Engagement at What Cost?

Examining the intersection of social media, Generative AI and gender-based violence

Scott Timcke and Zara Schroeder

Working Paper

August 2024

RESEARCH ICT AFRICA

Workshop17
17 Dock Road
V&A Waterfront
8001
Phone: +27 21 447 6332
www.researchictafrica.net

About the Resisting Information Disorder in the Global South project

Principal Investigator: Herman Wasserman, Stellenbosch University

Despite information disorder being a widespread problem in countries in the Global South, the study of this phenomenon remains dominated by examples, case studies, and models from the Global North. This project builds on this pre-existing research and its recommendations. It takes a thematic approach to identify the key drivers of information disorder in the Global South and evaluates appropriate responses and strategies. The goal is to support and influence future policy and governance interventions.

About Research ICT Africa

Research ICT Africa (RIA) is an African think tank that has operated for over a decade, working to fill a strategic gap in the development of a sustainable information society and digital economy. It has done so by building the multidisciplinary research capacity needed to inform evidence-based policy and effective regulation in Africa. RIA's dynamic and evolving research agenda examines the uneven distribution of the benefits and harms of the intensifying global processes of digitalisation and datafication.

Acknowledgements

We wish to thank our participants who generously shared their personal perceptions and experiences of online gender-based violence. We appreciate the time they dedicated to participating in our interviews.

We thank our colleagues at RIA for their contributions to this report, as well as Drew Haller and Alan Finlay for their editorial support.

Introduction

Michelle, a participant in our interviews, recounted: "Spaces where I have encountered GBV [gender-based violence] have mostly been online." Her words carry the weight of first-hand witness: "It would mostly be on podcasts, and I would find snippets of those podcasts on TikTok, where men are talking about women in a degrading way, making rape jokes, and just overall being really derogatory."[1] As we argue in this report, the rapid spread of this content is facilitated by the current design of platform engagement markets. Michelle's observations underscore the gender hierarchies in such markets, where misogynistic sentiments are not only tolerated but can also be actively encouraged by misogynistic users. "It is in a space where there are other men, and the other men are riling one another up", she noted, painting a vivid picture of the toxic dynamics at play. Generative Artificial Intelligence (GenAI), while offering many affordances, is challenging policymakers to re-examine the gender dynamics of information disorders within platform and AI governance.

Building upon previous Research ICT Africa (RIA) studies examining the systemic and commercial drivers of misinformation, this paper scrutinises the design of engagement markets that monetise and propagate information disorders across Africa (Timcke, Orembo, & Hlomani, 2023; Timcke et al 2023; Timcke & Hlomani 2023; Hlomani et al 2023).

Engagement markets play a central role in enabling the spread of coordinated harm, often through opaque incentive structures that prioritise viral dissemination regardless of veracity. We define an engagement market as a platform where many different sets of data users meet. The platform helps create value by allowing these groups to share, analyse, and use data together. These markets incentivise certain types of content by rewarding high engagement metrics like views, likes and shares. As a result, sensational, controversial, or emotionally charged content often gains more traction, even if it is misleading or harmful. This creates a feedback loop where creators and platforms are motivated to produce and amplify content that generates the most engagement, rather than focusing on accuracy or social benefit (see Timcke & Rens 2024).

In mentioning these matters, we do not wish to suggest an instrumental conception in which GenAI is changing the world; rather ours is a socio-technical conception that examines interests and benefits from the current deployment of GenAI. As we demonstrate throughout this working paper, issues of platform governance and the rollout of AI systems cannot be separated from social life.

For this paper, we seek to better understand how opaque engagement markets facilitate the creation of online gender-based violence (GBV). We follow the UN Declaration on the Elimination of Violence against Women (1993) in understanding that "'violence against women' means any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life."[2]

[1] See Table one for a list of participants and their demographics.

[2] This paper uses the term 'woman' broadly, though gender is fluid. We recognise the heightened vulnerability to GBV faced by transgender, queer, and gender non-conforming individuals within the LGBTQ community. We aim to be inclusive across the gender spectrum and highlight the need for further research on the unique challenges faced by gender-diverse populations.

Opacity is an important characteristic because it is one means by which hierarchical power hides the violence of domination. GBV can be understood as a manifestation of patriarchal systems of oppression that have subjugated women for centuries. It is not an isolated issue, but rather a symptom of deeply rooted gender inequalities ingrained in social, economic and political structures. While AI and its associated engagement markets become mediators of GBV, one must recognise that the structures and relations enabling GBV predate the internet era (Faith 2022; Baekgaard 2024). In South Africa, for example, the staggering statistics on femicide lay bare the harsh reality that mere awareness is insufficient. Michelle's exasperated words, "We know the statistics, we know femicide is a problem. I don't know what else we can do in terms of prevention", expose the urgent need for transformative action. Preventative measures must dismantle patriarchal norms and challenge systemic gender biases that enable and perpetuate violence against women across all spheres — physical and virtual.

First, we examine opaque AI-mediated engagement markets and research on the implications of AI-saturated social media systems. We then analyse the proliferation of non-consensual intimate images as an expression of patriarchy and the production of fear. Our empirical findings present interviews with South African women focusing on their experiences of platform-based sexism and the individual self-protection strategies that they use to mitigate this. However, we argue that individual actions are insufficient responses, partly because of how engagement markets are designed.

Socio-technical Entanglements and Opaque Engagement Markets

Too often, the discourse around technological innovation, automation, and AI presents these advancements as inevitable, placing primary focus on managing the subsequent social costs on citizens (See Timcke et al 2024). However, it is crucial to examine whether some of these 'adjustment costs' are simply too high or not worth paying. Indeed, public policy cannot be solely concerned with mitigating the negative consequences of technological change; it must also address the fundamental question of whether certain innovations or their applications are truly desirable.

The rise of information disorder campaigns, facilitated by engagement markets and platform dynamics, exemplifies the systemic challenges posed by new technologies when left unchecked or unregulated. These campaigns not only undermine the credibility of essential institutions like electoral bodies, but also threaten the integrity of democratic processes and the ability of groups to participate in political decision-making. At the same time, current political impasses have less to do with epistemology and more to do with economy. While there will always be competing interpretations of the world, its events, and their significance, what matters is strengthening the sense of accountability to one another through collective enterprise, and a shared sense of the stakes involved and investment in the success of politics.

Addressing the systemic challenge of information disorders requires an approach that goes beyond fact-checking or content moderation to examine wider socio-technical entanglements. This necessitates a deeper examination of the economic incentives, power dynamics, and regulatory frameworks that shape the digital ecosystem and enable the flourishing of engagement markets that profit from the spread of information disorders.

Engagement markets can pose challenges for platform governance, in part because private platforms have a proprietary interest in not revealing systems and code (Lindquist, & Weltevrede, 2024; Burrell, 2016). This makes rigorous external, independent third-party data collection and evaluation difficult, conditional, or near non-existent. Overall, the opaque nature of platforms can undermine effective governance. When optimised for engagement, incentives are created for harmful behaviour which in turn can compromise social cohesion.

Opacity can manifest in several ways within GenAI systems. It can range from the proprietary and closed-source nature of the code to corporate decision-making procedures. The issue also extends beyond technical obscurity. AI models are often riddled with biases along the lines of class, race, and gender, reflecting existing societal prejudices. Moreover, even the researchers developing these models frequently lack a comprehensive understanding of the intricate decision-making processes underlying their outputs. This epistemic opacity is particularly concerning in the context of AI content moderation systems, which wield considerable influence over the flow of information and discourse in online spaces. This lack of transparency raises profound ethical and epistemological questions about the role of GenAI in shaping information orders.

Distinctions and linkages between social media and GenAI

Platforms distort perceptions and norms by rendering moderate opinions practically invisible, often over-amplifying the most extreme and polarising voices (Robertson, del Rosario, & Van Bavel 2024). GenAI meanwhile threatens to have a homogenising effect by reducing the richness of human experiences and perspectives to a statistical mean. Functioning together, the tendencies of platforms and GenAI result in social media algorithms that inadvertently create echo chambers that cause fringe viewpoints to appear more widespread than they are, giving them as much standing and credibility as more typical perspectives. This distorted representation skews people's judgments about what beliefs and behaviours are prevalent.

By contrast, AI systems trained on datasets meant to capture the central tendencies in and of human behavioural data, creative works, and knowledge run the risk of flattening diversity into reductive types. By their design, machine learning models learn the statistical norms, minimising deviations as 'noise'.

The tensions between these modes of technological mediation will shape online publics in the next decade. Additionally, 'traditional' oversight methods like ensuring transparency become challenging when it comes to GenAI. This is because GenAI systems largely function as opaque 'black boxes' where the internal decision-making processes are obscured and difficult to interpret. The complex way GenAI operates complicates the ability to scrutinise and understand how these technologies arrive at their outputs.

The implications of an AI-saturated existence

Much like manufacturing power restructured the labour force at the onset of industrialisation, AI will reshape social, political and economic life in unexpected ways (Timcke 2021). People and policymakers are only starting to fully grasp the implications of an AI-saturated existence. Concurrently, an obsession with AI's 'existential threat' overlooks the current, real harms of these systems (Hanna & Bender, 2023). These systems are already exacerbating social inequalities. As Amnesty International explains:

> Numerous states across the globe have deployed unregulated AI systems to assess welfare claims, monitor public spaces, or determine someone's likelihood of committing a crime. These technologies are often branded as 'technical fixes' for structural issues such as poverty, sexism and discrimination (Amnesty International 2023).

Addressing the current and potential harms of social inequality require sustained efforts to identify the human rights violations that will occur if these systems are deployed without proper regulation, oversight and accountability mechanisms. It also requires addressing matters of market concentration so that the business interests do not trump public interest.

Persuasive GenAI propaganda

There is evidence from experiments that Language Learning Models (LLMs) like GPT-3 can generate persuasive propaganda in the form of news articles, particularly when a human curated the outputs (Goldstein et al 2024). Under certain conditions, the edited GenAI propaganda was found to be as persuasive as original human-written propaganda (Goldstein et al 2024). Additionally it appears that GenAI propaganda is hard to detect, regardless of whether human or computational detectors are used (Chen & Shu 2023).

When these factors are bundled together with existing AI personalisation and distribution algorithms (and the knowledge that micro-targeted ads can be more effective than non-tailored ads [Simchon, Edwards, & Lewandowsky, 2024]) there are real prospects for GenAI to create effective political ads tailored to an individual or group at scale and speed.

Gender-based Violence and Female Fear Factories

The proliferation of non-consensual intimate image distribution is facilitated by tools that are accessible, inexpensive and easily developed due to open-source code. There is a history of exploitation of women's bodies (Mies, 1986; Hanmer, 1990l Carlin, & Federici, 2014; Goldin, 2021), which has been especially visible in the realms of human trafficking where women are often used as sex workers (Alvarez, & Alessi 2012). GenAI provides a more accessible means for the broader public to manipulate and repurpose images. Once these manipulated images are uploaded to various platforms, they move beyond the original uploader's control, spreading across the internet and reaching vast audiences. This capability of GenAI to create new content from existing images poses significant concerns regarding the misuse and exploitation of women's bodies and images.

As Karen Hao (2021) reports, industry figures estimate that women are the subjects of between 90% and 95% of non-consensual intimate image distribution. This form of abuse leverages a person's biometric data, using readily available tools that can generate strikingly realistic content from minimal inputs. The data is often sourced from various platforms where user-generated content is uploaded. The low barrier to entry, exemplified by a Telegram user openly advertising the ability to produce fake nude images for USD 10 using just photos from social media, has enabled a market for visual sexual exploitation at an unprecedented technological scale (Maiberg, 2024). The proliferation of services offering to create GenAI pornographic imagery of any individual without their consent represents a severe and unacceptable violation of privacy and digital safety.

Content mills are stealing videos and images from real women on Instagram and using GenAI to superimpose synthetic 'influencer' faces onto the stolen content. These mills then monetise this

misappropriated and non-consensual material by using AI influencer bot accounts to drive paid subscriptions on platforms like OnlyFans (Koebler, 2024). This represents a pernicious convergence of content theft, lack of consent over use of individuals' likeness and bodies, and the leveraging of emerging AI capabilities to commit gender-based harms at scale for profit. Instagram's lack of systematic action to curtail this abuse of their platform amplifies the injustice against the predominantly women victims whose privacy and autonomy are being disregarded.

When considering the AI landscape in which online GBV operates, the current 'My Pussy in Bio' spam adequately depicts it (Herrman, 2024). The 'My Pussy in Bio' spam has been trending on X since January 2024, and X users have reported that it has been difficult to avoid. This spam demonstrates the challenging landscape women need to navigate online. Furthermore, it displays how women's digital content is stolen by people online and used for pornographic content created with generative AI without their consent.

One participant in our interviews, Grace, provided a similar example: "There are Instagram accounts, known as 'leak accounts' that purposefully post pornographic/inappropriate pictures of people without their consent, some purely dedicated for revenge. It's very difficult to get pictures removed from this site, and even if Instagram takes it down, they generally start a new account and repost all the pictures again."

Mindful of how "patriarchy relies heavily on the symbolic, on the arrangement of certain events as though their internal logic is natural, automatic and inevitable" (Gqola 2021, p. 13), these malicious attacks, which primarily target women, are fundamentally about exerting power. Victims of synthetic revenge porn, whether it involves still images, videos or audiovisual content, face limited legal recourse due to the "long-standing indifference that our law has toward crimes that disproportionately affect women and girls" (Moreau & Rourke, 2024). The circulation of this content harms the fundamental rights of women, individually and as a class.

Pumla Dineo Gqola (2021) has written convincingly about 'Female Fear Factories'. What Gqola refers to are the mechanisms that produce the disturbing reality of violence against women within patriarchal societies. "It is a theatrical and public performance of patriarchal policing and violence towards women and others cast as female, who are, therefore, safe to violate," Gqola explains (2021, p. 18-19). These mechanisms, structures, and relations actively manufacture, reproduce, normalise and perpetuate the subordination of women through fear. One expression of this fear is 'rape culture', which Gqola explains is part of the "routine terrorisation of women" (2021, p. 11).

Audiences are a central component of rape culture. Judgements about respectability and shame are often shared publicly to remind women that their bodies are objects of use. Another participant in our interviews, Nancy, concurred with this when she stated, "women get threats of rape on their posts and are sexually harassed through their comments". This fear acts as a tool for women to police themselves, restricting their freedom and keeping them under patriarchal control. The legacy of Southern Africa's history shapes the experience of GBV. As highlighted in the introduction, GBV is a pervasive issue in South Africa, one that is inextricably linked to various converging and diverging drivers of power. Despite these bleak dynamics, Gqola emphasises the resistance and agency of women to curtail patriarchy, aiming to dismantle the very structures that perpetuate this violence and empower women to claim their agency.

Methods and Ethics

To better understand the 'routine terrorisation of women' we asked a small purposeful sample of nine women in South Africa to share their perceptions of online GBV. The primary objectives were to explore their perceptions of the forms, contributing factors, and potential solutions to this issue. Specifically, we sought qualitative data through personal accounts, experiences, observations, and opinions regarding the kinds of manifestations of online GBV they had encountered; their views on the role and responsibilities of platforms in addressing this challenge; and the influence of norms on the manifestation of online GBV.

Participants were recruited through an open call circulated via WhatsApp and subsequent snowball sampling between March and June 2024. We employed asynchronous platform-mediated text conversations via WhatsApp to explore participants' understandings of online GBV. Given the dispersed geographic locations of our participants across South Africa, WhatsApp provided a convenient and appropriate platform for data collection. As other researchers have found, the platform offers "ease of use [and] convenience" (Gibson 2022). This method enables researchers to "study human interactions such as patterns of behaviours, experiences, and perceptions presented in the online world" (Salmons 2009) in a cost-effective manner for both researchers and participants, especially as the latter group do not have to take time off from work to contribute to studies like this one (De Gruchy et al 2021 ; Shahid, & Shaikh, 2019).

While employing WhatsApp for data collection offers several advantages, there are limitations that need to be considered, especially around privacy concerns, data security and access requirements. Having participants share their perceptions via WhatsApp raised privacy concerns regarding the ability of others, especially service providers, to access information on their mobile phones. The security measures taken by participants regarding their personal information could compromise the confidentiality of conversations. Using WhatsApp as an interviewing tool also requires participants and researchers to have access to mobile phones and reliable internet connectivity, which may be a barrier for lower socioeconomic groups (Mwanda, 2022).

Lastly, asynchronous text conversations creates the possibility that participants researched their views prior to responding. This could affect the research findings. In all instances we have included interviewee responses verbatim, despite grammatical errors in some responses. Participants' confidentiality was strictly maintained, and data was securely stored and handled in accordance with relevant research guidelines. The participants' names used in this paper are pseudonyms. The participants represented a diverse range of individuals, including unemployed persons and working professionals from various racial and class backgrounds. See Table 1 for a list of participants and their characteristics.

Participant	Age	Gender	Race	Daily online usage	No. of social media platform/app accounts	Name of social media platform/app used	Years spent online
Angela	Not Shared	Not Shared	White	4 hours	5	Instagram, Pinterest, Facebook, Reddit, Snapchat	15
Fiona	26	Female	Black	6-7 hours	6	Instagram, TikTok, Facebook, LinkedIn, Snapchat, WhatsApp	13
Freya	30	Female	Indian	2 hours	3	WhatsApp, Instagram, TikTok	9
Grace	20	Female	Coloured	3 hours	3	Instagram, TikTok, Snapchat	7
Michelle	27	Female	Coloured	4 hours	4	Instagram, Facebook, YouTube, TikTok	14
Nancy	27	Female	Coloured	3-5 hours	5	Instagram, TikTok, WhatsApp, Pinterest, Twitter	17
Naomi	29	Female	Coloured	8 hours	4	Instagram, LinkedIn, TikTok, Facebook	13
Rebecca	27	Female	White	2 hours	4	Instagram (two accounts), TikTok, Snapchat	17
Sabrina	23	Female	Indian	3-5 hours	4+	Multiple Instagram and TikTok accounts for public and private posting, WhatsApp, Locket	12

Table 1: Demographics of the participants

Experiences and Perceptions of Online Misogyny

This section examines participants' encounters with online misogyny and GBV, exploring their understanding of these issues and their perceived connection to power dynamics. It covers concerns about GenAI's impact on online GBV and recounts witnessed instances of such violence, along with their implications. The discussion covers the targeted objectification of marginalised groups and social media's role in proliferating online GBV. It also addresses how the male gaze reinforces unrealistic standards and gendered inequalities. The section then explores participants' self-protective measures, their limitations, and the influence of social, cultural, and religious contexts on online safety strategies. Finally, it examines participants' views on blame, desensitisation, and the impact of social position on experiences of online GBV.

The Understanding of online GBV

Most of the participants identified the non-consensual sharing of sexual images as a form of online GBV. The participants also made linkages between GBV and power. Using phrases like "subtle discrimination" and "outright aggression", Fiona stressed how cyberbullying was a common tactic of online GBV, while Sabrina stressed how "harassment [...] is done using social media"." Nancy understood online GBV as "harassment through comments" on platforms, but also as "things like identity theft, impersonation, hacking, stalking, revenge porn/sharing of private messages, picture, videos with consent take place."

Participants explained how platforms provided access "for perpetrators to engage in abuse", as well as reach to disseminate abusive content. "Social media provides a platform for sharing intimate pictures without consent. Perpetrators may use distribution of such photos to blackmail and harass victims," as Freya said. Rebecca was also of the opinion that "some of the comments on social media bully, shame, or threaten individuals"." This perception was shared by Fiona:

> These forms of violence permeate various online platforms like Tiktok, X, Instagram, and Facebook, where anonymity and limited moderation often prevail. The prevalence of gender-based harassment in these spaces creates a hostile environment that undermines users' sense of safety and well-being, impacting their mental health and online participation.

These accounts outline how social media enables power dynamics that play out in GBV by providing a platform for it to take place. Furthermore, the widespread use of social media means that the reach of GBV is amplified.

Understanding of GenAI

Grace explained that she is "cautious of generative AI being introduced into social media", and that "at times it feels like an invasion of privacy". She elaborated by stating that "there's a risk of misuse, especially in terms of creating fake images and deep fakes that can harm people's reputations and privacy". "The thought that someone could take my ordinary photos and manipulate them into something inappropriate without my consent is really distressing and triggering. I worry for myself because of past harassment experiences," she added. Grace's perspective suggests that her privacy concerns are primarily focused on the potential misuse of AI , particularly in creating fake images and deepfakes that could harm people's reputations and privacy.

For Naomi “generative AI can be dangerous”. Nancy echoed these sentiments: “I don’t think generative AI is something that can be very safe.” Angela shared how GenAI’s presence in social media “trigger[s] intense feelings of existential dread, depression, and self-questioning”. “[I’m] not a fan of it [generative AI],” said Sabrina.

In contrast to these participants’ reactions to GenAI, Rebecca confessed that she is “not very familiar with the topic” since she has not “done much research on it”.” However, she thought that GenAI is “something that has been introduced into our lives without our consent”.

The notion of consent provides a nexus to the issue of online GBV since victims’ privacy is violated through the widespread use of this tool. As Rebecca suggests, even though she says that she is “not familiar” with GenAI, she is aware of the threat of personal information being used unknowingly, and the potential harms that stem from this:

> We don’t even know how our information has been used. And I think it speaks a lot to how harmful this can be because we don’t really know much about it or the general public doesn’t know much about it, yet it is a part of our life.

Her statement about the lack of consent in the introduction of GenAI also touches on a wider phenomenon of the often-involuntary nature of technological adoption, and how the participants view the rollout of products in the market. These responses suggest a sense of distress, fear and concern around GenAI.

Witnessing GBV online

Participants shared their first-hand experiences of online GBV. These “very common” experiences have negatively affected their perceptions of social media. Freya detailed one experience: “Someone I did not know messaged me to be a sugar baby offering money so that I could talk to them.” Similarly, Sabrina said:

> I’ve witnessed online GBV, through unwanted sexual remarks from men on Instagram and TikTok, on Omegle when men flash their private parts, and I’ve seen gender discriminatory memes on Instagram. Often girls will post themselves on social media and men will assume that they want sexual attention, which would lead to them saying unsolicited sexual remarks and sending unsolicited nudes.

Grace shared the following:

> I’ve experienced harassment on Instagram before, both from people I knew and strangers, and the thought of genAI enabling something like this is really scary. I’ve also found a lot of males expect you to be open to conversation and willing to do certain things, even if you do not respond to them. I’ve even experienced many fake accounts trying to follow me after I’ve blocked certain people (men) from my social media, and although Instagram has put in place measures against this, it doesn’t really work.

Grace also provided an account of an instance where two of her friends’ “ordinary images” were manipulated and transformed into pornographic content without their consent. She spoke about the impact this had on their mental health:

> This experience was particularly traumatic for them, given my one friend’s history as a survivor of sexual assault. The violation of their privacy and the unauthorised use of their likeness in such a manner triggered painful memories and exacerbated their past traumas,

> she also suffers constant PTSD at the thought and does not like being photographed anymore. The incident not only caused emotional distress but also led to a prolonged struggle to have the manipulated images removed (which was not successful) and to address the damage to the victims' reputations.

This account illustrates the lasting impact of online GBV on victims, including mental strain and damage to their reputations. Misogynistic comments are directed "at women, trans, or non-binary queer people specifically," according to Rebecca. "They tend to get some comments on their appearance, comments on their identities, comments/threats to them [which are] shaming."

These experiences illustrate the various lasting implications of online GBV, and the distress caused by online violations. They also show how navigating engagement markets and unsolicited interactions from men online pose significant challenges for both women and marginalised gendered identities.

Targeted objectification

Sabrina suggests there is something inevitable about the objectification of women that happens online: "As a girl posting online, you're guaranteed for men to do these kinds of things (harassment)." "I think that's why girls feel more comfortable having private Instagram accounts," she said.

Sabrina suggests that objectification is encouraged by the way the platforms work:

> I think apps like TikTok and Instagram contribute widely to the sexualisation of women, which makes men view women as sexual objects and would make men more likely to comment on unsolicited sexual things or send them unsolicited nudes.

As Rebecca suggests, rather than GenAI offering an opportunity for empowering women, this objectification continues the historical marginalisation and undermining of women based on their gender:

> People are using AI to create harmful content and [they] are also using it as a way to continue to degrade women and take away women's consent. It is pretty worrisome to think that this [GenAI] is out there that can continue to enable the ways that women are being targeted. Unfortunately, us women are the ones who are the victims of these things.

Grace added that,

> Social media, especially Instagram and Twitter, provides a platform for exposing intimate pictures, especially female nudes. There are many accounts that are dedicated to slut shaming and exposing females which [are intended] to harm them.

In different ways, our participants explain how women's bodies are made available for gratification without repercussion, thereby becoming a permissible form of violence. This illustrates the wider bearing that platform power dynamics and gender hierarchies have on mediated social life.

The male gaze

According to the participants, comments around body and beauty standards online are another way that the gender hierarchies and systemic gendered inequalities manifest through the proliferation of online GBV. Rebecca frequently witnessed "men mak[ing] comments on women's

social media, making comments about how they dress". She drew attention to how women are meant to "look and act" online, underscoring how the male gaze shapes how women represent themselves. The male gaze is a concept that refers to how the media depicts the world from a heterosexual male perspective that presents women as objects of male desire. It involves depicting women's bodies and behaviours in a manner primarily for the pleasure and from the perspective of the assumed heterosexual man (see Mulvey, 1975).

Grace shared how this media representation of women's bodies plays out on social media: "I find that gender plays a significant role on [platforms]. There's a lot of pressure to present yourself a certain way… a curated version of oneself." She suggests that expressing oneself as a women online result in both positive and negative responses simply for being a woman:

> I've experienced both positive and negative interactions related to my gender, including supportive comments from friends and followers but also unsolicited messages and objectifying comments from strangers who feel entitled to being in your space because of the access they have to you through social media.

The participants also raised concerns about unrealistic expectations around 'ideal' body standards and how they must engage in performative behaviours to navigate these expectations and protect themselves from online GBV. "The reinforcement of unrealistic beauty standards through AI filters amplifies gender stereotypes," Grace said. This perception was shared by Michelle who believes that "societal expectations around a woman being petite in stature can fuel potentially derogatory and violent conversations around online GBV"." We understand Michelle to mean that women labelled as 'fat' are frequently fat-shamed online, which can be considered a specific form of online GBV.

Rebecca concurred with Grace and Michelle by sharing her worries around how GenAI will exacerbate unrealistic expectations around body image and standards for women and other gender identities online. This poses the risk of increasing violent rhetoric against women. Rebecca expressed that she would not be "surprised if images of women were generated within the context of what industries are trying to sell us in terms of what our bodies are supposed to look [like]".

She described these expectations for women as being "women with bigger breasts" and "smaller waists." She said these body standards were "misogynistic" and explained that they place unrealistic expectations and pressure on "everyday women"." She felt that this use of GenAI will "give men more power" and expressed concern about GenAI "feed[ing] into fetishes or fantasies that may not be appropriate"."

Naomi shared similar concerns around body standards and expectations that will be amplified through GenAI. She stated that we can expect to see "things like ideal women" created by GenAI, and that already we can "see some aspects of objectification happening and even patriarchy" that comes with an "undertone of male superiority". For Naomi, "GenAI makes these issues feel less real." She said this was "dangerous because it gives a sense of fiction to the content being generated by GenAI. This makes it easy to detach from the consequences". These remarks suggest that under certain conditions GenAI may be deployed in ways that foster the male gaze, and that already this is being experienced.

These experiences and remarks allude to the consensus that GenAI and AI systems in general enable misogynistic behaviour towards women. They also reinforce body and beauty standards that women are expected to adhere to for the sake of the male gaze.

Self-protecting actions

To counter unsolicited messages and requests, amongst other forms of online attacks, women have often resorted to making their social media accounts private. The purpose is to protect themselves as well as their content. "I've had a few experiences of GBV," Freya said, "but I made my account private and Instagram doesn't allow a message to come into your inbox if you don't follow the person." Sabrina said, "Women have two options, either they have a private account and post themselves freely, or have a public account and be more restricted with what they post. Having a private Instagram ensures that you know exactly who is viewing your content."

Grace contributed to Sabrina's statement by expressing, "Most of the time the negative aspects outweigh the positive which is why I've recently deleted my main Instagram account and started a new one with followers that are only my trusted friends and family." Grace added that "[t]here's a feeling of entitlement in the air that you have to respond [to] …you have to accept a follow request and they [men] persist until you block them". This viewpoint was shared by Naomi who stated that she keeps her online "images locked", meaning that limited people online have access to her pictures and videos. These comments draw attention to how participants need to adjust their online behaviour to ensure their safety, mostly by limiting who can access their content. For woman, being safe online therefore has implications for the extent to which they can speak freely publicly.

Societal norms and responses to GBV

Participants suggested that online GBV is normalised through gendered cultural expectations, which implicitly perpetuate the cycle of online violence. Societal norms also limit the options women have in addressing online violations. For example, Fiona was clear about the cycle of victim blaming that reinforces patriarchy:

> Cultural and societal norms heavily influence the response to online gender-based violence in South Africa. Patriarchal attitudes normalise misogyny and violence against women both online and offline, perpetuating a cycle of victim-blaming and minimising the seriousness of gender-based harassment. These entrenched beliefs about gender roles and power dynamics further contribute to the normalisation of online abuse.

Sabrina expressed similar concerns:

> Societal norms affect it in the way that society makes it feel like it's the girls' fault for men doing things. In my opinion, our culture, in the coloured Muslim community, people don't care much about issues like online GBV, as it's a problem that is a modern issue, so it's not spoken about and is heavily disregarded.

A similar perspective was expressed by Freya when she spoke about victim-blaming and shaming. She said:

> I think that often women/men are ashamed to share such an experience because it's not often spoken about or the victim is often blamed. Victims may not want to report such

> incidents to protect the reputation of the family. The normalisation of abuse in some societies may prevent individuals from speaking out.

"Culture and society can impact the response to GBV very negatively," Fiona added. She said:

> When a female gets hacked and her nudes are shared with the world, she won't be seen as the victim, she will be villainised and made to seem at fault when in reality the hacker is at fault for sharing something that was private. Culture and society's response to instances of online GBV almost makes it seem like a joke and it's not a real issue.

In addition to gender, race, ethnicity and religion also shape the experience of online GBV and the steps one can take to minimise exposure. The nexus between these factors directly impacts misogynistic behaviours and attitudes. "Cultural norms that prioritise male dominance contribute to the acceptance or minimisation of gender-based violence," Freya said. She goes on: "In the Indian Muslim community it's not something that is spoken of. There is a lot of shame around it and fear of spoiling the family's reputation." These sentiments were echoed by Fiona:

> As a woman with intersecting identities in South Africa, such as race, ethnicity, and religion, my experiences of online gender-based violence are compounded by multiple layers of discrimination and marginalisation. Intersectionality exacerbates the impact of harassment [through] intersecting forms of oppression, such as racial or religious discrimination, amplifying the trauma and making it more challenging to seek support or recourse within a society that already struggles with systemic inequalities.

This was shared by Rebecca:

> Society and culture frames how women should exist, perpetuates and influences the prevalence of violence. I think that a lot of cultural and societal norms are violence in and of themselves, specifically ones that take away the rights and autonomies of the targeted individual.

Nancy suggested how gendered expectations for women at the intersection of mediated cultural norms and religion can be complex, with the outcome being that women's free expression is repressed: "As a coloured, Muslim woman I can see that Muslim women who cover up get more flack than women who don't cover up". She added:

> For instance a woman who makes a TikTok in a bikini would get praise whereas a woman wearing a headscarf and a burkini will get ridiculed for merely obeying her religion even though both of these women are free to make their own choices.

We understand this remark to reflect the demands for women to objectify themselves and their bodies for the male gaze.

These responses highlight how cultural, social, and religious contexts influence participants' experiences of online GBV. They also suggest how these contexts restrict agency when women try to act to prevent GBV, such as reporting instances to the police. They suggest the need for broader social change to effectively address online GBV.

Desensitisation and trivialisation

In the view of our participants, the normalisation of online GBV can cause the issue to become trivialised even amongst women. Naomi spoke of how "pornographic and sexually explicit images

can promote violence but also overexposure can desensitise us". According to Rebecca, 'shaming and blaming' contributes to this desensitisation: "Society makes you feel as though you should be ashamed of yourself if you experienced online gender based violence." She attributes shame to the naturalisation of everyday sexism and its forces of coercion, saying,

> We see it so often that it doesn't really register anymore and it is kind of like oh well that's messed up then we just move onto the next thing. But because it is so common it is so easily overlooked. As women we have been conditioned to overlook it to the point where it is really difficult to recognise it as violence. It is almost like we are desensitised to it and I would say that it happens more frequently than I pay attention to.

Grace shared similar perceptions around desensitisation: "They (men) sometimes joke about it (online GBV), they also dismiss the issue or blame victims for posting photos online in the first place." She continued by saying: "Subtle misogynies include the trivialization of concerns raised by women about AI misuse, most of the time it gets dismissed as attention seeking or handled in a way that doesn't make the victim feel supported but rather like they're being a burden."

Participants' statements suggest how online acts of violence against women are overlooked and normalised by society.

Attitudes towards content moderation

There were concerns expressed about the lack of content moderation on platforms, with Nancy linking this to a lack of regulation of platforms: "Because social media isn't regulated properly, it's easy to get away with online gender-based violence as there is no/not enough accountability for these perpetrators as they are being behind an online persona." This concern was reiterated by several of the participants, who then suggested areas of improvement with respect to the regulation of GenAI on platforms. For Grace,

> Platforms should enforce clear and comprehensive content moderation policies to swiftly remove harmful content, especially speech, harassment and non-consensual intimate images. Platforms should enforce community standards consistently and transparently, holding users accountable for violating policies related to GBV.

Grace made the link to the role of engagement markets in facilitating online GBV when she said,

> Despite the perceived safety of social media platforms, the advancement of AI technologies has introduced new avenues for exploitation and abuse. Platforms perpetuate GBV as they allow for exposure and hate speech despite having programs to detect and block these things. Certain algorithms also promote provocative pictures of women or stereotypes. The ability to create anonymous accounts limits accountability and exposure of the perpetrators which just makes it seem easier to do.

The concerns the participants had around a lack of content moderation online was clear, as well as how this limits their safety online, especially with respect to anonymity, which encourages harassment and unsolicited engagement from men.

Analysis and Interpretation

The role of engagement markets in driving female fear factories

The combination of anonymity, lack of consent, commercial incentives, and GenAI's ability to produce volumes of hyper-realistic yet entirely artificial pornography poses a threat to personal privacy, digital trust and safety for all. When it comes to engagement markets being an enabling environment for online GBV, platforms confront a bind. The vitriol that proliferates on their platforms is a substantive driver of user engagement, posing a tension with content moderation initiatives to detoxify these digital spaces (Kumar et al 2018). Information disorders, whether they manifest as harassment, misinformation or online GBV, have proven to be powerfully addictive, igniting outrage that compels certain users to engage further on platforms and refresh their feeds (Rajadesingan, Resnick, & Budak 2021; Rajadesingan, Resnick, & Budak, 2020). This engagement sustains the data-mining advertising models that are the source of platform revenue.

Earnest efforts to address information disorders from platforms could undermine their revenues. However, the incessant exposure to information disorders can assert the male gaze on even casual users, incrementally acclimating them to a view of women that systematically disturbs the boundaries of acceptable online discourse. This dynamic allows bad-faith actors to leverage platforms' commercialised architecture to foment a hostile climate against women. Platform owners' and managers' fears of undermining their revenue limits their incentives to address information disorders such as online GBV, including GenAI non-consensual intimate images, and other parts of the female fear factory.

Political backlash through commercial platforms

Overall, women have found themselves ensnared in a vicious cycle perpetuated by patriarchal norms. Their worth has been inextricably linked to sexualised ideals crafted by men. Unless they conform to these 'ideals', women risk being overlooked or undervalued, as our interviews suggest. However, even when women meet such ever-changing 'ideals', they face the prospect of being objectified and trivialised by men. Rather than being seen as equals, women are reduced to mere accessories, robbed of the respect afforded to their male counterparts.

More broadly women are caught up in a backlash to the gains they have made. Women and marginalised groups have gained influence in many spheres and, crucially, have acquired the means to impact political debates, but there is also a backlash to their growing influence. Misogyny and online GBV perpetuated by men can be partly explained by their perceived loss of traditional masculine privileges and 'naturalised advantages' through the establishment of formal and procedural equality. Additionally, as gender roles evolve, some young men may feel resentful and uncertain about their social positions, making them susceptible to reactionary agendas which exploit these frustrations and channel them into ideologies that promise to restore gendered entitlements (Nilan 2021; Kaiser, 2022). These dynamics manifests in hostile online spaces targeting women's progress.

Content moderation and 'accountability theatre'

When it comes to content moderation, at least for big firms, many platforms' policies typically only go beyond any legal 'mandatory minimums' because advertisers demand a space where hate speech will not drive away users. Conversely, one reason for the lack of minimally sufficient

content moderation in some Global South spaces is because these markets are not deemed valuable enough (De Gregorio, & Stremlau, 2023).

Local nuance is also lost through moderation efforts. Much human content moderation is outsourced to India and the Philippines. Furthermore, people in the Global South are moderating content for markets in the Global North while there is no content moderation in their societies. Again, this is emblematic of global inequalities. Indeed, some researchers talk about non-white racial groups in the Global South 'cleaning up' after whiter societies in the Global North (See Ahmad & Krzywdzinski, 2022).

Another consideration is that the prevailing view of content moderation, which dominates both regulation and academic discourse on online speech governance, is misleading and insufficient (see Douek, 2022). This perspective portrays content moderation as an equivalent to adjudication of speech rights wherein legislative-style rules are applied repeatedly to individual pieces of content by a hierarchical structure of moderators. This understanding leads regulators and academics to believe that holding platforms accountable for online speech decisions is best achieved by ensuring users receive individual review like court hearings around due process rights. However, this approach results in mere 'accountability theatre' rather than true accountability.

Rather, the volume and speed of online content creation and distributions makes it impossible to view content moderation simply as a collection of individual adjudications and the correction of individual decisions. Instead, content moderation should be seen for what it is: a project of mass digital administration with dynamic, automated systems offering proactive, continuous decisions. Regulators need to adopt systems-thinking approaches which account for structural, procedural, and technical mechanisms that moderate content.

Conclusion and Key Recommendations

From our vantage point, opaque engagement markets have become part of the female fear factory, by which we mean that these GenAI systems form part of the apparatus that "attempts to silence, humiliate, or kill them [women] in public spaces," to refer to Gqola's (2021, p. 12).

Policymakers must recognise the emerging phenomenon of GenAI variations of online GBV which are fundamentally altering the experience of platforms and media systems. Strong regulatory guardrails and accountability measures are needed, such as the demonetisation of non-consensual sexual depictions online. Urgent multi-stakeholder action involving police investigatory capacity, platform governance and the actioning of AI ethics mandates are required to address this new phenomenon.

Policymakers must recognise that GenAI-enabled GBV fundamentally changes the media landscape for women. Regulators should work with platform owners to develop efficient procedures for submitting and evaluating requests for image removal. In addition, proactive detection of synthetic media should also be encouraged, where possible. Regulators must not lose sight of how misogyny, combined with the commodification of data and engagement markets, can drive the circulation of non-consensual intimate images.

If necessary, regulators should leverage their discretionary authority, as well as new legislative powers if needed, to compel platforms to adopt comprehensive policies enabling individuals to request the removal of fake intimate imagery. Consistent monitoring and evaluation of

compliance with these policies is paramount, with potential civil or criminal penalties for non-compliance.

There is value in governments updating legal frameworks to provide victims of deepfakes and involuntary synthetic media with new causes of action and pathways for legal recourse. Such laws could potentially establish new torts allowing civil suits for damages, as well as statutes criminalising the non-consensual creation and dissemination of synthetic intimate imagery. Even so, existing laws around privacy, defamation, and non-consensual pornography may provide some avenues for redress. The larger point perhaps is that doctrinal reviews are required to assess whether the proliferation of GenAI media has outpaced the ability of existing laws to adequately address the multifarious harms they engender.

Finally, policy interventions must acknowledge the complexity of technical systems and business models, while also engaging with gender hierarchies. This broader perspective can enhance discussions on information disorders, situating them within the ambit of general and established democracy, economic development and human rights promotion efforts. Concurrently, it is important to recognise and address the backlash faced by women and marginalised groups who have gained relatively more influence and acquired the technological means to affect political outcomes. Other interventions that address the root causes of this expression of patriarchy are needed.

References

Ahmad, S., & Krzywdzinski, M. (2022). Moderating in obscurity: How Indian content moderators work in global content moderation value chains. In M. Graham & F. Ferrari (Eds.), Digital work in the planetary market (pp. 77-95). MIT Press.

Alvarez, M. B., & Alessi, E. J. (2012). Human trafficking is more than sex trafficking and prostitution: Implications for social work. Affilia, 27(2), 142-152.

Amnesty International. (2023, September 28). EU: AI Act must ban dangerous, AI-powered technologies in historic law. https://www.amnesty.org/en/latest/news/2023/09/eu-ai-act-must-ban-dangerous-ai-powered-technologies-in-historic-law/

Baekgaard, K. (2024). Technology-facilitated gender-based violence. Georgetown Institute for Women, Peace and Security; Embassy of Denmark.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1).

Carlin, M., & Federici, S. (2014). The exploitation of women, social reproduction, and the struggle against global capital. Theory & Event, 17(3).

Chen, C., & Shu, K. (2023, November 9). Combating misinformation in the age of LLMs: Opportunities and challenges. arXiv:2311.05656.

De Gregorio, G., & Stremlau, N. (2023). Inequalities and content moderation. Global Policy, 14(5), 870-879.

De Gruchy, T., Vearey, J., Opiti, C., Mlotshwa, L., Manji, K., & Hanefeld, J. (2021). Research on the move: Exploring WhatsApp as a tool for understanding the intersections between migration, mobility, health and gender in South Africa. Globalisation and Health, 17, 1-13.

Douek, E. (2022). Content moderation as systems thinking. Harvard Law Review, 136(2).

Faith, B. (2022). Tackling online gender-based violence; understanding gender, development, and the power relations of digital spaces. Gender, Technology and Development, 26(3), 325-340.

Gibson, K. (2022). Bridging the digital divide: Reflections on using WhatsApp instant messenger interviews in youth research. Qualitative Research in Psychology, 19(3), 611-631.

Goldin, C. (2021). Career and family: Women's century-long journey toward equity. Princeton University Press.

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? PNAS Nexus, 3(2).

Gqola, P. D. (2021). Female fear factory. Melinda Ferguson Books.

Hanmer, J. (1990). Men, power, and the exploitation of women. Women's Studies International Forum, 13(5), 443-456.

Hanna, A., & Bender, E. M. (2023, August 12). AI causes real harm. Let's focus on that over the end-of-humanity hype. Scientific American. https://www.scientificamerican.com/article/ai-causes-real-harm-lets-focus-on-that-over-the-end-of-humanity-hype/

Hao, K. (2021, February 12). Deepfake porn is ruining women's lives. Now the law may finally ban it. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

Herrman, H. (2024, March 26). Who's behind all the 'pussy in bio' spam on X? New York Magazine. https://nymag.com/intelligencer/2024/03/whos-behind-all-the-pussy-in-bio-spam-on-x.html

Hlomani, H., Orembo, L., Schroeder, Z., Schültken, T., & Timcke, S. (2023). Policy reinforcements to counter information disorders in the African context. RIA Policy Brief, no. 2/2023. Research ICT Africa.

Kaiser, S. (2022). Political masculinity: How incels, fundamentalists and authoritarians mobilize for patriarchy (V. A. Pakis, Trans.). Polity.

Koebler, J. (2024, April 9). 'AI Instagram influencers' are stealing women's bodies. 404 Media. https://www.404media.co/ai-instagram-influencers-are-stealing-womens-bodies/

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the Web. In Proceedings of the 2018 Web Conference (pp. 933-943). ACM.

Lindquist, J., & Weltevrede, E. (2024). Authenticity governance and the market for social media engagements: The shaping of disinformation at the peripheries of platform ecosystems. Social Media + Society, 10(1).

Maiberg, E. (2024, March 28). 'IRL fakes:' Where people pay for AI-generated porn of normal people. 404 Media. https://www.404media.co/irl-fakes-where-people-pay-for-ai-generated-porn-of-normal-people/

Mies, M. (1986). Patriarchy and accumulation on a world scale: Women in the international division of labour. Atlantic Highlands.

Moreau, S., & Rourke, C. (2024, February 8). Fake porn causes real harm to women. Institute for Research on Public Policy. https://irpp.org/op-ed/fake-porn-causes-real-harm-to-women/

Mulvey, L. (1975). Visual pleasure and narrative cinema. Screen, 16(3), 6-18.

Mwanda, Z. (2022). Text, voice-notes, and emojis: Exploring the use of WhatsApp as a responsive research method for qualitative studies. Critical Studies in Teaching and Learning, 10(1), 78-92.

Nilan, P. (2021). Young people and the far right. Springer.

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In Proceedings of the 14th International AAAI Conference on Web and Social Media (pp. 968-979).

Rajadesingan, A., Resnick, P., & Budak, C. (2021). How social learning amplifies moral outrage expression in online social networks. Science Advances, 7(33).

Robertson, C., del Rosario, K., & Van Bavel, J. J. (2024, April 1). Inside the funhouse mirror factory: How social media distorts perceptions of norms.

Salmons, J. (2009). Online interviews in real time. Sage.

Shahid, S., & Shaikh, M. A. (2019). Impact of "WhatsApp Chaupal" on the academic performance of graduate students of Karachi–A case study. FWU Journal of Social Sciences, 13(2), 94-107.

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus, 3(2).

Timcke, S. (2021). Algorithms and the end of politics. Bristol University Press.

Timcke, S., & Hlomani, H. (2023, February). Decoding the ballot: How might AI reshape democracy on the African continent? Research ICT Africa.

Timcke, S., & Rens, A. (2023, August). Preventing platform decay: Regulatory solutions to counter treacherous turns. Research ICT Africa.

Timcke, S., Hlomani, H., Makumbirofa, S., Rens, A., & Nawal Omar, N. (2024). The political economy of African AI: A primer on concepts, contexts, considerations and capitalism. Research ICT Africa.

Timcke, S., Orembo, L., & Hlomani, H. (2023). Information disorders in Africa: An annotated bibliography of selected countries. Research ICT Africa.

Timcke, S., Orembo, L., Hlomani, H., & Schültken, T. (2023). The materials of misinformation on the African continent. Mid-year report. Research ICT Africa.

United Nations. (1993). Declaration on the elimination of violence against women.