



When harm goes viral: Governing online harm and platform power in South Africa



**MOXII
AFRICA**



**RESEARCH
ICT AFRICA**

Digital harms and conflicts in South Africa

The forthcoming *When harm goes viral: Governing online harm and platform power* report examines three interlinked harm areas that define South Africa's online and offline conflict environment: Gender-based violence, xenophobia and elections.

Gender-based violence (GBV) is a national crisis: femicide rates are nearly five times the global average, with reported cases of rape exceeding 11 000 in a single quarter, according to the South African Police Service's (SAPS) crime statistics (Third quarter 2025-2026). Online misogynistic communities, the sharing of non-consensual intimate images (NCII), sexualisation of minors, and AI-generated "deepfake" abuse compound offline gendered violence, even as platforms like Twitter/X have also enabled resistance movements such as #MenAreTrash and #AmINext. In response, the South African government declared GBV a national disaster in 2025.

Meanwhile, xenophobia, rooted in anti-immigrant sentiment and recurrent violence since the major attacks of 2008, has been transformed by social media, which enables rapid coordination, the viral spread of fabricated crime statistics, and organised movements such as Operation Dudula, Put South Africans First and March and March, all of which blur the line between legitimate concern, hate speech and incitement to violence. This hostility is particularly directed towards migrants from elsewhere on the African continent. Most recently, anti-immigrant mobilisation and violence in 2026 disproportionately targeted African nationals, forcing tens of thousands to flee their homes and communities, and prompting large-scale repatriations to neighbouring African countries.

Threats to democracy and election integrity have grown alongside this increasingly competitive and fragmented political system. The 2024 election, in which the ANC lost its parliamentary majority for the first time, intensified competition for voters, with political actors increasingly using social media to mobilise support and, in some instances, exploit ethnic and other social tensions or call democratic processes into question. Social media has become a primary battleground for political contest, raising concerns not only about disinformation and coordinated or amplified information operations, but about their potential to undermine citizens' access to reliable information and erode trust in democratic institutions such as the Electoral Commission.

The Centre for Analytics and Behavioural Change's (CABC) longitudinal social media monitoring supports these findings. On GBV, CABC's *Countering Misogyny* and technology-facilitated GBV (TFGBV) research documents doxxing, non-consensual image sharing and child predation/sexualisation on mainstream platforms, often routed to private channels on WhatsApp and Signal, alongside deep-rooted

narratives blaming women for violence against them (including recent conversation around the lobola/bride-price custom). On xenophobia, recurring narratives scapegoat foreign nationals for crime, unemployment, strain on social services and food safety, despite being built on misinformation. On democracy, the analysis (detailed further in stakeholder interviews below) identifies coordinated inauthentic networks and disinformation, including operations funded by actors outside of South Africa.

Legal analysis and platform governance

The legal and regulatory analysis assesses South Africa's framework against international human rights standards, such as the International Covenant on Civil and Political Rights, the Rabat Plan of Action, the UN Guiding Principles on Business and Human Rights, and UNESCO's Guidelines for the Governance of Digital Platforms. Freedom of expression is strongly protected under Section 16 of the Constitution, limited only narrowly under Section 16(2) (war propaganda, incitement to imminent violence, advocacy of hatred constituting incitement to harm) and subject to the general limitation clause in Section 36. Constitutional Court jurisprudence (notably *Qwelane v South African Human Rights Commission* (SAHRC), which struck down the vague term "hurtful" from the Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) hate speech provision, and *Islamic Unity Convention v Independent Broadcasting Authority*) requires that restrictions on expression be narrowly tailored, proportionate, and reserve criminal sanction as a last resort, consistent with the Rabat Plan of Action.

The Prevention and Combating of Hate Crimes and Hate Speech Act 16 of 2023 has been gazetted but had not commenced by the time of the report, pending regulations. Technology-facilitated harassment is addressed through a patchwork of laws including the Cybercrimes Act, Protection from Harassment Act, Domestic Violence Amendment Act, Films and Publications Amendment Act and PEPUDA. However stakeholders still describe implementation gaps such as under-trained police, slow and re-traumatising legal processes (one non-consensual-image case took seven years to reach judgment), and limited SAHRC subpoena power to identify anonymous "ghost accounts."

The Protection of Personal Information Act (POPIA) provides a complementary but partial privacy framework relevant to doxxing. The Films and Publications Amendment Act extends content classification to online material but has raised concerns about overbreadth and self-censorship. Electoral law (eg. the Electoral Commission Act and Electoral Act) was not designed for digital

political communication, leaving gaps around online disinformation and coordinated influence operations. Overall, the report concludes that South Africa's challenge is not a lack of legal protection, but weak implementation, fragmentation, and limited accountability for the platform systems that amplify harm.

On platform governance, the report argues platforms are not neutral hosts: algorithmic recommendation, advertising and monetisation systems actively shape what content is amplified. Using UNESCO's Guidelines for the Governance of Digital Platforms, the report assesses five dimensions:

1. Human rights due diligence (platforms' global policies do not adequately capture South Africa-specific risks);
2. Transparency (significant information asymmetry between platforms and users, researchers and regulators, with limited independent research access);
3. Local linguistic and contextual capacity (illustrated by the SAHRC's 2024 submission to the Meta Oversight Board arguing that the apartheid-era flag constitutes hate speech under domestic jurisprudence, which global moderation standards failed to register);
4. Accountability and effective remedy (opaque reporting and appeals processes);
5. Meaningful multistakeholder participation (which should complement, not replace, platform and state responsibility).

The report also cautions against treating all platforms uniformly, since public platforms, messaging services and recommendation-driven platforms carry different risks.

Evidence from stakeholder interviews

Stakeholder interviews add texture to each harm area. On GBV, interviewees described harassment and doxing of women in public life (especially journalists), NCII and sextortion (with a gendered pattern where boys are extorted for money, and girls for further images), AI-generated "nudify" abuse, trafficking and romance-scam exploitation, and normalised "manosphere" radicalisation of boys as young as ten. Disproportionately affected groups include female journalists, girls and young women, LGBTQIA+ people, and people with disabilities (particularly the deaf community, who rely on WhatsApp and are frequently targeted). Stakeholders repeatedly cited weak, under-trained policing (victims told simply to "block" perpetrators, or blamed for images originally shared consensually), judicial capacity gaps, platform accountability failures (including jurisdictional evasion by South African offices that function mainly as advertising arms), eurocentric moderation blind to isiZulu, Sesotho and other local-language abuse, and business models that reward "rage bait."

On xenophobia, stakeholders described a deliberately sustained "machinery" of narrative production rather than incidental content, algorithmic amplification (one academic implicated TikTok's geolocation-based recommendation system), failed pre-publication moderation (demonstrated by a media stakeholder's test advertisement that went unflagged), and South Africa's lack of jurisdiction over foreign-funded actors.

On democracy, stakeholders described a coordinated "machinery" of mis-, dis- and mal-information, including a tracked operation from actors operating outside of South Africa, who were paying local creators to spread "white genocide" narratives, and a "hybrid" economy paying ordinary people small sums (as low as ZAR 50) to amplify divisive content. Threats extend to journalists, whistleblowers and human rights defenders (some of whom have removed public contact details in response to harassment), and to religious and racial minorities (e.g., online attacks around Diwali, and Indian-African tensions in KwaZulu-Natal). Across all three harm areas, stakeholders cited outdated and fragmented legislation, insufficient police and regulatory capacity (including SAHRC's lack of subpoena power), jurisdictional limits over actors operating outside of South Africa, and platform accountability failures including reactive moderation, eurocentric standards, and monetised amplification of divisive content.

Access to remedies and media and information literacy

On access to remedies, the report finds South Africa's remedy landscape fragmented across platforms, regulators and public entities (such as the Film and Publications Board, SAHRC, Equality Courts), courts/police, and civil society, with each offering only partial and often inaccessible pathways. Platform reporting mechanisms are opaque; regulatory mandates overlap without a single coherent pathway; courts and police lack technical capacity and speed to match digital harm; and civil society, while active in monitoring and litigation, operates largely outside formal platform governance structures with limited data access.

Regarding media and information literacy (MIL), the report finds activity across government, schools, civil society, media and platforms, but no coordinated national strategy. Government-led work centres on election-specific tools like Real411 (including the Independent Electoral Commission, Media Monitoring Africa/Moxii Africa, and platforms). Schools lack a nationally mandated media literacy curriculum, with the Western Cape's Google-partnered pilot being the only province-level program. Civil society (led by Moxii Africa) and media organisations provide informal MIL functions.

However, needs are highly differentiated: young people need literacy addressing sextortion and AI-generated abuse; xenophobia-affected communities need tools to recognise manufactured statistics and paid amplification; voters need to recognise coordinated inauthentic behaviour; journalists need literacy paired with digital security training; and marginalised and disabled communities need materials designed in local languages and accessible formats from the outset. A notable recent development is the Competition Commission's November 2025 *Media and Digital Platforms Market Inquiry* final report. Using competition rather than content law, this now legally requires Google/YouTube, Meta, TikTok and Twitter/X to fund digital literacy programmes (including a ZAR 688 million Media Support Package and vernacular-language training via the *Media Development and Diversity Agency*), converting MIL from voluntary corporate responsibility into an enforceable remedy.

Key findings

The report's seven key findings are:

- 1 Online harms are rooted in existing social and political inequalities, but platforms alter their scale, speed and reach, so governance must coordinate with institutions addressing offline drivers;
- 2 Platforms actively shape amplification through design, so accountability must move beyond individual content to systems and processes;

- 3 South Africa has extensive but fragmented laws and institutions, requiring better coordination rather than an entirely new regulatory regime;
- 4 The gap between formal legal rights and effective, accessible remedy is a critical weakness;
- 5 Global platform governance does not consistently reflect South Africa's linguistic, cultural and political context;
- 6 Different forms of online harm require differentiated, risk-sensitive responses rather than one-size-fits-all regulation; and
- 7 MIL is essential but cannot substitute for institutional and platform accountability, and responsibility for systemic harms must not shift onto individual users.

Recommendations

The findings indicate that South Africa's central governance challenge is not simply the absence of legal or institutional mechanisms, but fragmentation between existing frameworks, uneven platform accountability and limited access to effective remedies. The recommendations therefore prioritise coordination across government, regulators, platforms and civil society, while recognising that different forms of online harm require differentiated responses. They seek to strengthen implementation and accountability within the existing governance architecture, while identifying areas where further regulatory intervention may be warranted.

Government

Recommendation 1: Establish a coordinated government response to online harms

Action: Government should strengthen existing coordination arrangements and establish a standing mechanism for coordinating policy and institutional responses to online harms. The existing Information, Communication Technologies and Media Regulators Forum of South Africa should be leveraged as a key coordination avenue, complemented by a dedicated interdepartmental mechanism where the Forum's existing mandate does not encompass the full range of institutions implicated in online harms.

Purpose: The mechanism should develop a shared understanding of different forms of online harm, clarify institutional mandates, identify regulatory and implementation gaps, coordinate responses to cross-cutting harms and establish clear referral and escalation pathways. Coordination should include not only regulators, but also relevant departments, law-enforcement and prosecutorial authorities.

Partners: Information Regulator, FPB, SAHRC, relevant departments, SAPS, National Prosecuting Authority and other relevant statutory bodies.

Output: A government-wide coordination framework setting out institutional responsibilities, referral pathways, escalation procedures and priority areas for joint action.

Government

Recommendation 2: Develop a coherent, rights-based governance framework

Action: The government should develop a coherent policy framework for harmful online content that better aligns existing legal and institutional approaches with constitutional rights and international human rights standards.

Purpose: The framework should clarify the distinction between unlawful content, harmful but lawful expression and constitutionally protected speech; establish proportionality as a guiding principle; and identify appropriate safeguards against over-removal and censorship. It should also account for the differentiated nature of online harms and the heightened risks faced by particular groups and communities.

Output: A common set of policy principles and institutional guidance for the prevention, regulation and remedy of online harms.

Recommendation 3: Invest in prevention and resilience without shifting responsibility to users

Action: The government should expand coordinated MIL, civic education and digital safety initiatives, including programmes responsive to groups disproportionately affected by online harms. These interventions should be developed as complementary measures rather than substitutes for platform and institutional accountability. Public education should equip users to identify, report and respond to harmful content while ensuring that responsibility for systemic harms remains with the institutions capable of addressing them. This should include integrating MIL into formal education and training programmes, based on a comprehensive and inclusive MIL strategy.

Output: Coordinated national and community-level prevention and resilience initiatives; comprehensive and inclusive strategy for youth, civil society, and public officials.

Regulators

Recommendation 4: Formalise regulatory cooperation and referral pathways

Action: Relevant regulators and enforcement institutions should establish formal cooperation protocols covering referrals, overlapping mandates, information sharing, investigations and escalation of serious or cross-cutting online harms.

Purpose: Formal cooperation should reduce duplication and jurisdictional gaps while ensuring that individuals are not required to navigate fragmented institutional structures to obtain assistance.

Output: Joint referral and escalation protocols, including designated institutional contact points and procedures for urgent cases.

Recommendation 5: Strengthen coordinated oversight and transparency of platform governance

Action: Regulators should develop coordinated approaches to overseeing platform governance, including transparency regarding content moderation, recommendation and amplification systems, enforcement practices and responses to complaints. Regulatory approaches should be proportionate and rights-respecting, distinguish between different categories of harm, and apply proportionate measures that protect freedom of expression while addressing unlawful and systemic harms. These should include the creation and oversight of mechanisms that enable multistakeholder participation and a more active role for academia, civil society and digital platforms.

Output: Coordinated regulatory guidance and, where appropriate, common transparency or reporting requirements.

Regulators

Recommendation 6: Strengthen access to effective remedies

Action: Regulators should improve the accessibility, visibility and coordination of complaint and redress mechanisms for individuals experiencing online harms, particularly TFGBV, NCII, hate speech, threats and coordinated harassment. Mechanisms should provide clear information on where to lodge complaints, how cases are referred between institutions, and what remedies are available.

Output: Clear and publicly accessible complaint, referral and escalation pathways.

Digital platforms

Recommendation 7: Strengthen transparency and accountability for platform systems

Action: Platforms should increase transparency regarding content moderation, recommendation algorithms, amplification and visibility-reduction measures affecting South African users. Transparency should extend beyond the number of content removals to provide meaningful information about how systems identify, prioritise and distribute potentially harmful content and how users can challenge enforcement decisions.

Output: Meaningful, South Africa-specific transparency reporting.

Recommendation 8: Strengthen local contextual capacity

Action: Platforms should invest in South African language capacity, local moderation expertise and contextual knowledge of the country's social and political environment. Engagement with South African civil society, researchers and affected communities should inform the development and evaluation of moderation systems, particularly where global policies may not adequately capture locally specific forms of harm.

Output: Improved local language and contextual moderation capacity, and structured stakeholder engagement mechanisms.

Recommendation 9: Apply human rights due diligence to platform design and enforcement

Action: Platforms should conduct ongoing human-rights due diligence assessing how platform design, recommendation and amplification systems may contribute to discrimination, polarisation or offline harm in South Africa. Moderation and enforcement should incorporate meaningful due process, including accessible appeals, clear explanations and effective remedies, with enhanced safeguards for particularly serious or high-risk harms.

Output: Context-specific risk assessments and strengthened appeals and remedy mechanisms.

Civil society and academia

Recommendation 10: Strengthen independent monitoring and evidence generation

Action: Civil society and academia should continue to generate independent evidence on emerging online harms, platform practices and their effects on democratic participation, with particular attention to groups and harms that remain under-researched. This evidence should support regulatory oversight, policy development and platform accountability.

Output: Regular independent monitoring, research and evidence products.

Recommendation 11: Institutionalise affected community participation

Action: Civil society, academia, regulators and platforms should establish meaningful mechanisms for affected communities to contribute to the development and evaluation of online-harm governance responses. Participation should extend beyond consultation to enable communities to identify emerging harms, assess whether interventions are effective, and inform platform and regulatory responses.

Output: Structured multistakeholder engagement mechanisms.

Recommendation 12: Expand MIL as a complementary resilience measure

Action: Civil society and academia should expand MIL and digital safety initiatives that enable users to critically assess information, recognise manipulation and understand reporting and remedy mechanisms. These initiatives should explicitly complement, rather than replace, institutional and platform accountability.

Output: Targeted MIL and digital safety programmes.

Conclusion

The report concludes that South Africa's digital environment reflects the country's broader post-apartheid social, political and economic inequalities. Digitalisation has expanded civic participation but also intensified existing tensions through the rapid spread of misinformation, hate speech and digitally facilitated violence. Current legal and regulatory frameworks, designed largely for analogue communication, remain ill-equipped to address the speed, scale, multilingual character, and contextual nuance of online harm, particularly where global platform moderation fails to account for local context. The report argues that addressing this requires not just additional content restriction, but a coordinated, rights-based governance ecosystem spanning government, regulators, platforms, civil society and communities, so that digital technologies strengthen rather than undermine democratic participation, social cohesion and human rights protection in South Africa.

Background

This study extends UNESCO's global Social Media 4 Peace initiative into South Africa, examining how digital platforms shape freedom of expression, hate speech and content governance in one of the world's most socially complex and politically significant digital landscapes. The research aims to build whole-of-society capacity for rights-based digital governance, deepen understanding of how harmful content affects marginalised communities, and produce practical recommendations adaptable to other African contexts, while contributing to South-South learning within UNESCO's broader digital governance agenda.

The study uses a mixed-methods approach combining secondary literature review, social media analysis, legal and policy analysis, and primary research through interviews with 29 stakeholders across government, regulators, the legal sector, civil society, media, academia, the technology and platform sector, and affected communities.



Disclaimer: This publication was produced with the financial support of the European Union and UNESCO. The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO or the European Union concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries. The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions expressed therein, which are not necessarily those of UNESCO or the European Union and do not commit the Organizations.