<AI & Equality> African Toolbox | Case study

Technology-Facilitated Gender-Based Violence in Africa: When AI Becomes a Weapon

Watch the video

This case study is part of the **African <AI & Equality> Toolbox**, which builds upon the methodology of the global <AI & Equality> Human Rights Toolbox—an initiative of Women At The Table in collaboration with the United Nations Office of the High Commissioner for Human Rights (OHCHR). The African Toolbox is a collaboration between the <AI & Equality> initiative and the African Centre for Technology Studies (ACTS).
To learn more visit **aiequalitytoolbox.com**

Introduction

In January 2025, Ethiopian Mayor Adanech Abiebie woke to find her face digitally grafted onto intimate videos with political leaders—deepfakes so convincing that 90% of viewers believed the fabricated narrative. Within hours, the AI-generated content linking her to Prime Minister Abiy Ahmed had garnered over 562,000 views, spreading the false claim that her political success stemmed from sexual relationships rather than competence. Meanwhile, in Cameroon, President Paul Biya's daughter Brenda faced a coordinated avalanche of harassment after publicly disclosing her sexual orientation—92 Facebook posts using identical templates reached 8.9 million people with before-and-after photos designed to mock her appearance and identity.

These aren't isolated incidents. They're part of a sophisticated, continent-wide campaign of Technology-Facilitated Gender-Based Violence (TFGBV) that weaponizes AI systems, exploits algorithmic amplification, and leverages cultural tensions to silence women and LGBTQ+ individuals across Africa. What Code for Africa's research reveals is both the staggering scale of these attacks—individual campaigns reaching millions—and their increasing sophistication as perpetrators learn to game AI systems designed to maximize engagement.

In Nigerian livestreams, young women are coerced into sexual acts through coordinated mass reporting threats. In Uganda, AI-powered content moderation systems fail to detect local language slurs like "woubi" and "lélé" that flood social media with anti-LGBTQ+ hatred. Across eleven African countries—Burundi, Cameroon, Côte d'Ivoire, Ethiopia, Ghana, Kenya, Nigeria, Uganda, South Africa, Senegal, and Zimbabwe—digital platforms have become battlegrounds where artificial intelligence amplifies rather than prevents systematic harassment targeting gender and sexual minorities.

The human cost is devastating: women political leaders withdrawing from public life, LGBTQ+ individuals silenced by fear, and democratic discourse degraded by campaigns that achieve massive reach through algorithmic promotion of controversial content. But this case study reveals something more troubling: current AI architectures, optimized for engagement rather than human dignity, create structural vulnerabilities that make such attacks not just possible but profitable for platforms and effective for perpetrators.

The Weaponization of Engagement: How AI Amplifies Hatred

The Ethiopian Mayor: When Deepfakes Target Democracy

The attack on Addis Ababa Mayor Adanech Abiebie began with a single TikTok account that had mastered the art of viral manipulation. On January 2, 2025, the account posted an AI-generated video showing Abiebie kissing Ethiopian Prime Minister Abiy Ahmed—a fabrication so seamless that it required technical analysis to identify as synthetic media. The video's caption suggested she had secured her mayoral position through sexual relations, tapping into deeply rooted biases about women in leadership.

What happened next reveals the terrifying efficiency of AI-driven harassment campaigns. Within the first 20 comments, 90% supported the video's false narrative, often responding with laughing emojis that signal high engagement to TikTok's algorithm. The platform's recommendation system, interpreting emotional reaction as user interest, began promoting the content to wider audiences. By November, a second deepfake video linking Abiebie to the Equatorial Guinea sex scandal had been created and distributed by the same account, demonstrating how successful harassment campaigns evolve and expand.

The technical sophistication was matched by cultural precision. The videos didn't just use AI to create convincing forgeries—they leveraged existing social attitudes about women in politics, transforming cutting-edge technology into a weapon for ancient prejudices. The mayor's actual governance record, including controversial urban development projects, became secondary to fabricated sexual narratives designed to undermine her authority through gendered attacks.

But the most chilling aspect wasn't the technology—it was how the platform's own AI systems became unwitting accomplices. TikTok's engagement-optimized algorithm treated the high emotional response as a signal to promote the content further, turning artificial intelligence into an amplification engine for artificial lies.

Brenda Biya: The Anatomy of Coordinated Digital Violence

When Cameroon's First Daughter Brenda Biya publicly came out as lesbian, she unknowingly triggered one of the most documented coordinated harassment campaigns in African digital history. The response wasn't spontaneous outrage—it was a precisely orchestrated attack that revealed the industrial scale of modern TFGBV operations.

Code for Africa's analysis of the campaign reads like a blueprint for digital violence. Ninety-two Facebook posts contrasted her "before and after" appearance, collectively reaching

8.9 million people and generating 17,745 interactions. But the devil was in the details: thirty-four of these posts used identical copy-paste techniques, featuring the same captions and layouts with surgical precision. This wasn't organic community response—it was coordinated inauthentic behavior designed to maximize algorithmic amplification.

The campaign's efficiency was staggering. The 34 identical posts alone generated 8.05 million views and 14,651 interactions, demonstrating how template-based attacks could achieve massive reach through minimal effort. Comments like "before she started sleeping with girls" reduced her changed style to sexual stereotypes, while others used her image to symbolize national decline: "She reflects the country's progress." A review of 4,600 comments found that 98% mocked or ridiculed Biya—a level of unanimity that suggested orchestrated rather than organic sentiment.

The cross-platform coordination was equally sophisticated. Between September 2024 and March 2025, approximately 50 TikTok videos—mostly posted by Ivorian users—continued the mockery as part of a "Cameroon vs Côte d'Ivoire" social media trend. Individual videos received hundreds of thousands of views, with coordinated timing patterns that maximized algorithmic visibility across platforms.

What made this campaign particularly devastating was how it exploited legitimate cultural discourse. The "country comparison" trend provided plausible cover for harassment, allowing attackers to frame systematic targeting as playful regional rivalry. This cultural camouflage made the content harder for automated systems to identify as harmful while ensuring it resonated with audiences predisposed to anti-LGBTQ+ sentiment.

The Nigerian Livestream Economy: AI-Enabled Sexual Exploitation

In Nigeria's TikTok ecosystem, Code for Africa documented something even more disturbing: the emergence of an AI-enabled sexual exploitation economy that uses platform features to coerce young women into performing sexual acts for online audiences. The case reveals how live streaming platforms become venues for real-time digital violence that combines technological coercion with economic manipulation.

The system operates with industrial efficiency. Hosts like @❤RICHARD DP and @SpecialPoint use phrases like "view once" to suggest content will only be visible temporarily, exploiting young women's concerns about permanent exposure. But viewers routinely record these sessions, preserving and redistributing content across platforms to maximize harm. One recording of a SpecialPoint livestream posted on X received 1.4 million views, transforming a moment of coercion into lasting digital violence.

The coercion mechanism reveals sophisticated understanding of platform vulnerabilities. When women set boundaries around what they're willing to do, hosts coordinate mass reporting campaigns to threaten account suspension—essentially weaponizing platform

safety mechanisms to enable abuse. In March 2025, researchers documented a host threatening to disable a young woman's account when she refused to expose herself, while viewers coordinated pressure tactics through coordinated messaging.

The AI dimension becomes clear in how these operations evade detection. Hosts use sequential username variations after suspensions—adding letters or numbers to return under slightly modified handles. The platforms' automated systems, designed to detect spam or commercial manipulation, consistently fail to identify these harassment networks that operate at the intersection of sexual exploitation and coordinated inauthentic behavior.

Perhaps most troubling is how platform algorithms reward this content. The high engagement generated by controversial livestreams—driven by a combination of sexual content and audience participation—signals to recommendation systems that this content should be promoted to broader audiences. The platforms' own AI systems become enablers of exploitation, transforming human trafficking into algorithmic success.

Cultural Warfare: Anti-LGBTQ+ Campaigns as Information Operations

Uganda's Legislative Hatred: When Laws Become Content

Uganda's Anti-Homosexuality Act, enacted on May 29, 2023, didn't just criminalize LGBTQ+ identities—it provided a legal foundation for coordinated digital harassment campaigns that achieved massive reach through AI-driven amplification. Code for Africa's analysis reveals how legislative hatred translates into viral content that spreads across borders and platforms.

Seven TikTok videos supporting the Act achieved a combined 868,030 views and 39,452 interactions, but their timing reveals strategic coordination. These posts appeared from March to April 2023—before the Act's enactment—indicating pre-existing anti-LGBTQ+ discourse designed to build support for criminalization. The content used phrases like "homosexuality is a sin," "sodomised," and "say no to homosexuality (LGBTQ)" that became viral hashtags amplified across X, TikTok, and Facebook.

The campaign's cross-border reach demonstrated how local legislation becomes regional propaganda. Ugandan content celebrating criminalization spread to Kenya following their Supreme Court's LGBTQ+ rights ruling, to Tanzania during parliamentary debates about LGBTQ+ support funding, and to Burundi where President Évariste Ndayishimiye suggested stoning homosexual people. Each national moment became an opportunity for coordinated amplification that transcended borders.

The technical sophistication was hidden beneath cultural authenticity. Videos featured local influencers, religious leaders, and politicians speaking in native languages about preserving African values against foreign corruption. This cultural resonance made the content highly engaging for target audiences while providing plausible cover for coordinated campaigns. TikTok's recommendation algorithm, unable to distinguish between genuine cultural expression and manufactured hatred, promoted the most engaging content to broader audiences.

Tanzania's Parliamentary Theater: Transforming Hatred into Headlines

Tanzania's parliamentary debate on May 17, 2024, reveals how TFGBV campaigns exploit democratic institutions to generate viral content. When MPs condemned ministry support for LGBTQ+ projects as threats to Tanzanian cultural values, their speeches became raw material for coordinated digital amplification that reached over a million people.

MP Mwita Waitara's declaration that "We do not want homosexuality in Tanzania. We do not want filthy behaviour here" became the centerpiece of a sophisticated content operation. Nine TikTok clips sharing his homophobic comments received 1,070,640 views and 67,877 interactions, while X posts supporting the MPs' statements reached 17,769 views through coordinated resharing.

The campaign's effectiveness stemmed from exploiting democratic legitimacy. Parliamentary speeches provided authoritative sources for anti-LGBTQ+ content that platforms couldn't dismiss as hate speech—after all, these were elected officials speaking in official forums. This institutional cover enabled massive amplification of harmful content under the guise of political reporting.

The algorithmic amplification patterns revealed how AI systems inadvertently promote institutional hatred. Parliamentary debates generate high engagement because they involve political conflict and controversial topics. Recommendation algorithms, optimized for user interest rather than social harm, promoted the most controversial clips to audiences likely to engage with anti-LGBTQ+ content. The result was democratic institutions becoming content factories for coordinated harassment campaigns.

Burundi's Presidential Violence: When Leaders Incite Digital Mobs

President Évariste Ndayishimiye's December 29, 2023 suggestion that homosexual people "should be put in a stadium and stoned" demonstrates how TFGBV campaigns exploit the highest levels of political authority. The statement generated 3,650 mentions on X, receiving approximately 25,500 engagements and 980,000 views, while 188 Facebook posts shared the president's comments between December 2023 and May 2024.

The viral amplification revealed sophisticated coordination mechanisms. Rather than simple resharing, the campaign involved strategic timing patterns that maximized algorithmic visibility. Posts appeared at optimal engagement times across different platforms, suggesting coordinated scheduling designed to maintain momentum over months rather than days.

The content moderation challenges became apparent when TikTok searches for related content triggered community guideline warnings, yet the material continued circulating through screenshot sharing and indirect references. This cat-and-mouse dynamic demonstrates how sophisticated harassment campaigns adapt to platform policies while maintaining their reach and impact.

The Technical Architecture of Digital Violence

Algorithmic Amplification: How AI Rewards Hatred

Code for Africa's research reveals a disturbing pattern: AI recommendation systems consistently amplify TFGBV content because emotional provocation generates the high engagement that algorithms interpret as user satisfaction. Analysis across platforms shows that controversial content targeting women and LGBTQ+ individuals achieves 15-20% higher engagement rates than baseline content, leading to algorithmic promotion that multiplies reach exponentially.

The Ethiopian mayor case provides a clear example of this dynamic. The initial deepfake video achieved 562,138 views not through paid promotion but through organic algorithmic amplification driven by high engagement rates. Users commenting with laughing emojis, sharing the content, and spending time viewing the fabricated material all sent positive signals to TikTok's recommendation system. The AI interpreted coordinated harassment as user interest, promoting the content to broader audiences who might not have encountered it otherwise.
This creates a feedback loop where harmful content becomes self-amplifying. Initial engagement drives algorithmic promotion, which increases reach, which generates more engagement, which triggers further promotion. The result is that well-executed harassment campaigns can achieve viral status without significant financial investment—they simply need to generate enough initial engagement to trigger algorithmic amplification.

The temporal patterns are equally concerning. Code for Africa found that 80% of TFGBV campaign engagement occurs within the first 48 hours, suggesting that algorithmic promotion decisions made in the crucial early period determine ultimate reach and impact. This creates a narrow window where intervention might be effective, but current content moderation systems consistently fail to respond quickly enough to prevent viral amplification.

Evasion Technologies: Gaming the System

The sophistication of TFGBV technical tactics reveals how perpetrators have developed systematic approaches to circumvent content moderation while maximizing algorithmic amplification. The research documents a comprehensive toolkit of evasion strategies that exploit specific vulnerabilities in AI-driven platforms.

"Spamouflage" techniques represent the most basic level of evasion. Attackers replace letters with symbols or numbers—writing "Us£less" instead of "Useless" or "w0n" instead of "won"—to bypass keyword-based detection systems. These modifications are subtle enough that human readers understand the meaning while automated systems fail to recognize harmful content.

More sophisticated is the exploitation of cultural and linguistic gaps in AI training data. Terms like "woubi" and "lélé"—French slurs targeting LGBTQ+ individuals—pass through content moderation systems trained primarily on English-language datasets. This cultural blindness creates systematic vulnerabilities that attackers exploit to spread harmful content in African contexts where local knowledge is essential for harm recognition.

Account management strategies reveal industrial-scale coordination. When harassment accounts face suspension, they return with slightly modified usernames—adding numbers or letters to maintain brand recognition while evading automated detection. Pre-registered backup accounts enable immediate resumption of activities, while cross-platform coordination ensures campaign persistence even when individual accounts face enforcement action.

The temporal coordination demonstrates sophisticated understanding of algorithmic systems. Coordinated campaigns time their posts for maximum algorithmic visibility, leverage trending topics to increase reach, and use engagement manipulation to trigger recommendation system promotion. This isn't amateur trolling—it's professional information warfare adapted for gender-based violence.

Content Moderation Failures: When AI Can't See Culture

The systematic failures of content moderation reveal fundamental limitations in how AI systems understand cultural context and coordinated behavior. Code for Africa's research documents specific cases where sophisticated harassment campaigns evaded detection despite clear coordination patterns.

The Brenda Biya case provides the starkest example. Thirty-four Facebook posts using identical copy-paste techniques should have triggered automated detection systems designed to identify coordinated inauthentic behavior. Yet these posts collectively achieved 8.05 million views while evading platform enforcement. The identical captions, synchronized

timing, and template-based sharing patterns represent textbook examples of coordination that current AI systems fail to detect.

The linguistic gaps are equally problematic. Content moderation systems trained primarily on Western datasets demonstrate reduced effectiveness with African cultural contexts and language patterns. Local slurs, cultural references, and context-dependent harmful content consistently pass through automated systems designed for different linguistic and cultural environments.

Real-time detection capabilities prove inadequate for the speed of viral content. TFGBV campaigns achieve massive reach before content moderation systems can respond effectively. The Ethiopian mayor's deepfake video reached over 500,000 people before any intervention, while livestream exploitation in Nigeria occurs in real-time with minimal possibility for protective intervention.

Perhaps most concerning is how algorithmic promotion outpaces human review. Content that violates platform policies still receives algorithmic amplification during the period between posting and moderation review. This creates a window where harmful content can achieve viral status even if it's eventually removed, making content moderation reactive rather than protective.

Human Rights in the Age of Algorithmic Violence

1 Stage 1: Objective and Team Composition
The Foundation of Harm

The human rights violations documented in TFGBV campaigns begin with fundamental design decisions made during AI system development. Platform objectives optimized for user engagement create structural incentives that reward controversial content regardless of social harm. Code for Africa's analysis demonstrates how these engagement-focused metrics systematically promote harassment campaigns targeting women and LGBTQ+ individuals.

The Ethiopian mayor case illustrates this dynamic clearly. TikTok's algorithm interpreted high emotional engagement with deepfake content as user satisfaction, promoting fabricated harassment material to broader audiences. The platform's objective function—maximize user engagement and time on platform—directly conflicted with human rights principles of dignity and non-discrimination. Yet the technical system had no mechanism for recognizing this conflict.

Team composition during AI development reveals systematic exclusion of affected communities and human rights expertise. Platform development teams lack meaningful representation from women, LGBTQ+ individuals, or African communities who bear the consequences of system design decisions. This exclusion isn't accidental—it reflects broader power structures that prioritize technical capability over social responsibility.

The absence of human rights considerations in this stage has cascading effects throughout the AI lifecycle. When systems are designed to maximize engagement without considering dignity, participation, or equality, they become vulnerable to exploitation by sophisticated harassment campaigns. The technical architecture embeds these values from inception, making later interventions inadequate for addressing fundamental structural problems.

Human Rights Alignment Requirements:

- **Community Agency in Objective Setting**: Affected communities must have genuine decision-making power in defining what AI systems should optimize for, not just feedback on predetermined technical goals.
- **Dignity-Centered Metrics**: Success measurements must include human dignity, democratic participation, and community safety alongside engagement and revenue metrics.
- **Representative Development Teams**: Meaningful inclusion of women, LGBTQ+ individuals, and African communities in technical decision-making roles.
- **Human Rights Expertise Integration**: Systematic inclusion of human rights practitioners in technical architecture and objective-setting processes.

2 Stage 2: Defining System Requirements
Building Safety into Technical Specifications

Current system requirements demonstrate fundamental inadequacy in addressing coordinated harassment campaigns targeting specific demographics. The Brenda Biya case reveals how 34 identical Facebook posts evaded automated detection systems designed primarily for spam or commercial manipulation rather than gender-based violence.

The technical requirements gaps extend beyond simple detection failures. Platforms lack demographic-specific harm monitoring, cultural context understanding, and rapid response capabilities for coordinated campaigns. The Nigerian livestream exploitation demonstrates how real-time TFGBV occurs faster than current moderation systems can respond, requiring fundamentally different technical architectures.

Cross-platform coordination represents another systematic requirement failure. TFGBV campaigns operate across TikTok, Facebook, X, and other platforms simultaneously, but current systems lack information-sharing capabilities to detect distributed harassment networks. Individual platforms optimize their own metrics while remaining blind to coordinated campaigns that span the digital ecosystem.

The absence of affected community input in requirements definition creates systems optimized for metrics that conflict with human rights. Engagement maximization, viral amplification, and recommendation system effectiveness become requirements without consideration of how these features enable systematic harassment of marginalized communities.

Human Rights-Aligned System Requirements:

- **Real-time Coordination Detection:** Technical capabilities to identify synchronized posting patterns, template sharing, and cross-platform campaign coordination.
- **Cultural Context Integration:** Content evaluation systems that understand local languages, cultural references, and context-dependent harmful content.
- **Demographic-Specific Harm Monitoring:** Systematic tracking of system impacts on women, LGBTQ+ individuals, and other marginalized communities.
- **Community-Defined Safety Standards:** Requirements development that includes affected community input on what constitutes harm and appropriate intervention.
- **Rapid Response Architecture:** Technical systems capable of intervention before viral amplification occurs rather than reactive content removal.

3 Stage 3: Data Discovery
Bias and Representation in Training Systems

Training data bias contributes systematically to TFGBV amplification through cultural blindness and representation gaps. Content moderation models trained primarily on Western datasets demonstrate significant effectiveness gaps when deployed in African contexts, failing to recognize local language slurs and culturally specific harmful content.

The linguistic bias is particularly severe. Terms like “woubi” and “lélé”—slurs targeting LGBTQ+ individuals in French-speaking African countries—pass through moderation systems that lack training data from these linguistic and cultural contexts. This isn’t simply a technical oversight—it reflects systematic underrepresentation of African voices in AI training data collection and curation.

Recommendation algorithm training demonstrates similar bias patterns. Models optimized on datasets that don’t include sophisticated harassment campaigns fail to recognize coordinated TFGBV tactics when deployed in African contexts. The algorithms learned to maximize engagement from data that didn’t capture the specific ways that marginalized communities face systematic digital violence.

The data collection process itself violates human rights principles by excluding affected community consent and participation. Training datasets include harassment content targeting women and LGBTQ+ individuals without their consent, while failing to include community knowledge about harmful content recognition and appropriate intervention strategies.

Human Rights-Aligned Data Practices:

- **Community Consent and Participation:** Affected communities must have agency in determining how their data is collected, used, and represented in training systems.
- **Cultural Representativeness:** Training data must include diverse African languages, cultural contexts, and community-defined examples of harmful content.
- **Participatory Dataset Curation:** Community experts should be involved in identifying harmful content patterns and appropriate intervention strategies.
- **Bias Impact Assessment:** Systematic evaluation of how training data representation affects different communities and intervention effectiveness.

4 Stage 4: Selecting and Developing Models
Technology in Service of Human Rights

Model selection and development decisions directly enable TFGBV through engagement optimization that rewards controversial content. Recommendation algorithms trained to maximize user engagement systematically promote harassment campaigns because emotional provocation generates the high interaction rates that models interpret as success.

The technical architecture embeds these harmful incentives throughout the system. Content that generates strong emotional responses—including coordinated harassment targeting women and LGBTQ+ individuals—receives algorithmic promotion regardless of social impact. Models optimized for engagement metrics lack mechanisms for recognizing when high interaction rates indicate harm rather than user satisfaction.

Explainability limitations prevent affected communities from understanding how algorithmic systems make decisions about content promotion and moderation. When harassment campaigns achieve viral reach through algorithmic amplification, victims and advocates have no insight into why these decisions occurred or how to challenge them effectively.

Fairness considerations remain absent from model development despite documented evidence that current systems systematically amplify harassment targeting specific demographics. The lack of intersectional fairness metrics means that platforms cannot identify when their systems disproportionately harm women, LGBTQ+ individuals, or other marginalized communities.

Human Rights-Aligned Model Development:

- **Community-Defined Success Metrics:** Models should optimize for community-identified values like safety, dignity, and democratic participation rather than purely engagement-focused metrics.
- **Harassment-Aware Architecture:** Technical systems must be designed to recognize when high engagement indicates coordinated harassment rather than organic user interest.
- **Transparent Decision-Making:** Affected communities must be able to understand how algorithmic systems make decisions about content promotion and moderation.
- **Intersectional Fairness Integration:** Models must include systematic evaluation of impacts on multiply marginalized communities and intersectional harm recognition.

5 Stage 5: Testing and Evaluation
Community-Centered Validation

Current testing frameworks demonstrate insufficient consideration of TFGBV scenarios and community-defined harm. Evaluation protocols focus on technical performance metrics rather than community safety outcomes, missing systematic ways that platforms enable harassment campaigns targeting marginalized groups.

The absence of affected community participation in testing creates systems optimized for metrics that conflict with human rights. Platforms measure success through engagement rates, user growth, and retention without systematic evaluation of impacts on women, LGBTQ+ individuals, and other vulnerable communities.

Real-world testing limitations mean that harassment scenarios receive inadequate evaluation during development. The sophisticated coordination tactics documented by Code for Africa—template sharing, cross-platform campaigns, cultural code-switching—represent attack patterns that current evaluation frameworks fail to anticipate or address.

Performance measurement systems lack demographic-specific assessment capabilities. Platforms cannot identify when their systems systematically amplify harassment targeting specific communities because they lack evaluation frameworks designed to detect these patterns.

Human Rights-Aligned Testing Approaches:

- **Community-Defined Harm Assessment:** Testing protocols must include affected community evaluation of what constitutes harmful system behavior and appropriate intervention.
- **Adversarial Harassment Scenario Testing:** Systematic evaluation against documented TFGBV tactics and coordination patterns.
- **Demographic-Specific Performance Monitoring:** Regular assessment of system impacts on different communities with particular attention to marginalized groups.
- **Real-World Impact Evaluation:** Testing that goes beyond technical metrics to assess effects on human dignity, democratic participation, and community safety.

6 Stage 6: Deployment & Post-Deployment Monitoring
Accountability and Continuous Improvement

Post-deployment monitoring reveals systematic gaps in platform capabilities to detect and respond to coordinated harassment campaigns. TFGBV operations achieve massive reach before intervention because current monitoring systems are reactive rather than proactive and lack real-time coordination detection capabilities.

The response capability limitations demonstrate how platforms prioritize technical performance over community protection. Average response times for content moderation exceed viral content spread times, meaning that harassment campaigns consistently achieve their objectives before any protective intervention occurs.

Community feedback integration remains inadequate despite sophisticated systems for collecting user reports and appeals. Affected communities report coordinated harassment campaigns that platforms fail to recognize as systematic threats rather than individual content violations.

Systematic learning from TFGBV incidents is limited by platforms' reluctance to acknowledge that their technical architectures enable harassment. Without recognition of structural problems, platforms focus on reactive content removal rather than proactive system design changes that could prevent future campaigns.

Human Rights-Aligned Monitoring and Response:

- **Proactive Threat Detection:** Real-time monitoring systems capable of identifying coordinated campaigns before they achieve viral amplification.
- **Community Agency in Intervention:** Affected communities must have mechanisms to rapidly escalate threats and influence platform response decisions.
- **Systematic Impact Assessment:** Regular evaluation of how platform systems affect human rights with particular attention to marginalized communities.
- **Structural Learning Integration:** Platform commitment to modifying technical architectures based on documented human rights impacts rather than limiting response to content removal.

Building Human Rights into AI Architecture

Technical Interventions That Center Dignity

The path forward requires fundamental architectural changes that embed human rights principles into AI system design rather than treating them as external constraints. Code for Africa's research provides a roadmap for technical interventions that could effectively mitigate TFGBV while maintaining platform functionality and innovation.

- **Engagement Quality Assessment Systems** represent the most critical intervention. Instead of optimizing purely for interaction quantity, platforms must develop technical capabilities to distinguish between positive engagement (learning, community building, democratic participation) and negative engagement (harassment, discrimination, coordinated attacks). This requires training models on community-defined examples of constructive versus harmful interaction patterns.

- **Coordination Detection Integration** must become a core platform capability rather than an afterthought. The Brenda Biya case demonstrates how 34 identical posts can evade detection despite clear coordination patterns. Platforms need real-time network analysis capabilities that can identify template sharing, synchronized timing, and cross-platform coordination before viral amplification occurs.

- **Cultural Context Recognition** requires systematic integration of African languages, cultural references, and local knowledge into content moderation systems. The failure to detect slurs like "woubi" and "lélé" isn't a minor oversight—it reflects systematic exclusion of African voices from AI development that must be corrected through participatory dataset development and community expert integration.

- **Rapid Response Architecture** must enable intervention before viral spread rather than reactive content removal. This requires predictive systems that can identify potential harassment campaigns in their early stages and protective measures that can be activated within minutes rather than hours or days.

Community Ownership and Platform Governance

Technical solutions alone cannot address TFGBV without corresponding changes in platform governance that give affected communities genuine agency in system design and operation. The documented harassment campaigns succeed partly because platforms operate as closed systems where community voices have minimal influence on technical decisions.

- **Community Advisory Integration** must go beyond tokenistic consultation to include affected communities in technical architecture decisions, policy development, and evaluation criteria. The Ethiopian mayor's experience with deepfake harassment could

have been prevented if platform design had included Ethiopian women's organizations in identifying potential harms and appropriate interventions.

- **Transparent Algorithmic Decision-Making** requires platforms to provide affected communities with meaningful information about how recommendation systems promote content and why specific moderation decisions occur. Currently, harassment victims have no insight into why coordinated campaigns achieve viral reach or how to effectively challenge algorithmic amplification of harmful content.

- **Community-Defined Success Metrics** must supplement or replace engagement-focused optimization with measurements that reflect human rights principles. Platform success should be evaluated based on community safety, democratic participation, and dignity rather than purely technical metrics that may conflict with human rights.

- **Cross-Platform Coordination** requires industry-wide cooperation to address harassment campaigns that span multiple platforms. Individual platform optimization creates systematic vulnerabilities that sophisticated campaigns exploit through distributed coordination.

Regulatory Frameworks and International Cooperation

The transnational nature of TFGBV campaigns documented across eleven African countries requires coordinated policy responses that can address cross-border digital violence while protecting legitimate communication and democratic participation.

- **TFGBV-Specific Legal Frameworks** must address the sophisticated coordination mechanisms that current laws don't adequately cover. The harassment campaigns targeting the Ethiopian mayor and Brenda Biya represent forms of coordinated digital violence that require legal recognition and enforcement mechanisms designed for AI-enabled coordination.

- **Platform Accountability Standards** must include specific requirements for TFGBV prevention rather than generic content moderation obligations. Platforms should be legally required to maintain systems capable of detecting coordinated harassment campaigns and providing rapid protective intervention for targeted individuals.

- **International Cooperation Mechanisms** are essential for addressing campaigns that exploit platform coordination across different jurisdictions. The viral anti-LGBTQ+ content spreading from Uganda to Tanzania to Kenya demonstrates how local legislation becomes regional propaganda that requires coordinated response capabilities.

- **Community Participation Requirements** must be embedded in regulatory frameworks to ensure that affected communities have genuine agency in defining harm and appropriate intervention rather than having technical solutions imposed by external authorities.

Conclusion:
Reclaiming AI for Human Dignity

The documented patterns of Technology-Facilitated Gender-Based Violence across Africa reveal both the devastating human cost of AI systems optimized for engagement over dignity and the potential for technical architectures that serve human rights instead of undermining them. The Ethiopian mayor whose fabricated sexual scandals reached over half a million people, Brenda Biya whose harassment campaign generated 8.9 million views, and the countless women coerced in Nigerian livestreams represent not isolated tragedies but systematic failures of AI systems designed without meaningful consideration of human rights principles.

Yet their experiences also illuminate the path forward. Every documented harassment campaign reveals specific technical vulnerabilities that can be addressed through AI architectures designed to center community safety over engagement maximization. Every coordination pattern that current systems fail to detect provides blueprints for more effective intervention mechanisms. Every cultural blindness in content moderation identifies opportunities for more inclusive AI development that includes African voices in technical decision-making.

The choice facing the AI development community is stark: continue building systems that systematically amplify digital violence against marginalized communities, or fundamentally restructure technical architectures to embed human rights principles throughout the development lifecycle. The research demonstrates that sophisticated harassment campaigns will continue exploiting engagement-optimized algorithms until platforms prioritize dignity over viral growth.

But this case study also reveals reasons for hope. The technical interventions required to address TFGBV—coordination detection, cultural context recognition, community participation mechanisms—represent advances that would benefit all platform users, not just those targeted by harassment campaigns. Building AI systems that protect the most vulnerable creates more robust, democratic, and sustainable digital environments for everyone.

The women political leaders, LGBTQ+ individuals, and marginalized communities targeted by these campaigns are not asking for special protection—they're demanding equal access to digital spaces free from systematic harassment that undermines their fundamental human rights. Their calls for justice provide blueprints for AI development that serves human flourishing rather than exploitation.

The deepfakes targeting the Ethiopian mayor continue circulating, but her experience has contributed to growing recognition that current AI architectures are fundamentally incompatible with human rights principles. The coordinated harassment of Brenda Biya reaches millions, but the documented coordination patterns provide technical specifications for detection systems that could prevent future campaigns. The exploitation documented in Nigerian livestreams continues, but the evidence of systematic coordination offers pathways for protective intervention.

Their experiences, documented through Code for Africa's meticulous research, transform individual trauma into collective knowledge that can reshape how AI systems relate to human dignity. The women who faced these attacks have become inadvertent experts in the vulnerabilities of engagement-optimized algorithms and the possibilities for technical architectures that center community safety.

The next phase of AI development will be defined by whether the technical community learns from their experiences or continues building systems that amplify the very forms of digital violence these women have endured. The choice is between AI that serves engagement metrics regardless of human cost and AI that treats human dignity as the ultimate optimization target.

In the end, the women whose harassment campaigns are documented in this research are not just victims of algorithmic violence—they are unwitting pioneers of a more democratic approach to AI development that centers community needs over technical convenience. Their suffering demands nothing less than fundamental transformation of how artificial intelligence relates to human rights.

The technology exists to build these better systems. The legal frameworks can be developed to ensure accountability. The community knowledge is available to guide more inclusive development processes. What remains is the political will to prioritize human dignity over engagement maximization and community safety over viral growth.

The women of Africa who have faced these attacks are still speaking, still leading, still demanding digital spaces that honor their humanity. Their voices, amplified not by engagement-hungry algorithms but by principled solidarity, point toward AI futures that serve human flourishing rather than exploitation. The question is whether the technical community will listen.

About the case study

This research uses behavioural and narrative analysis to examine technology-facilitated gender-based violence across 11 African countries, drawing on social media data to identify disinformation and coordinated harassment patterns. Code for Africa (CfA), the continent's largest civic technology and data journalism initiative, supported the research through its expertise in open-source intelligence and data-driven investigations.

This report was compiled by Code for Africa's Hanna Teshager, a senior investigative data analyst at the iLAB team, using ML to combat disinformation, map online coordinated inauthentic behaviour, and influence operations. The report is based on her research and analysis with investigative data analysts, including senior investigative data analyst John Ndung'u, Chike Odita, Fatimaelzahra Saeed, Moffin Njoroge, and Vanessa Manessong. The research ongoing since 2023 examines TFGBV patterns across 11 African countries. About the authors Hanna Teshager is CfA's senior Investigative Data Analyst at the iLAB team, with 4+ years of experience using ML to combat disinformation, map online coordinated inauthentic behaviour, and influence operations.

Other contributors to this case study are Caitlin Kraft-Buchman, Emma Kallina, and Sofia Kypraiou, authors of the original *Framework to AI Development: Integrating Human Rights Considerations Along the AI Lifecycle* upon which the Toolbox structure is based. Additional contributors are Amina Soulimani and Pilar Grant, from Women at the Table and the <AI & Equality> Human Rights Initiative.