/2184-news-agency/news/41029-internet-was-accessed-in-72-5-million-brazilian-households-in-2023

by Brazilians as "zap zap", acts as a problem solver, useful not only to access information[184] or conduct small business[185] but also to exchange messages in real-time with friends, family, and colleagues.[186] It is the preferred channel for intimate communication due to low-cost access and zero-rating policies.[187] While telecom companies allow Brazilians to access WhatsApp "for free", they also create a large dependence on WhatsApp as a digital infrastructure.[188]

In order to understand the challenges of disinformation and extreme speech on WhatsApp in Brazil, a comparison can be drawn between prevalent scenarios during the 2018 and 2022 general elections. WhatsApp played a key role in the rise of digital authoritarianism in Brazil. Public and private groups have allowed for weaponized strategies to spread malicious content to citizens with the help of other social media platforms such as Kwai and TikTok.[189]

In 2018, Bolsonaro's campaign efficiently spread disinformation on WhatsApp as opposed to traditional mainstream television campaigns.[190] The campaign appealed to those dissatisfied with previous governments and to economically poor communities in Brazilian society. They were reached via WhatsApp groups, including private family groups where Brazilians experience interpersonal trust and exchange all types of everyday information. Distribution of disinformation via WhatsApp bypassed Brazilian electoral law.

One of the most prominent examples of disinformation claimed that Fernando Haddad (presidential candidate for the left-leaning worker's party) distributed "baby bottles with penis-shaped nipples" at kindergarten schools while serving as Minister of Education.[191] The false claim spread in image and text format from Facebook to private and public WhatsApp groups. The social impact was so tremendous that the Ministry of Education had to clarify that the picture was fake. Another prominent example of disinformation was the so-called "gay kit". Bolsonaro claimed on his social media accounts that Haddad had

184 Gil De Zúñiga H, Ardèvol-Abreu A and Casero-Ripollés A (2021) WhatsApp political discussion, conventional participation and activism: Exploring direct, indirect and generational effects. *Information, Communication & Society* 24 (2): 201–18. https://doi.org/10.1080/1369118X.2019.1642933

185 Lapowski I (2024) How WhatsApp ate the world. Rest of World. Available at: https://restofworld.org/2024/how-whatsapp-for-business-changed-the-world/

186 Barbosa (2021) COMUNIX WhatsAppers: The Community School in Portugal and Spain.

187 Lorenzon L (2021) The high cost of 'free' data: Zero-rating and its impacts on disinformation in Brazil. Data-Pop Alliance. Available at: https://datapopalliance.org/the-high-cost-of-free-data-zero-rating-and-its-impacts-on-disinformation-in-brazil/; Belli L (2017) Net neutrality, zero rating and the minitelisation of the internet. *Journal of Cyber Policy* 2 (1): 96–122. https://doi.org/10.1080/23738871.2016.1238954

188 Barbosa S (2021b) WhatsAppers for social good: Local community response to Covid-19 in Brazil. ICLD. 22 November. Available at: https://icld.se/en/publications/sergio-barbosa-2021-whatsappers-for-social-good-local-community-response-to-covid-19-in-brazil/

189 Data from WhatsApp public groups were collected during September-October 2022. These domains were analyzed through a Palver WhatsApp monitor that collected data across 26 states plus the Federal District in Brazil.

190 Bolsonaro had brief appearances on open and public TV: 8 seconds and 11 short insertions daily. Alckmin, currently the vice-president, had the longest time: 5 minutes and 32 seconds and 434 insertions daily. See: https://www.hannaharendt.net/index.php/han/article/view/429/565

191 Barbosa and Back (2020) The dark side of Brazilian WhatsAppers.

created a "gay kit", which he supposedly introduced in primary schools so that children as young as six years old would be encouraged to "become gay". This claim mainly circulated on evangelical WhatsApp groups.[192] As former Minister of Education, Haddad had been involved in educational initiatives against homophobia, but pro-Bolsonaro groups twisted this and distorted the purpose of the educational materials.[193] Bolsonaro himself appeared on the most traditional Brazilian TV news program, "Jornal Nacional", and showed a book he claimed was proof of the existence of the "gay kit". Later, a TV presenter revealed Bolsonaro's claims were false.[194]

In January 2019, Bolsonaro became Brazil's first far-right president. He was also the first president following the 1988 transition to receive an absolute majority in the Brazilian parliament and the strong support of local councilors and mayors aligned with his political party. His campaign was polarized and fragmented, mobilizing a sense of disappointment with the workers' party.

By comparison, the 2022 general election between leftist candidate Lula da Silva and Bolsonaro was a tense battle. With experts reporting that Bolsonaro posed a threat to democracy,[195] all signs pointed to him losing the election. Bolsonaro, however, insisted the polls were wrong and that he was on track to win.

After losing the election, he invoked a rhetoric of voter fraud, denying the election results. Brazilians faced an insurrection on 8 January 2023—mirroring the one the US suffered in January 2021 when a mob of Donald Trump supporters stormed the US Capitol. On the day the election results were announced, a series of riots broke out across Brazil, mainly organized on Telegram and WhatsApp groups. For example, Bolsonaro's supporters' "riot" was first announced and shared on WhatsApp groups. More than 1,800 people were later detained after causing serious damage to government buildings.

192 Davis S and Straubhaar J (2020) Producing antipetismo: Media activism and the rise of the radical, nationalist right in contemporary Brazil. *International Communication Gazette* 82 (1): 82–100. https://doi.org/10.1177/1748048519880731

193 Ozawa, Woolley, Straubhaar, Riedl, Joseff and Gursky (2023) How disinformation on WhatsApp went from campaign weapon to governmental propaganda in Brazil.

194 For more details, see: https://www.aljazeera.com/opinions/2018/10/31 bolsonaro-gender-ideology-and-hegemonic-masculinity-in-brazil

195 Barbara V (2022) Bolsonaro is afraid of going to prison, and he's right to be. *The New York Times.* 8 August. https://www.nytimes.com/2022/08/08/opinion/bolsonaro-brazil-prison-election.html

Pro-Bolsonaro groups also circulated inaccurate videos that claimed to show suspicious electronic voting machines to promote the false narrative that the elections favored Lula over Bolsonaro. This narrative was further aggravated after the election, with standstills on highways throughout Brazil, in which truck drivers would organize protests and street actions to reclaim Bolsonaro's victory. Some shared messages and videos on public WhatsApp groups suggesting support from the military and police. According to circulating messages, the population would wait for 72 hours to claim Article 142 of the Brazilian Constitution, which would order the arrests of court officials, suspend parliament, and transfer power to the military, as Bolsonaro supporters executed a "state of siege".[196] As the examples from the Brazilian case show, WhatsApp weaponization has been combined with a cross-platform communication strategy to target various sections of Brazilian society.

196 For more details, see https://factcheck.afp.com/doc.afp.com.32QY3RK

RECOMMENDATIONS

Finding ways to make online encrypted messaging platforms safe and secure for users while protecting human rights and democratic values is critical. This report has discussed how encrypted messaging platforms are located within complex structures of power, social norms, and political cultures, even as they are intertwined with technological architectures and corporate policies. In the context of Meta's recent changes in content moderation policies and the continued importance of encrypted messaging platforms such as WhatsApp, especially in the Global South, this report proposes a set of measures for governments, platforms, and civil society to address extreme speech and disinformation. They highlight the need for developing approaches that are grounded in lived realities of specific contexts and international human rights standards. They call for close knowledge of diverse and dynamic social and political practices that have emerged around encrypted messaging platforms, which often contradict promises of privacy and secure communication signaled by encryption technology as well as undermine regulatory efforts with the clever use of campaign tactics. At a time when platforms are rolling back trust and safety protocols, this report serves as yet another call to take platform governance and content moderation seriously while also cautioning that removing encryption is not a solution to address extreme speech and disinformation.

Below, we provide a list of measures for a contextualized and qualified approach to encryption, addressing different stakeholders, challenges, and opportunities. Multiple stakeholders, with the support of UN entities and other multilateral agencies, should focus on finding whole-of-society solutions to online harms and challenges. This means working with relevant expert groups, civil society, and the technical community to develop and implement technical and nontechnical solutions which are lawful, necessary, proportionate, and informed by expert opinion.

We first outline general categories of intervention, followed by key steps within each category.

The general categories are as follows:

- Platform governance
- Mitigating digital influence operations
- Supporting research
- Strengthening fact-checking
- Awareness raising and capacity building
- Leveraging artificial intelligence responsibly

Platform governance

Encrypted messaging has offered the possibility of safe and secure communication for fact-checkers, journalists, political critics, and marginalized communities. However, in authoritarian and non-democratic regimes, encryption offered by the platforms is not always a guarantee that they will not be harassed, threatened, or forced to divulge data or the identity of sources. Encrypted messaging platforms have also been used for spreading extreme speech and disinformation. Governments, platforms, and civil society need to work together to protect human rights online, including on encrypted platforms and offline.

Key steps:

- **Safeguards against limiting encryption:** Overly broad, informal, or extra-legal measures to access user information and content on encrypted platforms, as highlighted in this report, immediately undermine the safety and security that these spaces can offer. Governments should not use spyware and should ensure surveillance practices and other measures towards platform accountability adhere to international human rights standards, including the principles of necessity, legality, and proportionality. Governments can commit to principles such as the Necessary and Proportionate Principles[197] and the Freedom Online Coalition Principles on Government Use of Surveillance Technologies.[198]
- **Due process:** Create or strengthen existing legal frameworks that provide remedies. For example, to protect against indiscriminate takedowns and infringement of freedom of expression, Article 18 of the DSA proposes the establishment of certified dispute settlement bodies to which online users can lodge complaints and seek redressal after failing to find redressal through platforms' internal complaint procedures. Similar measures can be explored in the context of encrypted messaging globally. Platforms should also prioritize messages that contain content that is clearly illegal as well as content that is harmful, including exclusionary extreme speech and gendered harassment.
- **Alignment with broader principles:** The regulation of encrypted messaging platforms can align with the broader principles of platform governance as enunciated by regulations such as the DSA while taking into account context and assumptions of "data universalism".[199] While being aware of regulatory overreach and misuse that can occur when regulations developed in Western democracies are copy-pasted for repressive agendas, regulatory systems for encrypted messaging in the Global South should explore context-appropriate measures and build on existing traditional media regulation infrastructures for a holistic regulatory approach.

197 https://necessaryandproportionate.org/principles/

198 https://freedomonlinecoalition.com/guiding-principles-on-government-use-of-surveillance-technologies/

199 Loukissas Y A (2019) *All data are local: Thinking critically in a data-driven society.* Cambridge MA: MIT Press.

- **Monitoring business models:** Multistakeholder and publicly accountable regulators should monitor advertising and other revenue-generating models of platforms, and whether business APIs, payment services (such as WhatsApp Pay), cross-advertising across different services of companies, and other monetization models are being manipulated for political propaganda and extreme speech.
- **Human rights due diligence:** In line with the UN Guiding Principles on Business and Human Rights,[200] platforms should conduct ongoing human rights due diligence of their services across the markets they operate in to understand and address emerging risks to human rights in different contexts.
- **Trust and safety and human rights teams:** Trust and 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. This is particularly important in light of multiple reports of budget and personnel cuts in these teams.[201] They should have adequate resources, including language capabilities, funding, and personnel, in all the countries they operate in. They should not circumvent this requirement in the Global South.
- **Code of conduct:** Encrypted messaging platforms should participate in applying a contextually responsive industry-wide code of conduct grounded in international human rights principles.
- **Metadata analysis and user reporting:** Rather than requiring content moderation that would undermine encryption, governments and platforms should explore interventions that do not undermine encryption, such as metadata analysis (see p. 25), and develop strong user reporting mechanisms in place to identify and address online harms.[202] However, user reporting is not without challenges. We discuss these limitations and propose steps to address them in the section below.

200 https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

201 https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435

202 See also https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf

User reporting and community moderation

As mentioned earlier, user reporting is currently the main form of content moderation on encrypted messaging platforms. Meta has taken this to a new level by introducing "Community Notes" as a mechanism to replace professional fact-checking across its social media platforms in the US; however, this is not applicable to WhatsApp.

While offering users the opportunity to evaluate content and exercise the choice to influence what they see on social networks, including messaging applications, user reporting has severe limitations. When users assume responsibility in moderation—as group administrators, community moderators, or fact-checkers—this comes with social implications. For example, on the one hand, if group admins have strong ties to the people in the group they are managing, they may be more permissive of content that violates rules. On the other hand, if a group consists of more weak-tie contacts, it may be easier for administrators to authoritatively enforce the rules.[203] Similarly, when individuals encounter false information shared in family WhatsApp chats, they may be more reticent to correct false content from individuals who are socially senior to them. In order to not embarrass them, they may end up not correcting at all or choose to correct in other, outside channels (see the section on deep extreme speech, p. 27). When community members take up fact-checker roles, this may have prosocial effects but can also introduce bias.[204]

Even more, in the context of systematic and organized campaigns to spread extreme speech and disinformation, the proposed "Community Notes" model of Meta and other platforms cannot fully ensure protection against harms. While user reporting infrastructures should be improved, organized disinformation campaigns that misuse and weaponize user reporting to overwhelm platform systems are not uncommon. The full reliance on community moderators raises several questions about how this system would be structured and the potential misuse of this infrastructure by bad actors, such as to delay rating posts with potential misinformation. X's "Community Notes" model has been criticized since for a note to become publicly visible, it requires consensus

203 Shahid, Agarwal and Vashistha (2024) 'One style does not regulate AI'.

204 Garimella K (2022) Community-driven fact-checking on WhatsApp: Who fact-checks whom, why, and with what effect? Available at: https://gvrkiran.github.io/content/WhatsApp_community_factchecking____ICWSM_May_2023.pdf; Pearce KE and Malhotra P (2022) Inaccuracies and izzat: Channel affordances for the consideration of face in misinformation correction. *Journal of Computer-Mediated Communication* 27(2). DOI: 10.1093/jcmc/zmac004

from people from across the political spectrum. Critics have observed that achieving this consensus in a partisan environment is "nearly impossible".[205]

Additionally, from the perspective of the platform interface, the steps to report content on WhatsApp groups to admins are cumbersome. The first condition is that the group admins must have turned on the "send for admin" review option.[206] Only after enabling this can users notify admins on a group about messages they want to report, and following this the admin can take a decision. Users can also report messages by reporting them to WhatsApp directly; however, fact-checkers, users, and civil society groups have experienced that there is no clear timeline for when the report is resolved.[207] Platform measures are critical, while community interventions can bring cultural context, especially to address the challenge of culturally coded images, videos, and texts that circulate on WhatsApp. Community interventions can also bring organic traction for moderation efforts.

Trusted flaggers: Since WhatsApp and other encrypted messaging platforms are promoting "Channels" and because of the popularity of public WhatsApp groups, the regulatory mechanism of "trusted flaggers" should be explored, provided that a robust mechanism for an independent, publicly accountable regulatory authority is in place to protect against intentional or unintentional bias and platforms are transparent about content removed based on reports from trusted flaggers. For example, the DSA requires providers of internet hosting services to implement "user-friendly notice and action mechanisms" and internal complaint handling systems through which users can report violations. Through the category of "trusted flaggers", the regulation proposes to expedite this process for the greater public good. Online platforms are obligated to process and decide on the notices submitted by trusted flaggers "on priority and without delay". Trusted flagger status is "awarded to entities and not individuals that have demonstrated...that they have particular expertise and competence in tackling illegal content, that they represent collective interests and that they work in a diligent and objective manner".[208] Such measures

205 Mahadevan A (2025) Meta will attempt crowdsourced fact-checking. Here's why it won't work. Poynter. 5 January. Available at: https://www.poynter.org/commentary/2025/meta-community-notes-crowdsourced-fact-checking-x/

206 "How to send a message for admin review", WhatsApp. Available at: https://faq.whatsapp.com/286279577291174/?cms_platform=android&helpref=platform_switcher (accessed 23 January 2025).

207 In 2021, Pro Republica reported that, "WhatsApp reviewers have three choices when presented with a ticket for either type of queue: Do nothing, place the user on "watch" for further scrutiny, or ban the account. (Facebook and Instagram content moderators have more options, including removing individual postings."https://www.propublica.org/article/how-facebook-undermines-privacy-protections-for-its-2-billion-whatsapp-users

208 https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (accessed 23 January 2025).

are intended to complement rather than replace professional fact-checking (see the section on fact-checking, p. 31).

Misuse of user reporting: WhatsApp and other encrypted messaging platforms should implement necessary safeguards and monitoring mechanisms to prevent misuse of user reporting and community moderation by vested interests, including ruling governments.

Improve user reporting: Platforms should work with and fund civil society initiatives and researchers to develop infrastructure to support user reporting, community moderators, and admins through culturally relevant tools, such as templates that can be used to identify and communicate to group members when content is violating of guidelines or culturally appropriate nudges that can be shared by group admins and community moderators.[209] The interface for user reporting should be easily accessible on the platform.

Information verification tools: WhatsApp should revive and extend initiatives like the "Search the web" function.[210] Tested in March 2020, the feature gave users the ability to quickly search the web for the text or image they had received for more context. This tool could empower users to verify information and provide necessary context.

Community moderators: Community moderators should be supported through clear and transparent rules, recognition, and responsive communication with companies. They should be protected against repressive state actions. Platforms should develop a robust system to monitor threats and misuses. In doing so, they should consider anti-hate and anti-disinformation initiatives of various UN entities and other multilateral agencies and civil society organizations that are aligned with international human rights standards.

Prosocial interventions: Beyond tools and interfaces, platforms should develop initiatives around user reporting and community-based content moderation to promote prosocial intervention strategies that mitigate bystander apathy and further collective action.[211]

209 Shahid, Agarwal and Vashistha (2024) 'One style does not regulate AI'.

210 Singh M (2020) WhatsApp tests new feature to fight misinformation: Search the web. 21 March. Available at: https://techcrunch.com/2020/03/21/whatsapp-search-web-coronavirus/

211 Riedl M (2020) Content moderation and volunteer participation. In M Baker, BB Blaagaard, H Jones and L Pérez-González (eds) *The Routledge encyclopedia of citizen media* (pp. 93–98). Routledge; Matias, JN (2019) The civic labor of volunteer moderators online. *Social Media + Society* 5(2) 1–12. https://doi.org/10.1177/2056305119836778; Friess D, Ziegele M and Heinbach D (2021) Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication* 38(5) 624–646. https://doi.org/10.1080/10584609.2020.1830322; Draper NA (2019) Distributed intervention: Networked content moderation in anonymous mobile spaces. *Feminist Media Studies* 19(5) 667–683. https://doi.org/10.1080/14680777.2018.1458746

Mitigating digital influence operations

As this report has outlined, political weaponization of WhatsApp Channels, microtargeting and segmentation, coordinated manipulation, and gender-based violence are constantly evolving on encrypted messaging platforms. Political propaganda spread through family and community networks is a striking feature of WhatsApp communication. The role of commercial players who offer "disinformation services" has amplified the instrumental use of WhatsApp channels for political agendas in different global contexts.

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 in encrypted messaging platforms, including WhatsApp, through the use of grey networks, clickbait operators, and digital influencers. Defined as "industry/actor accountability",[212] this type of regulatory intervention entails implementing stricter rules to "ensure transparency in election expenditure, regulation of campaign finance, professional code of conduct and co-regulatory models for digital influence operations".[213]

At the same time, following an assessment of systematic risks that arise from manipulative digital influence operations, platforms should implement risk mitigation measures. Such processes should allow for independent expert verification.

Key steps include:

- **Transparency measures:** Platforms should implement robust transparency standards that require them to publish reports that inform policymakers, users, regulators, and researchers about how they moderate content, carry out proactive investigations and act when notices are issued by governments, and monitor political advertising and coordinated manipulation. While Meta publishes quarterly reports which include Adversarial Threat Reports, platforms should streamline regular reports which are accessible and available for all the regions they operate in.
- **Collaborate and coordinate across platforms:** Platforms can explore mechanisms for sharing identified harmful content, similar to hash-sharing databases, and best practices among each other, as well as in close collaboration with outside organizations such as researchers, journalists, and fact-checkers.

212 Ong J C (2021) Southeast Asia's disinformation crisis: Where the state is the biggest bad actor and regulation is a bad word. *Items, Social Science Research Council.* https://items.ssrc.org/ disinformation-democracy-and-conflict-prevention/ southeastasias-disinformation-crisis-where-the-state-is-the-biggest-badactor-and-regulation-is-a-bad-word/; Caplan R (2018) Content or Context Moderation? Data & Society Research Institute. https://datasociety.net/ library/content-or-context-moderation/ (accessed 3l October 2021)

213 Udupa (2024) Shadow politics: Commercial digital influencers, "data" and disinformation in India, p.9.

Supporting research

To ensure regulation, policy, interventions, and awareness campaigns by governments, platforms, and civil society are relevant and reflect on-the-ground realities, research into the use of online platforms, the spread of harm across platforms, and the effectiveness of interventions needs to be supported.

Key steps include:

- Governments should develop legal frameworks for researcher access to data, including data donation initiatives.[214]
- Platforms should provide access to data and support research through funding without interfering in its outcomes. Grants need to be awarded equitably across jurisdictions. At the same time, such funding schemes should not be used as a way to evade other platform governance measures detailed in this report.
- Platforms should share vital information that can be useful for researchers while maintaining user privacy. This includes transparency around internal moderation practices and design interventions that the platform implements to curb harmful content.[215]
- Platforms should provide researchers access to viral content—specifically, content that has surpassed a predefined exposure threshold, such as 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. Research support for programs such as CrowdTangle should be reinstated.
- Multiple stakeholders should come together to build capacity and support research and researchers working in the field of extreme speech and disinformation studies. Lessons can be learned from the creation of institutional committees at universities to fight disinformation.[216] Efforts to undermine the field of extreme speech and disinformation studies through the instrumental use of "free speech" and other discourses should be monitored and challenged by means of stakeholder engagements, especially at various UN bodies and other multilateral agencies.

214 Garimella K and Chauchard S (2024) WhatsApp explorer: A data donation tool to facilitate research on WhatsApp. arXiv preprint arXiv:2404.01328

215 Barbosa S and Milan S (2019) Do not harm in private chat apps: Ethical issues for research on and with WhatsApp. *Westminster Papers in Communication and Culture* 14(1) 49–65.

216 University of Brasilia launched the committee to fight disinformation in January 2025. https://noticias.unb.br/institucional/7782-unb-lanca-comite-de-enfrentamento-a-desinformacao

Strengthening fact-checking

Fact-checkers are critical actors in bringing contextual understanding to content on online encrypted platforms. However, they face a number of challenges when operating on encrypted messaging platforms, including the cross-platform spread of content, reliance on user reporting, the need to work in multiple languages, and the need to scale their work.

Key steps include:

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

Awareness raising and capacity building

Awareness raising and capacity building are key tools in shaping end-user actions on a platform and addressing contextual nuances of how they produce, consume, and share content.

Key steps include:

- Multiple stakeholders need to support the creation of digital literacy educational initiatives as mandatory components of educational curricula, with long-term goals, evaluation metrics, and prospective action plans. These efforts need to provide information about: how encryption works; reflect the realities of how information is created, shared, and consumed across online platforms, including encrypted messaging platforms; and include steps users can take to protect themselves. They should also include literacy around cross-media campaigns and sociohistorical contexts of speech.
- Literacy and educational initiatives specifically for digital influencers must be developed, which hone in on fact-checking skills.[217] These initiatives should take into account the challenges of engaging with influencers, including partisan interests and expectations of incentives.
- Address corporate monopoly in encrypted messaging and encourage the creation of alternative community-based and non-profit messaging applications, especially in the Global South.
- Civil society and researchers can support awareness raising through continued research that compares platform features and policies, such as Ranking Digital Rights, and research that documents the actual use of a platform across contexts and communities.

217 See some relevant links in the toolkit.

- Governments should provide funding for relevant authorities and prosecutorial offices so that they are technically equipped and trained to respond meaningfully to online harm and violence.

Leveraging artificial intelligence responsibly

AI has the potential to exacerbate harm through the creation and amplification of harmful synthetic content at scale. At the same time, AI can be an important tool to address disinformation and propaganda. It is, therefore, imperative that AI is leveraged responsibly.

Key steps include:

- Platforms/companies should support fact-checkers in leveraging AI to develop and share easily understandable and consumable 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 community moderators and fact-checkers with free access to such models.
- Initiatives such as the "AI4Dignity" project to create spaces for collaborative coding among AI developers, fact-checkers, and ethnographers should be implemented at local and national levels for AI-assisted content moderation.[218] This project has developed a process model of facilitated discussions and reflexive iterations to develop categories and training datasets that can address the cultural, linguistic, and political complexity of extreme speech contexts in a grounded way. Fine-tuned open-source models and interfaces to classify content that emerge through such initiatives should be made available for fact-checkers and anti-hate initiatives.
- An AI-enabled reporting mechanism can be integrated into platforms for flagging harmful content in multiple languages. This would allow users to report items in various regional languages, enhancing accessibility for non-English-speaking users. On WhatsApp, this can be integrated into Meta AI as currently, the "forwarded many times" tag on a message on WhatsApp only functions in English. The reported items could then be sent to a centralized interface available to fact-checkers.

218 Udupa S (2021) Digital technology and extreme speech.

DATA ACCESS TOOLS AND LIST OF RESOURCES

The non-exhaustive, indicative list below highlights some initiatives that provide useful resources and possible directions for further development of support structures.

Gathering WhatsApp data using data donations

To address the complex challenges of collecting WhatsApp data, one available tool developed by a group of researchers, including a co-author of this report, is WhatsApp Explorer (https://github.com/gvrkiran/WhatsAppExplorer). It helps researchers access WhatsApp data using data donations. The tool is designed to collect data from WhatsApp groups and individual chats for research purposes. It relies on a data donation model whereby users donate their WhatsApp data securely and anonymously. It is designed to simplify the data donation process while addressing the privacy, legal, and ethical concerns inherent in such research.

WhatsApp Explorer enables participants to donate their data with minimal effort. The approach primarily focuses on in-person interactions with "gateway users"—those who consent to share entire data (chat history) from groups they belong to—especially in regions like India and Brazil, where face-to-face engagement fosters trust. However, it is also possible to adapt this to an online donation process, making participation more flexible. Participants' privacy is paramount, and the tool ensures that no personal data, such as one-on-one conversations, is collected. All collected data is immediately anonymized using advanced algorithms, such as Google's Data Loss Prevention API, ensuring no sensitive information is accessible. Visual content undergoes additional anonymization processes to protect users' identities, ensuring a secure and ethical approach to WhatsApp data collection for social science research.

A schematic flow of the data collection flow is shown in Figure 2.

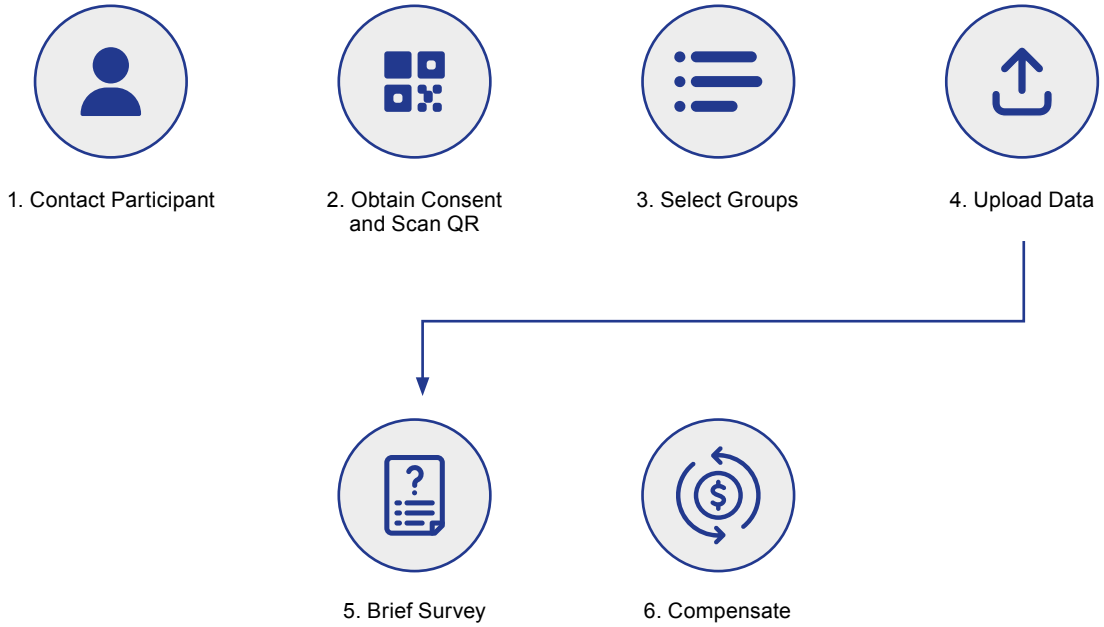

Figure 2: *Schematic flow of the data collection*

This open-sourced tool helps to gather and examine WhatsApp data. The code for the tool is accessible to academics upon request. Though technically the tool can democratize data collection from WhatsApp, even for academics with little technical background, in practice the tool still requires technical knowledge to set up and maintain. Additional funding could help in developing a research deployment framework where researchers interested in the tool could contribute resources (for example, computation, engineering time, or maintenance) to enable the tool to be open and accessible to everyone.

While the tool helps researchers gain valuable insights, it is not without its limitations. Its approach to studying WhatsApp communication is primarily quantitative, which may lead to the collection of data that isn't always relevant. One significant drawback is the tool's lack of selectivity in data gathering. Users are required to donate entire group conversations without the option to choose specific messages. This broad-brush approach expects to find disinformation that might be present even in seemingly harmless contexts, such as family group chats where private information might be shared. However, this strategy may not be the most effective or appropriate method for identifying and analyzing disinformation on the platform.

Fact-checkers, end-users, and whistleblowers

- **Fact-Checker WhatsApp Bot:** This tool, built using Flask, Twilio, and OpenAI API, serves as an SMS-based fact-checking tool that can evaluate the authenticity of information provided in user queries. https://github.com/Yashism/Fact-Checker-WhatsApp-Bot
- **On-Device Fact-Checking Solution:** Using advances in state-of-the-art techniques to find similar image and text messages, an on-device fact-checking solution could identify up to 40 per cent of the shares of potential misinformation in public WhatsApp groups while preserving end-to-end encryption if content can be prioritized appropriately and responded to quickly. https://misinforeview.hks.harvard.edu/article/can-whatsapp-benefit-from-debunked-fact-checked-stories-to-reduce-misinformation/
- **WhatsApp Tiplines:** These are dedicated services where users can forward potential misinformation for fact-checking. Tiplines have proven to be highly effective in identifying viral content on WhatsApp, often before it appears in large public groups. A list of International Fact-Checking Network-accredited organizations that offer fact-checking services on WhatsApp can be found here: https://faq.whatsapp.com/505912054085566 4

- **Deepfake Analysis Unit:** A tipline that aims to verify AI-generated misinformation on WhatsApp. https://www.dau.mcaindia.in/
- **Whistleblower Protection:** This non-profit initiative offers support for whistleblowers with relevant resources as well as engages in advocacy for their protection. https://thesignalsnetwork.org/about-us/
- **United Nations Development Program's iVerify Platform:** iVerify is a UNDP fact-checking platform to combat misinformation. It is a support package that has a range of open source digital tools for monitoring, fact-checking, and responding. It has been implemented in nine countries, including Zambia, Sierra Leone, and Pakistan. https://www.undp.org/digital/iverify

Companies

- Metadata Analysis: While respecting encryption, WhatsApp can utilize metadata analysis to identify and address online harms without undermining user privacy. https://misinforeview.hks.harvard.edu/article/research-note-tiplines-to-uncover-misinformation-on-encrypted-platforms-a-case-study-of-the-2019-indian-general-election-on-whatsapp/
- Dedicated Fact-Checking Channel: This proactive approach allows users to receive verified fact-checks, access awareness materials, and learn about spotting false information directly within WhatsApp.

Researchers

- Metadata from the WhatsApp tipline and public groups: These are available at: https://doi.org/10.7910/DVN/ZQWG02
- Researcher Support Consortium: https://researchersupport.org/
- Scicomm Support: https://scicomm-support.de/en/
- Democracia em Xeque Institute (IDX)[216]: https://en.institutodx.org/sobre/
- Bereia Collective[217]: https://coletivobereia.com.br/proposta-bereia/

Civil Society

- Senior citizens digital media literacy program: Program aiming to raise awareness among senior citizens. https://english.jagran.com/india/sach-ke-sathi-seniors-fact-check-training-held-for-senior-citizens-in-delhi-malviya-nagar-10210725
- Literacy and fact-checking training for influencers: Programs aiming to equip influencers with critical skills. https://journalismcourses.org/free-online-course-on-influencers-and-journalists-starts-with-8000-participants-from-149-countries-registration-is-still-open/ and https://www.redescordiais.org.br/en/
- Teen fact-checking awareness programs offered by fact-checking organizations: https://www.poynter.org/mediawise/programs/tfcn/ ; https://www.boomlive.in/tfcn

216 Brazilian Research Institute, with the mission of expanding the production of knowledge to fight disinformation campaigns, hate speech, and violent political extremism.

217 The Bereia Collective is a non-profit journalistic initiative focused on fighting misinformation in religious digital environments. Created in 2019, as a result of a study conducted at Federal University of Rio de Janeiro (UFRJ), the initiative focuses mainly on the evangelical segment of Brazilian society.

AUTHORS

Anmol Alphonso is a senior fact-checker at Boom Factcheck, India.

Sérgio Barbosa is a postdoctoral researcher at the Institute of Citizenship Studies (InCite), University of Geneva.

Cayley Clifford is Deputy Chief Editor of Africa Check.

Kiran Garimella is Assistant Professor in the School of Communication and Information, Rutgers University.

Elonnai Hickok is an independent expert on technology policy.

Martin Riedl is Assistant Professor at the School of Journalism and Media, University of Tennessee, Knoxville.

Erkan Saka is Professor at the Institute of New Media and Communication, Istanbul Bilgi University.

Sahana Udupa is Professor of Media Anthropology at Ludwig-Maximilians-Universität München (LMU Munich).

Herman Wasserman is Professor and Chair at the Department of Journalism, Stellenbosch University.