Feminist Perspectives on Social Media Governance

February 2023

IT for Change

INTERNETLAB
law and technology research center

Acknowledgements

Editors: Avantika Tewari and Amshuman Dasarathy

Editorial Inputs: Anita Gurumurthy

Editorial Review & Copyedit: Sohel Sarkar, Sadaf Wani, and Intifada P. Basheer

Proofreading: Intifada P. Basheer, Sadaf Wani, and Viraj Desai

Design & Visual Support: Intifada P. Basheer

Report Layout: Jojy Philip

This is a curated compendium of essay submissions from a two-day closed roundtable on 'Feminist Perspectives on Social Media Governance'. The roundtable was held from 10-20 April 2022, and was organized by IT for Change and InternetLab.

Both the roundtable and this compendium are part of IT for Change's 'Recognize-Resist-Remedy' project, supported by IDRC (Canada), the World Wide Web Foundation, and the Ford Foundation.

Contents

Editorial

AVANTIKA TEWARI and AMSHUMAN DASARATHY

The original promise of freedom and democracy offered by the internet as a space for networked communication, self-expression, political participation, and public interaction beyond national borders has swiftly come and gone. The past decade has witnessed a steady rise in the amount of extreme content on online platforms in the form of cyberstalking, bullying, harassment, hate speech, disinformation, and propaganda. The intensity, scale, and ubiquity of these problems has brought to the fore the need to reimagine pathways for building a global communicative justice paradigm along feminist principles. Today, we are confronted with misogynistic, hateful, and dangerous content on an unprecedented global scale. Developing a feminist ethic of social media governance to counter such harmful rhetoric, is therefore an urgent and pressing need. An important step in this direction is to recognize the function that extreme speech plays in deterring equal political participation and promoting further marginalization of historically oppressed sections of society. The lack of recourse to address violence inflicted by hate speech further pushes us to probe the problem through an examination of the harms that have become normalized as part of digital culture.

As part of our efforts to challenge and check the normalization of such extreme speech, we at IT for Change have been working to tackle and understand the scale of the problem of pervasive online misogynistic speech. It is within this context that the research project titled, 'Recognize-Resist-Remedy' was initiated in August

2019 with the overall objective of evolving a comprehensive socio-legal strategy to address the proliferation of sexism and misogyny in the internet-mediated public sphere, as a collaborative endeavor between IT for Change (India) and InternetLab (Brazil). The project explores how women's first-order right to participation can be reclaimed in the platformized publics of the internet age. One of the many project outputs interrogating digital gender-based violence was the two-day closed roundtable on 'Feminist Perspectives on Social Media Governance' organized on 19-20 April 2022.[1]

Through this roundtable, we sought to envision new imaginaries of social media governance to eradicate the unfreedoms arising from misogyny in online communications agora. The two days of engaging discussion, and the cross-pollination of ideas from diverse disciplinary standpoints provided us with innovative ways to think about online misogyny – its social embeddings and the trajectories of legal responses to it. This global roundtable on feminist social media governance facilitated a comparative analysis of platform governance experiences across different contexts in the Global South. The roundtable witnessed participation from feminist scholar-practitioners from different parts of the world. A compendium with initial inputs from attendees was produced prior to the roundtable and this is the publication of a final volume of essays with a subsequent synthesis report.

The roundtable hosted lawyers, academics, scholar-practitioners, and activists, whose conversations centered around the following three questions:

1. What do empirical studies of platform regulation (and self-regulation) tell us about addressing sexism and misogyny online?
2. What national legal-institutional frameworks may be appropriate to check gendered censorship and promote gender-equal participation in the online public space?
3. What kind of global responses may be relevant towards nurturing gender-inclusive online publics?

[1] This roundtable was part of our Recognize-Resist-Remedy project, supported by IDRC (Canada), the World Wide Web Foundation, and the Ford Foundation.

The focus was on chalking out the broad contours of future directions for platform regulation and techno-design overhaul to effectively address sexist hate from an intersectional perspective, without intensifying gendered censorship cultures. This collection of essays is a reflection of the enriching debates and conversations shaping the direction of efforts towards ensuring platform accountability and political action on addressing online gender-based violence (OGBV).

Part 1: Gendered Political Violence

Mardiya Siba Yahaya's essay, 'Creating Feminist Online Spaces by Conceptualizing the Experiences of Visible Muslim Women', examines how online public spheres have been characterized by masculine hegemonies which inhibit women's online participation. Women's conditional access to the online public sphere is mediated by the imminent danger of violence, policing, and harassment, thus shifting the onus of privacy or protection onto women, as individuals liable to maintain their own safety on the internet. However, this expectation is often countervailed by the gendered violence that is pervasive in society. In this essay, the author explores the patriarchal terms of access that define women's mode of (in)visibility in the digital public sphere. Studying the implications of moral surveillance and scrutiny of African Muslim influencers online, the essay critically interrogates the various tools to address gendered surveillance and argues for the importance of re-grounding technology in a social context, embodiment, and place.

Yasmin Curzi de Mendonça's essay, '"I Silenced Myself": Brazilian Female Activists on Online Gender-based Violence', presents partial results of her doctoral research at the State University of Rio de Janeiro, where she interviews Brazilian activists, politicians, and journalists who have experienced OGBV. The study explores the strategies of resistance embraced by women in the face of adversarial situations, also contending with the insufficiencies in the current Brazilian legal framework. The essay also points to the culpability of social media platforms in perpetuating OGBV.

Part 2: Contextual Content Governance

Divyansha Sehgal's essay, 'Gendered Hate in a Video World: An Analysis of Gender-based Harassment and Community Guidelines for Short-form Video', offers a study of community guidelines that are officially documented by platforms to establish ground rules for users, lay out acceptable uses of their technology, and outline the recourse available to users when these guidelines have been violated.

The essay explores community guidelines for short video sharing apps in India (Moj, MX Takatak, etc.), to answer whether and how these apps incorporate local contexts, and what the sector can do better to make the platform safer for all users.

Arjita Mittal's essay, 'Moderating Bodies: Reproducing Systems of Power through Platform Policies', attempts to articulate the unspoken, yet active, cultural work performed by community guidelines that frame people's participation on platforms. It highlights the need for a context-specific regulatory document on the behaviors of platform users. The essay brings to the fore the doublespeak of Big Tech platforms that profess to develop community guidelines for the benefit of their users, while actually curbing user expression and autonomy by upholding hegemonic normative standards of the community.

Part 3: Community-led Moderation Systems

Arnav Arora, Mahalakshmi J., and Cheshta Arora's essay, 'Designing for Disagreements: A Machine Learning Tool to Detect Online Gender-based Violence', recounts the authors' experience of building a user-facing browser plug-in to detect and mitigate instances of OGBV in three Indian languages: Indian English, Hindi, and Tamil. It describes the nature of data collected across languages and forms, how it represents OGBV, and how the 'problem' of building an OGBV detection tool for journalists, activists, leaders, community influencers, and celebrities translates into reality on Twitter. The essay offers an expanded definition of OGBV to include instances of misogyny, sexism, transphobia, as well as hate speech, 'creepy' comments, and abusive political/electoral speech, and presents an overview of the algorithmic responses to such instances of violence.

Quito Tsui's essay, 'Feminist Communities of Care and a Transnational Response to Redesigning Social Media Governance', goes beyond the usual inquiry of the deeply flawed system of social media governance that attempt to answer: Who is being harmed in this space? What legal mechanisms are there for accountability? How can we remove or punish those engaging social media's purposefully misogynistic design? The author argues, while these questions are important, they also limit our ability to be emboldened in the kind of interventions we seek. The author urges that we interrogate what it will take for us to be able to account for the varied ways in which misogyny and platformized hate arise. Exploring the ways in which feminist care is practiced, and understanding communities of care, she submits, can offer us a holistic lens for redesigning of social media

governance. Transnational feminist work on international solidarity is seen to provide possible ways forward for a global response to the vagaries of social media governance.

Part 4: Legal Responses to Online Gendered Violence

Kim Barker's essay, 'The Responsibility Gap and Online Misogyny: Beyond Misogyny, through Violence, Towards (Legal) Safety?', reveals how online misogyny is a phenomenon encompassing one of the many forms of violence perpetrated against women online. It is a pernicious, silencing experience designed to subjugate women online and limit their participatory rights. The essay argues that legal systems fail to define online violence against women (OVAW), frequently excluding gender (and/or misogyny) from the ambit of hate crime frameworks and rendering women online to the lowest priority for protection. It demonstrates the ways in which responsibility gaps persist beyond the law as well, and how the fragmented approaches of policy bodies, international organizations, online platforms, and the law compound the problem. It articulates the need for a holistic, 'joined up' approach to tackling misogyny, evaluating its impact and harms, and also suggests approaches to provide online safety by bridging the existing responsibility gap.

Esther Lee's essay, 'Internet Safety, for Whom? South Korea's Post-Nth Room Platform Governance', offers a close look at the May 2020 revisions to South Korea's Telecommunication Business Act, that challenged the assumptions of platforms as passive intermediaries and uninhibited open space for content and information exchange. These revisions belatedly recognized the consumption and circulation of non-consensual intimate images (NCII) by networked publics, and expanded the regulatory obligations of domestic social media and open community forums to directly filter, report, and takedown exploitative materials. A year later, this imperfect, yet landmark legislation faces an imminent threat of scale-back under the new conservative administration. This essay interrogates the rejection of expanded intermediary liability under the guise of favoring anti-censorship regimes. The author argues that the instrumental invocation of the right to free expression conceals the kernel of sexism that gets reinforced by uncritically rendering the internet as an ungovernable space. An unqualified anti-censorship stance, in fact, comes at the expense of women's rights to safe and full digital participation. By examining South Korea's current legislative impasse, therefore,

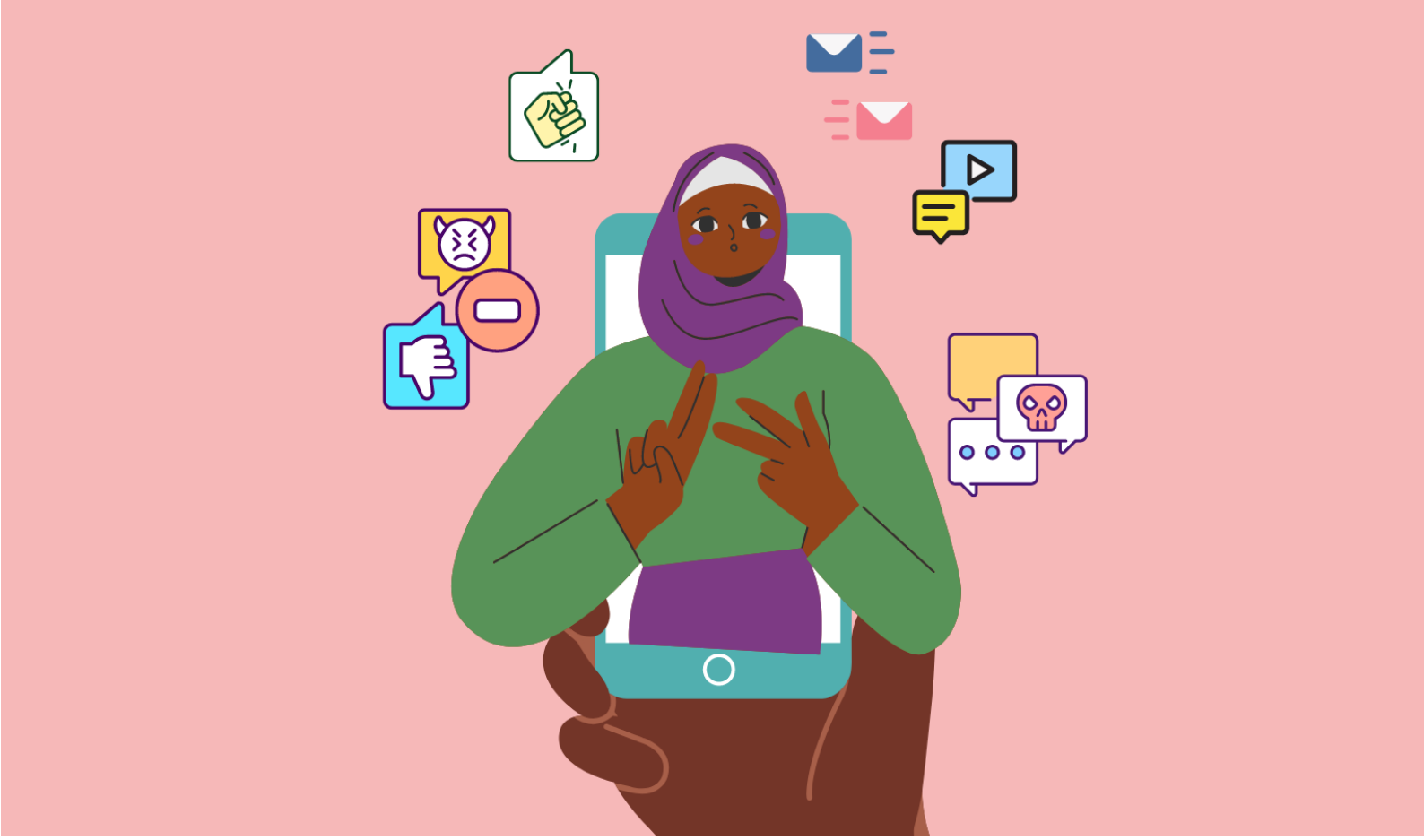
Lee asks: What are the feminist conceptions of platform standard-building; and what are the strategies for countering the fallacy of gender-inclusive online publics as being mutually exclusive from internet freedom?

Part 5: Racialized Techno-political Organization of the Internet

Anne Njathi and Rebeccah Wambui's essay, 'A Digital Patriarchal Bargain: Violence against Women in Kenya', is predicated on the quantitative increase in the use of the internet, particularly among women, whose qualitative participation, however, is undercut by new weapons such as shadow banning, trolling, and cyberbullying among others, to silence and shame them. Through case studies, the findings of the essay indicate that while violence against women and queer folks continues unabated on social media platforms; platforms' regulation policies purported to correct the harms, actually, reinforce gender and sexual hierarchies and class-racial inequalities to reify differential and unjust treatment of people belonging to marginalized sections. The essay also offers a close reading of the Kenyan Computer Misuse and Cybercrimes Act, offering constructive criticism to envision governance and policies that dismantle systemic gender inequalities unfolding within digital platforms in the context of developing African countries.

Graciela Natansohn's essay, 'Against Internet Coloniality: Notes on Misogyno-Racist Violence on the Internet', is a theoretical exploration, providing a structural critique of violence against women in digital environments. It invokes the work of Rita Segato on the etiology of online-offline gender violence to establish the constitutive inseparability of the two seemingly disparate spheres. It also unpacks the media rationale governing platforms and algorithms that structure communications on the internet. The author argues that this misogynist-racist violence cannot be interpreted as deviations or flaws of the digital communication processes. Instead, they are rational and grounded manifestations of power, knowledge, being, and internet coloniality, that are a necessary component of the current colonial, modern, and binary socio-technical design.

Gendered Political Violence



Creating Feminist Online Spaces by Conceptualizing the Experiences of Visible Muslim Women

MARDIYA SIBA YAHYA[1]

Abstract

In imagining a feminist online 'public' space, this essay draws from the histories of violence to explain and investigate current trends of technology-facilitated violence. The essay centers visibility as a conceptual framing to explain gendered surveillance and how African Muslim women creators experience social media. In this work, the understanding of surveillance considers the key role of moral shaming in controlling Muslim women's free use of online spaces and how that is

[1] Mardiya Siba Yahaya is a feminist digital sociologist, researcher, and community movement builder, whose work examines the internet and its cultures from the margins of gender, sexuality, race, geo-political location, and religion. She contributes to creating safer and secure digital societies using intersectional feminist standpoints and frameworks.

closely related to women's experiences of the public domain in offline spaces. The essay also shows how surveillance today involves multiple actors and scenarios which make it more complicated than earlier conceptions of discipline and control. By drawing from historical trends of violence, gendered moral surveillance, and the design of female fear, the essay grounds cases of violence and safety online within the social contexts of Muslim women in Africa. The same framing is used to argue and explain how technology design and policies can center gender, safety, privacy, and inclusivity. Most importantly, by centering gender and feminist frameworks in technology research and work, this essay communicates that a feminist social media governance practice must acknowledge how patriarchal hegemonies and structures offline play out online, and how our current actions become limited because of these realities.

1. Introduction

Some histories of gendered violence can be traced through the transition from feudalism to capitalism. During the transition, Silvia Federici, explains that the domestication of women was an intentional design to enable the advancement of capitalism. Federici demonstrated through her books, 'Caliban and the Witch' and 'Witches, Witch-hunting, and Women' that the spectacle created by burning 'witches' at the stake was a patriarchal tactic to instill fear and force women to police and surveil each other.

Federici further explains in 'Caliban and the Witch' that gender, reproduction, and sexuality became the patriarchal tactic for control and instilling the language of fear. Here, women self-regulated to avoid and navigate murder through witch hunts, which have evolved into femicides, and gender-based violence. In line with Federici's work, Pumla Dineo Gqola conceptualizes the female fear factory as an intentional patriarchal organization of society that limits women and gender non-conforming persons' interaction in public spaces. For Gqola, the female fear factory exists to remind us that public spaces are masculine turfs and people who are not cis-hetero-men should not freely exist within it. In this way, our bodies are 'made' public, and morality becomes the lingering source of shame. Both Federici and Gqola agree that the public display of violence is used to create a collective trauma that communicates to women that their safety lies in regulating their excessiveness.

It is along these lines that the concept of visibility, concerning how certain people exist and enjoy 'public' goods and as a tool of discipline and surveillance, becomes

particularly intriguing because of the paradoxes it produces. On one hand, it is an effective protest tool against patriarchal structures and hegemonies that use violence to remind women that the pleasures of public spaces do not belong to them. Feminist theorists and activists such as Mona Eltahawy suggest taking up space, seeking attention and intentionally centering pleasure in public spaces to challenge patriarchal hegemonies. At the same time, it becomes weaponized against online creators with claims that they are agentic users who make money, and contribute to the platform's gains and visibility and thus should not expect safety, security, or privacy.

Meanwhile, the design and organization of violence against women, queer people, and non-white cis-hetero men centers moral surveillance and shaming as the panoptic warden that controls and punishes people who transgress normative frames of belonging.

During the transition into capitalism, as referenced, women faced threats of witch hunts, and this was not reserved for only poor people. Engaging in economic activity, reproductive and lifestyle decisions that did not include a marriage and giving birth became a moral failure and source of shame that warranted a woman to be burned.

The same logic governs our contemporary public spaces, where women's paid labor of any form is diminished and is made to become a source of shame. To work and be adequately paid for our work requires us to freely engage, interact and enjoy these 'masculine public spaces'. Social media thus far has sought to create a space that allows 'all' people to engage and connect, providing the opportunity for the online content creation industry to emerge and progress. However, this industry mostly relies on women's labor and creative expression and centers their visibility to gain social currency which in the long run becomes a source of income. Here, women, directly and indirectly, challenge the female fear factory by becoming the content or the creators who enjoy the online public space. However, the failure of product designers to center gender, class, and racial justice causes certain groups such as African Muslim women to face technology-facilitated violence due to their position at the intersections of structural, hegemonic, disciplinary, and interpersonal oppression.

Visibility continues to become a fascinating concept because it reinstates Gqola's explanation that to challenge and eventually dismantle the female fear factory, feminism becomes a dangerous job. Similarly, Gqola explains that "women

are terrorized into appearing to choose passivity", which they also choose as a "safety measure". Thus, in acknowledging the danger posed by visibility, Eltahawy references the need to seek attention as a 'necessary sin for women'. Meanwhile, it is through their understanding of such dangers that Muslim creators identify ways to personally navigate while existing and engaging online. Meanwhile, the lingering fear of violence through policing and moral surveillance forces Muslim women content creators to self-censor, and these actions help surveillance systems and cultures control what individuals will do in the future. This work centers Muslim women content creators as their experience in online spaces exemplifies the tension between requiring women to be visible while shaming their visibility.

The power relations reproduced online are embedded in structural forms of oppression where the universality of design creates biases that privileges certain groups over others. These structures also make it difficult for African Muslim women content creators to experience the pleasures of their online presence and engagement. Similarly, disciplinary oppression is what organizes the surveillant gaze that watches and polices every action of the Muslim woman. Meanwhile, the hegemonic system of oppression places morality as a reason for surveillance to occur, and the interpersonal sphere is where Muslim women through their daily online interactions experience biases and discrimination for being women, African and Muslim. This essay explores how online violence at the intersections of design injustices, the matrices of domination, and the female fear factory can be addressed through design-based frameworks and platform governance policies.

2. Content Creation, Visibility, and Moral Surveillance

Content creation for Muslim women represents hobbies, artistic freedom, and professions. For instance, Layla,[2] a UK-based Nigerian creator, whom I interviewed for my research in 2021 and who began her journey four years ago shared that being a content creator is her full-time job. Meanwhile, Latifa, a photographer and food blogger based in Cape Town, explained that she always loved taking photos and social media presented an opportunity to showcase her work besides her daily profession. In these two stories, we come to understand the pleasures of using social media for Muslim women. Further, the creator economy has advanced and come to play a significant role in shaping digital cultures, marketing, advertising, and the artistic consumption industry.

[2] This essay uses pseudonyms for all participants to protect their privacy.

Online, Muslim women have found a way to express themselves; directly and indirectly challenging patriarchal hegemonies. Content creation also comes with some form of visibility as platforms require influencers to retain engagement to effectively monetize their content. In this domain, visibility yet against presents paradoxes that place Muslim women between the contours of surveillance and control. On one hand, Muslim content creators are exposed to forms of monitoring, shaming and violence that are mostly carried out by male patriarchs online. Yet, on the other hand, to be a successful content creator, one must continue to share themselves and their work regardless of the violence they face.

Revisiting Federici, the collaboration between capitalism and patriarchy ensured that women's and gender nonconforming peoples bodies became the concern of the public, allowing our bodies and choices to be sites for surveillance. Yet, a significant aspect of online content creation is to center oneself through fashion and beauty, which directly challenges patriarchal hegemonies that demand women to be 'hidden'. Muslim women get to make money from 'selling beauty', which further defies the fear factory that seeks to limit women's existence within public spaces. By positioning themselves as the content, Muslim women intentionally and unintentionally resist the very systems of discipline that make them available for scrutiny and violence. Thus, to restrict women from further benefiting from online space, their source of income becomes a subject of ridicule; they are attacked physically and bullied online. The spectacle of the continuous violence against Muslim women is why some Muslim creators shared that they prefer to do "what is right" to avoid violence. Doing "right" for the creators involves self-policing, and adhering to patriarchal scripts that are performances of a certain 'modest' and 'non-excessive' womanhood that should afford them safety.

For instance, Layla explained that she has struggled with her journey as a content creator, partly because she faces industry biases against black women who are Muslim. She also finds she and other influencers in her community are constantly being shamed and undermined for the work they do. Layla added that while influencers carried the creative industries during the Covid-19 lockdowns and produced content that required entire creative teams, their work continues to be the subject of ridicule. People still do not consider influencers to have 'real jobs' as their work relies on virtual environments. Layla's concerns relate to Kryss Networks' argument that the fact that violence happens online does not make it less real because it is in a 'virtual' space. Acknowledging violence within its different

contexts is the first step towards addressing technology-facilitated violence on a structural level. Just like Layla, Latifa too expressed her concern for the disregard for influencer labor and shared her experience of becoming the face and subject of a hate train for supposedly receiving free products.

How does the disregard for influencer labor explain moral surveillance? Women's labor that centers their bodies and beauty is considered a patriarchal moral failure that deserves disciplinary action. First, Muslim influencers are expected to become faith-based role models and strict standards of behavior are imposed on them. When they fail to meet these standards, they are shamed for threatening Islamic morality and values by centering themselves as content online. Second, by alluding to the idea that their work is not 'real', moral surveillance represents patriarchal pushback which in turn limits their will to engage and retain their reach. This, in the long run, inhibits their free use of online public spaces. In this way, as argued, morality becomes the panoptic warden that forces Muslim women to self-censor and control others in their space with the promise that they may be safe. For instance, Layla shared that she does not wish to become the subject of scrutiny and thus far has not, which she believes is a result of being 'lucky'. She explained, however, that people within the Muslim influencer community have been subjected to violence for merely existing in the space or for deciding to no longer wear the hijab.

The controversies surrounding 'selling beauty' according to both Latifa and Layla, are mostly embedded in misogynoir.[3] They also acknowledged that their hopes to do the 'right thing' for safety in a patriarchal system will never be enough. In this line of thinking, Bakkar argues that to escape any form of gendered surveillance and violence would require Muslim women to become white cis-gendered men which is something they are unable to do. Bakkar's point also brings our attention to the intersection of the biological essentialism associated with gender, the social privileges attached to being part of certain communities, and whose body is declared worthy of protection and whose requires more surveillance.

Furthermore, online content creators' push to 'put themselves out there' is used as a reason to justify harassment against them. Similarly, people who police

[3] Misogynior is a word coined by feminist Moya Bailey in 2008 to refer to anti-black racism and sexism that black women experience. The word uses the concept of misogyny, which is sexism and hate against women, and the French word for black, which is *noir*. Misogynior has since been used on Twitter, and in scholarly articles to highlight the specific kinds of sexism black women experience.

women do so believing that they are rightfully keeping excessive women in check. Simultaneously, surveillance as argued by Foucault relies on visibility, and "discipline is enforced through the creation of various forms of supervision". Thus, being a visible Muslim woman online makes surveillance possible. Yet, people who shame Muslim women online for being 'seen' and not 'hidden' use that same visibility to harass and attack them. At what point has patriarchy afforded women and people who defy gendered binaries their 'privacy'?

3. Design Injustice and Online Violence

Social media companies, in response to community and capacity-building needs, began to host a series of creator gatherings. Some of the sessions within these gatherings included how online creators could grow into being successful entrepreneurs and provided opportunities for some form of collaboration and networking. Meanwhile, when asked whether they were familiar with these gatherings, most of the women who engaged in my study shared that they were not usual guests or had never heard of these spaces. This made me ask two questions; first, who gets access to these spaces? Most importantly, are these spaces where creators are educated on community safety guidelines and provided support on their challenges? Can they share their concerns and needs with social media companies regarding the violence and harms they face?

The first assumption is that the content creator spaces and gatherings focus on teaching influencers to make more money, retain engagement and be better 'entrepreneurs' by understanding algorithmic updates. Influencers have arguably been at the core of the growth in modern popular culture, and supported the rise of different social media such as YouTube, Instagram, Twitch, and most recently TikTok. Technology companies and influencers converge at a point where both parties want to make profits from their product and their creativity respectively. It is assumed that safety and protection are in-built features for most users, and are not the main agenda for gatherings or conferences with creators. What is more worrying is if creator gatherings are spaces where everyone pretends violence does not exist and focuses on the 'good' affordances. Meanwhile, for African Muslim women creators, their pleasure and ultimately financial gain are diminished because of their position in the matrix of domination. In this way, as argued by Sasha Costanza-chock, this group of people comes to experience microaggressions and violence facilitated through design.

For instance, unmarking users through 'universally standardized' designs invisibilizes people who are to use these platforms and services from their context. If a woman who is a content creator, a Muslim, African and Black, must constantly consider social media hiatuses, censoring herself or limiting her creative expression to navigate the violence of online spaces, then creator gatherings teaching influencers to gain more engagement, traction and ultimately profit are counterproductive. First, the Muslim woman is unable to achieve the full potential of space because of her algorithmic biases, and her position in the matrix of domination. Second, most of these spaces remain inaccessible to women such as Layla and Latifa, who might benefit from the community yet have no idea such spaces exist in the first place. Collectively, these biases in access, product design, and technologically mediated harassment they face cause physical, emotional, and financial violence against Muslim women creators.

While Muslim women actively contribute to the influencer economy and platform capital, they remain one of the least protected or prioritized groups regarding algorithmic and product design. For example, TikTok is a particularly difficult space for Layla because using hashtags such as '#Blackcreator', or '#BlackHijabiCreator' gives her a lower reach and engagement rate. Yet, to accurately identify herself or her work, she indeed is a black hijabi creator. This phenomenon is what Wittkower and Costanza-chock describe as 'dysaffordances', where people have to misidentify themselves to fit within design standards. Layla also expressed concerns about the constantly changing algorithms that they are unable to keep up with, thus affecting how they work and what they ultimately achieve. Meanwhile, algorithmic and design updates are one of the main purposes of creator gatherings which women such as Layla do not have access to in the first place. Further, people who specialize in influencer and content creator management rightfully require fees for their services, bringing us back to our first argument that a person who is marginalized by the matrix of domination may not be able to afford or utilize such services in the first place.

4. Does Reporting Violence Work?

On 26 May 2022, I experienced what I believe was a coordinated attack by my Uber Eats delivery person and two men in a white car. This incident was avoidable on Uber's part since I had reported this specific delivery person not more than 30 minutes before Uber reassigned him to me when I placed another order. The

occurrence made me question what 'reporting' a case really does for the customer and what the redressal options are available for them and this time, Uber's failure to adequately address my issue led to me being attacked and robbed by this very person.

Why reference a food-delivery incident while speaking of moral surveillance, public spaces, and Muslim content creators? The incident allows us to question whether Uber has the systems and designs in place to identify that someone who has clearly reported a rider does not want further interactions with the said person. Through my conversation with Latifa who had experienced harassment online, she shared that 'reporting' accounts were rarely useful because the platforms could claim they do not see an issue. Both examples demonstrate that the failure of a designer to ground inclusive gender and safety within their product design and policies causes physical threats and violence against marginalized users. For instance, Muslim content creators should not have to endure disciplinary action through violence and comments that blame them, claiming they represent "bad Muslims".

Just as gig workers such as Uber delivery riders contribute to Uber's profit, content creators contribute to social media platform profits and reach. Hence, if Uber penalizes riders based on customer concerns and reports, in the long run they risk a loss since the fewer delivery people they contract, the fewer people they can serve. Likewise on social media, if safety, inclusion, and context are at the core of their design it will mean that it will inherently restrict interactions that harm others where harm will consider power dynamics that exist and not center the uncritical notion of 'freedom of expression' without considering power. In other words, given that online spaces mirror our offline social dynamics and Muslim women represent minoritized communities, their interests and concerns would be considered unprofitable. Similarly, Costanza-Chock argues that design decisions that use A/B testing will always privilege one group over the other, and these decisions often reinforce our biases.

Currently, many safety and security features are either too complex or overwhelming for users. Content creators are often required to turn on multiple security features such as limiting comments, direct messaging or content sharing which ultimately affects their engagement, interaction, and reach while doing very little to address the broader issues.

5. Design Justice: Grounding Technology in Social Context, Embodiment, and Place

The argument is not to demand that technology be used to solve social issues, but rather to point out that technology should not exacerbate these issues either. In this way, it is important to identify that one-size-fits-all solutions will always "erase difference and produce self-reinforcing spirals of exclusion". For Muslim content creators, the constant violence they face makes them understand that their visibility and use of online space is a source of frustration for male patriarchs. The paradoxical claims that their bodies should be something they should hide but at the same time be put on display, seek to control women's free use of public spaces, and pleasures within these spaces. The threats of being a visible woman lie at the foundation of the female fear factory which uses the language of violence, fear, and the promise of safety to control women and LGBTQ persons. In grounding technology in the social context, designers should "center the voices of those who are directly impacted by the outcomes of the design process". Outcomes here mean that the existing power dynamics must be evaluated to understand how design processes and decisions will affect specific communities; for instance, Muslim women content creators who are disabled. In addition, at the intersection of technology, gender, and religion, Muslim content creators addressed that some engagement requirements such as using trending music to drive metrics puts them in very controversial positions within their communities. A technology that is grounded in social contexts will not require people to use certain music or content to be successful in the space. Regarding biases, Layla would not have to misidentify through hashtags to gain more traction on her posts.

Social context, embodiment, and place also extend to content moderation not being solely automated. Users at the margins of the domination matrix often face forms of bigotry such as Islamophobia and homophobia. This means that a queer Muslim content creator may be reported by multiple users for expressing themselves within social media spaces or creating content. For instance, Muslim women content creators are de facto expected to be 'role models' of the community, so when they decide to either no longer wear the hijab, or be in solidarity with queer people, community members may consider these as harmful deviations that can 'negatively' impact their followers. In this case, morality is being centered yet again by the prison warden, and people who transgress set standards are punished using the very tools that are created to protect people online. Here,

the concept of protection ignores the existing power relations that are being reproduced through safety features. Without a contextual moderation policy, guidelines, and training, we find that people who are victims of microaggressions, violence, and bigotry are unjustly banned or suspended. This is how technology comes to reinforce discrimination by abstraction, where people face biases and discrimination because designs are abstracted from context. Within this argument, Costanza-Chock asks whether algorithmic designs should be structured using the logic of fairness, which they consider synonymous with color and gender blindness. Alternatively, they ask if "design should be according to racial, gender, and disability justice". This essay asks if design should also follow the logic of the religious context, geopolitical location, and language justice?

Grindr recently published a white paper on gender-inclusive content moderation that addresses most of what technology companies claim is 'difficult'. The paper shares Grindr's moderation process, but most notably they acknowledge that "careful product design sets the moderation team up for success". Their five-step moderation process provides a framework for how product designers can center gender justice. They suggest, for instance, separate moderation queues for trans and non-binary people because they frequently get discriminated against. Grindr's suggestion to separate moderation queues also requires additional and continuous training with content moderators to provide them with knowledge bases and context. What they achieve with this process addresses Costanza-Chock's question of blind 'fairness' and their concern for universalization. Here Grindr acknowledges through multiple scenarios how communities marginalized within the matrix of domination may experience aggression and harm through standardized design that does not consider the context of people, their location, culture, and the fluidity of gender. Regarding location and culture, Grindr's moderation white paper also provides a framework and reference to how gender and gender identities may differ in different areas and regions.

What Grindr achieves with their moderation policy and framework contributes to a foundational step into how feminist principles on social media governance can be successfully implemented. Feminist social media governance seeks to effectively ensure that technology does not further harm or discriminate against people. However, the Grindr example does not suggest they do not face challenges in achieving their goal because, as put by Gqola, patriarchy fights back. In the same sense, challenging the female fear factory is a dangerous job. As

Gqola explains, such feminist platform governance centers the multiple possible interactions among users and the fluid power relations that exist within our social engagements; it centers the voices of minoritized communities. Spaces such as creator workshops can provide an opportunity for community-led outcomes and engagement with content creators if they are accessible and do not focus only on maximizing profit and promoting entrepreneurial spirits.

Designing feminist online public spaces, and safer internets, as argued throughout this essay, is not independent of our offline histories, structures, and hegemonies. Most of the violent actions today could be viewed as an evolution of gendered historical trends of discipline and control. Public spaces have always been communicated as a domain where women should not freely exist; thus our online spaces and digital societies are not free of these structures either. Our digital lives are governed by our will to remain safe and practice security, external design and policy decisions, and sociocultural and locational contexts. For Muslim women creators who engage and attempt to exist outside of the norm, they are reminded through violence and the shaming of their supposed 'excessiveness'. At the same time, they experience design biases that put them at risk. Surveillance today involves multiple actors including us who participate in watching, controlling, punishing, sharing, and collecting information, making it much more complicated than Foucault's panopticon. Thus, in creating a feminist social media governance practice and framework, digital safety, security, and privacy must be considered as a social justice issue, which is a reiterative process of learning and reshaping our policies and designs to mirror dynamic social and power relationships. In addition, it is essential to acknowledge that to create our imagined feminist public space online, we must ultimately challenge patriarchal social and power structures.

"I Silenced Myself": Brazilian Female Activists on Online Gender-based Violence

YASMIN CURZI DE MENDONÇA[1]

Abstract

This essay aims to present the partial results of the author's doctoral research at the State University of Rio de Janeiro. It contains analyses of interviews with Brazilian activists, politicians, and journalists who have experienced situations of online gender-based violence (OGBV). The essay tries to analyze their strategies of resistance, the insufficiencies in the current Brazilian legal framework, and the obstacles social media platforms pose in properly addressing OGBV.

---

[1] Yasmin Curzi de Mendonça is a PhD candidate at the Institute of Social and Political Studies, Rio de Janeiro State University. She researches online gender-based violence against Brazilian women politicians, journalists, and activists. She also serves as a Coordinator of the Dynamic Coalition on Platform Responsibilities (DCPR) at the Internet Governance Forum of the United Nations.

1. Introduction

This essay aims to contribute to discussions around the phenomenon of online gender-based violence (OGBV) and its victims' resistance strategies based on data from in-depth interviews. It follows José Medina's theorization on the "epistemology of resistance", where he expands on the necessity of examining resistances from both analytic and normative lenses. Analytically, this study aims to shed light on the "epistemic aspects of oppression" regarding the lack of recognition of victims of OGBV by law enforcers. As argued below, the traditional approach to free speech, embedded into law enforcement, often neglects discourse as material violence. Normatively, the author tries to point out possibilities to address this epistemic oppression – i.e., the neglect of discursive violence – through the victims' standpoints and suggestions.

What provokes this inquiry is the offensive attacks in the online environment that disrupt the communicative possibilities of minorities who seek participation in the "networked public sphere".[2] How do online attacks impact victims' lives and digital participation? More importantly, how do survivors resist such attacks?

At the time of writing, the author conducted interviews with 12 women journalists, politicians, and activists who have faced online abuse. These categories were selected based on the following understanding:

(1) journalists, especially fact-checkers, are more exposed to the risk of hateful attacks and trolling because of the need to contact conservative and radicalized groups in the country as part of their work activities, as well as the need to maintain active profiles on social networks;[3]

(2) politically exposed women such as politicians, candidates, and activists challenge power structures and gender roles by actively participating in the public sphere that has historically been male-dominated, which also makes them preferred targets of discrimination and silencing attempts; and

---

[2] An updated understanding of the Habermasianian concept (where political participation and civil society organization mainly occur) for the digital age.

[3] This understanding stems from the author's observation of what is happening in the Brazilian context, and is also supported by data from a global survey of women journalists conducted by UNESCO. See, https://en.unesco.org/news/unescos-global-survey-online-violence-against-women-journalists.

(3) activists challenging the established norm of 'technosilencing'[4] by digital platforms face the 'risk of abuse'.

The essay also draws from feminist theories of technology which reject both technological determinism and gender essentialism when thinking about the role of technology in embedding gender-power relations, emphasizing that "the gender-technology relationship is fluid and situated", but also highlighting that "the materiality of technology affords or inhibits particular gender power relations". Given that the technology-related workforce has been historically dominated by men, especially in the higher echelons of companies, discourses on technological development have been framed by the hegemonic male standpoint. However, there is still space for resistance and contestation of gender imbalances, with the inclusion of women in the design and configuration of technologies and, more importantly, in the possibilities of women users to produce new socio-technical arrangements and change their social circumstances while using technologies.

Recently, IT for Change and InternetLab submitted a report on gender justice to the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression at the United Nations, where they stated that:

> The exclusion arising from misogyny on online platforms not only reflects the wide-ranging physical, psychological, and functional harms to individuals, but it also signifies the impoverishment and contamination of the public sphere and a systemic squelching of the aspirational ambitions of those silenced to belong and be heard.

The report gathers data from India and Brazil on cyberviolence against women journalists, politicians, activists, and academics, pointing out that due to stereotypes and gender roles, women still have a limited presence in the technology field. Fewer women than men are hired and promoted by technology companies and, as a consequence, very few women occupy leadership roles. The prevalence of the male gaze in these companies impacts how devices are

---

[4] Platforms can regulate users' speech and visibility through myriad techniques and features in their content moderation practices. As described by James Grimmelmann and Myers West, content moderation techniques involve, but are not limited to, the removal, filtering, labelling, and deprioritization (algorithmically or by human decisions) of content, shadow banning, and account suspensions. They are often opaque to bystanders and to users' themselves. In most cases, users lack appropriate information about content moderation applied to their accounts, and platforms don't have proper due process mechanisms for users, such as appeals.

produced and what kind of features are deemed the most relevant. For example, the 'people you may know' feature on Facebook has often led to stalking cases but this did not emerge as an issue for men who designed the feature. The study concludes that attacks have increased in recent years due to extremist ideologies and the amplification of online violence on social media networks. Such a scenario makes it necessary to understand how women survive and resist such attacks, using their agency to identify shared concerns and understand how platforms amplify hate speech at the cost of users' safety and wellbeing.

2. Online Gender-based Violence: A Typology

This essay draws from Michel Wieviorka's work, which acknowledges that:

> Violence can be physical, moral, or intellectual. It damages or aims to damage the integrity of those who are its victims, and most often, all of these constitute a whole: to injure, kill, rape, steal or attempt to do so; for example, it is also to bring into play meaning, significations, it is to weaken or destroy the "capabilities" […] which are concrete nor physical.

Wieviorka builds upon an interactionist approach to understand violence through the dynamics of interactions between subjects in specific situations. He conceptualizes violence as a constant state of "pure violence" that cannot be justified, differentiating it from conflict, which he understands as a resistance strategy against power asymmetries, for example, class struggle or the competition for dominance in a society.

Violence against women (VAW) is a form of collective violence – it consists of specific acts of violence that affect women as a group. Gender-based violence (GBV) refers to acts of violence against women and LGBTQ+ people – it targets their identities, affective orientation, and other socio-subjective elements linked to their subjectivities. As Judith Butler highlights, GBV often acts as a regulation mechanism in which the perpetrators' primary intent is to regulate bodies and ways of living that, in a manner, deviate from the male heterosexual status quo. According to her:

> The social punishments that follow upon transgressions of gender include the surgical correction of intersexed persons, the medical and psychiatric pathologization and criminalization in several countries, including the United States of "gender dysphoric" people, the harassment of gender-troubled persons on the street or in the workplace, employment discrimination, and violence.

Online gender-based violence (OGBV) is the range of behaviors in cyberspace that allow discrimination and threats against women and LGBTQ+ people's integrity and dignity. Offenses and attacks are also due to and based on an individual's gender identity and/or affective orientation. Literature often refers to online violence against women as online harassment, digital abuse, technology-facilitated gender-based violence, and online misogyny. In this essay, the author uses OGBV solely because it is a broader concept; it not only applies to acts of violence against women but those faced by LGBTQ+ people likewise profoundly affected by heteropatriarchal misogyny in the manosphere.

OGBV can manifest as (1) censorship, with individuals coercing people to delete their profiles/websites or pages, or engaging in coordinated flagging,[5] so that platforms take down their content and/or suspend them; (2) mob attacks, defamation, injury, coordinated attacks, sometimes by fake profiles and bots; (3) hate speech, with widespread misogyny, trans and LGBTphobia, and racism; (4) threats of physical violence and death; (5) stalking; (6) doxxing; (7) unconsented use of personal images; (8) unconsented use and dissemination of intimate images; and (9) extortion.

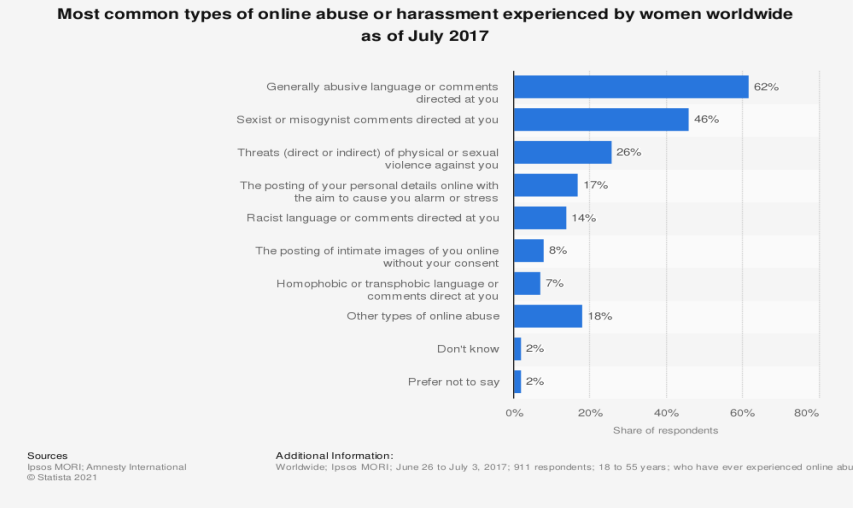
According to data from Statista, globally, the most prevalent form of online abuse or harassment directed at women is the use of abusive language followed by sexist/misogynist comments, direct or indirect threats of physical or sexual violence, exposure of personal data, racist language, circulation of unconsented intimate images, and homophobic or transphobic language.

The internet allows for several possible tactics to affect online communications. Groups of masculinists use several tactics against women online, such as, (1) coordinated attacks by various profiles on the same person; (2) trolling on posts; (3) dogpiling, in which many users post messages to the victim, not necessarily with offenses, but to make her feel vulnerable on the network; (4) dog-whistling, which involves the use of messages that do not necessarily ring violent to spectators, but represent something to the community; (5) doxxing; (6) ideological falsehood, meant to gain the victim's trust; (7) multiple SNSs, that is, attacks on multiple social networks at the same time; among others.

---

[5] The concept of coordinated flagging is better explored in Belli, L., Zingales, N., & Curzi, Y. (2021). *Glossary of platform law and policy terms*. Internet Governance Forum. https://www.intgovforum.org/en/filedepot_download/45/20436

Figure 1: Most common types of online abuse or harassment experienced by women worldwide as of July 2017



Source: Statista

These tactics are also continuously evolving. Unfortunately, legislation to address specific behaviors often does not go hand-in-hand with technological developments, enabling multiple possibilities for violent acts. In Brazil, for instance, Article 21 of the Civil Framework of the Internet (Bill 12.965/2014) was quite innovative in allowing victims of non-consensual intimate imagery dissemination – a practice commonly known as "revenge porn"[6] to ask for the removal of the content from platforms only with the notification of the victim or their legal representative (notice-and-takedown), not requiring a court order as is the general rule of Article 19 (judicial notice-and-takedown).

However, technologies have evolved to enable users to create deep nudes – false images produced from deep machine learning – which broadens the scope for other understandings about victims' image rights. These evolving tactics coupled with the inability of legislation to encompass such violence explain why victims often cannot get proper reparation in the judicial system. Of the seven interviewees who faced various forms of online attacks, five stated that they decided not to take any legal action against their attackers because they did not know how to do so, or whether they had any legal rights.

[6] The term "revenge porn" is highly contested in feminist debates. A proper reconceptualization is "non-consensual intimate images".

Online violence acquires new forms with the development of new technologies. The forced digitalization due to the Covid-19 pandemic, for example, made room for zoombombing[7] to become more frequent. El País reported[8] that zoombombing is often used by masculinist groups to disrupt and silence women's conversations at events linked to gender issues. The author has been a victim of zoombombing at the Dynamic Coalition on Platform Responsibilities, a coalition coordinated at the Internet Governance Forum of the United Nations in 2021, due to a security breach. Trolls invaded and "took control" of the room, starting to "bomb" the session with extreme pornographic and violent content, Nazi propaganda, and other harmful content through video and audio.

The virtual environment is full of features that facilitate the spread of violence: anonymity, for example, may enable harassers to hide their identities and makes it more difficult for victims to search for legal remedies.[9] Other elements may facilitate the ease and speed with which illegal or outrageous content is disseminated, including by 'non-human' actors, such as bots.[10] A relevant characteristic of the internet that constantly refreshes violent experiences is the 'permanence' of a post, which can be continuously revisited, reposted, and shared without temporal delimitation. Another is the possibility of coordinated attacks by 'online mobs' that can "easily overwhelm a single victim".

Online violence can also be directed at a specific victim or toward a broader community that shares the identity traits such as sexual orientation, gender, race, and ethnicity. Studies have pointed out that group-based harassment has wider consequences, in turn targeting individuals for their core identity. In this sense, for members of a given group who share similar characteristics, the experiences of others in the group influence their perceptions of risk and safety. Empirical research has also shown how women are constantly afraid of suffering online violence, even though they were not, in their personal experiences, victims of any direct violence, which generates a sort of chilling effect on their online participation.

---

[7] Practice constituted by the invasion of Zoom sessions by accounts.

[8] According to El País, attacks on Zoom happened more often in events that discussed gender issues.

[9] Also, according to Mary Ann Franks, "Burner e-mail accounts and anonymizing software such as Tor can prevent a victim from discovering a user's true IP address or other identifying information".

[10] The definition of 'Bots' is more explored in: Belli, L., Zingales, N., & Curzi, Y. (2021). *Glossary of platform law and policy terms*. Internet Governance Forum. https://www.intgovforum.org/en/filedepot_download/45/20436

Traditionally, chilling effects occur when an unjustified restriction of a licit speech by a rule or a measure taken by a private or public party causes a ripple effect, inhibiting other actors from proffering similar speech. The legal uncertainty is what would make individuals avoid expressing themselves. Concerns about chilling effects constitute one of the three false premises[11] about free speech that lead toward "an absolutist approach to the First Amendment". To Mary Ann Franks, this approach makes freedom of speech, including speech that violates others' dignity and psychophysical integrity, disproportionately more protected than other rights. In traditional comprehension, "regulations of unprotected speech are dangerous because of their potential to discourage protected speech".

Despite the appeal of such an argument, Franks, drawing upon Leslie Kendrick's work, notes that the US Supreme Court has found the chilling effect on nothing more than "empirical guesswork". Nevertheless, the lack of evidence on this approach has not affected the extent to which it influences law and policy decisions. Public actors have insisted on only considering the abstract possibility that people may be silenced or inhibited from speaking if any restrictive measure comes into effect. Franks calls attention to the fact that the traditional chilling effects approach neglects the "chilling of women and non-white men's speech" since "harassment, threats, genocidal rhetoric, hate speech, doxing, and revenge porn all have silencing effects". On the one hand, the absolutist approach to free speech does not rely on empirical research or data but keeps informing the mentalities of policymakers, judges, and law enforcers. On the other hand, several studies have empirically stated how such harms may impact the free speech, mobility, and freedom of association of minorities.

On street harassment, for example, Robin L. West states that it is:

> …the earliest – and therefore the *defining* – lesson in the source of a girl's disempowerment. If they have not learned it anywhere else, street hassling teaches girls that their sexuality implies vulnerability. It is damaging to be pointed at, jeered at, and laughed at for one's sexuality, and it is infantilizing to know you have to take it.

---

[11] In Kendrick's words, "The danger of chilling effects, the merits of the marketplace of ideas, and the importance of protecting freedom for the speech we hate, are settled tenets of First Amendment orthodoxy."

According to studies, in their everyday lives, victims typically respond to street harassment by changing their schedules for going out, ceasing to go out at night alone or taking specific pathways, and even thinking multiple times about their clothing choice. Yet, despite that, free speech absolutism undermines the negative impact verbal attacks have on minorities, often leading to devastating restraints on their freedom.

Drawing parallels with street harassment, the following sections explore how online attacks impact women's online activities and highlight the need to address the issue adequately. More importantly, they aim to show that, in the absence of the lack of proper support from institutions, women are developing their own strategies of resistance.

3. "My Choice was to Silence Myself": Online Gender-based Violence and Impacts on Journalistic Activity

According to a UNESCO report, women, and especially journalists, have undergone targeted attacks on online platforms, affecting their participation in these spaces. Converging with such data, the Brazilian Association of Investigative Journalism (ABRAJI) stated that, in Brazil, women journalists were the group most susceptible to attacks in digital environments in 2020 – out of 72 reports of digital attacks against journalists, 40 targeted women, constituting 56.76% of the total.

Thaís, a former fact-checker interviewed for this research, told the author that traumatic events related to her fact-checking job caused her to abandon her career. She is still a journalist but currently works in a feminist organization and does not want to go back to fact-checking. In her account, Thaís mentioned two experiences of coordinated attacks. The first, led by the Brazilian right-wing organization Movimento Brasil Livre, involved the creation of a dossier on hundreds of fact-checkers, exposing their personal data and images.

The second stars the Terça Livre channel which has recently been banned from YouTube for participation in anti-democratic demonstrations and attacks against public institutions following a Supreme Court investigation. After Thaís contacted the organization chief, Allan dos Santos, he ironically insulted her, and shared her email with thousands of subscribers of the channel on the private messaging application, Telegram. The subscribers subsequently started a mob attack against her.

> (...) In 2020, I was still with the fact-checking agency, and we were going to conduct fact-checking on YouTube. We were looking at the videos of some channels and

> seeing what kind of false information existed about the pandemic. Moreover, as usual in journalistic work, we ask for the "other side" to verify the story. If I am saying that you are spreading false information in your video, I will ask for your version. I will come and say, "oh, I am doing this story; I identified that there is this false information in your content; how do you position yourself concerning it?". At the time, I sent this email to Alan from Terça Livre, asking if he wanted to state his position and so on, and then he answered, "shove it in your ass". He has a Telegram group with his supporters. Then he shared a print of the email that I had sent, with my email address, which at the time was my working address (thank God that it was the professional account), to make a joke, as if to say, "look what I answered this reporter". Then came more emails from his supporters insulting me.

The activity of fact-checking, thus, appears to make journalists more vulnerable than others to possible attacks since they must reach individuals and groups that are often linked to radicalized or controversial activities in order to verify a story. After the episode with Terça Livre, Thaís abandoned her fact-checking career for fear of more such attacks.

Online harassment can be understood as a siege tactic. It takes place on multiple networks to silence speech seen as 'inimical'. Juliana dal Piva, another journalist, reported being the target of attacks after publishing a story about a film produced by the production company Brasil Paralelo. According to her:

> I do not remember when this was; I think I saw this movie on a Wednesday; the story came out on Thursday, and they started cursing me on Thursday morning, and then it was about two or three days of a very high amount of cursing and offenses. Thousands of name-calling and swearing.

Due to the lack of appropriate mechanisms for reporting, filtering, and moderating offensive content on platforms, the journalist's option was to silence herself immediately. As she stated:

> That day I discovered several things: silence a conversation where people are only swearing at you without constructive criticism, etc. You do not know which account [handle] is a robot and which is not, on most occasions. In addition, I also learned to mute notifications of accounts that do not have verified email or phone, which decreases interactions with robots and fake accounts.

However, silencing notifications was not sufficient to stop the offenses. On the contrary, the ecosystem of violence on the journalist's social media platforms

kept increasing. According to her, several overwhelming attacks, including death threats, were made in a short period. The journalist contacted the legal department of her newspaper agency for help, but the company did not take any action.

> When the death threat came, I informed my supervisor, who finally communicated to the legal team to see if they would do anything. They did not do anything. I regret it. I think I should have insisted more that they do something. I became fragile after that because I was in a situation where I did not feel comfortable demanding that something be done.

At least three interviewees reported the lack of support to journalists by their respective agencies. In the absence of any assistance, many did not know what they could do. Many interviewees (not only journalists) also stated that they did not trust the judiciary to achieve reparation, underscoring how women still lack recognition in Brazilian society.[12]

Other interviewees stated they did not know exactly how to sue their attackers and fulfill their rights. Some also noted that they did not feel confident in the outcome of the process. In her account, dal Piva said that, despite the result, victims must go to the judiciary to try and receive reparations.

> I tend to think that this is an obligation… to sue. I do not know if the judiciary will always answer the way we want, the judiciary is the judiciary, right, but I think we must sue. Furthermore, I believe that experiences are generally fruitful, especially in easy cases, when it is an injury or a blatant slander. I think that for the scenario to change, it is necessary "to make hurt in the pocket", to cause an embarrassment for the person, to force them to make a public apology for what they did (...).

She also stated that a positive decision by the judiciary on behalf of the victims might usher in cultural change, leading to a decrease in the incidence of such violent acts. More importantly, it can give the victims a new perspective on the validity of their demands. In her words:

> It is good, too, because you sometimes get so many negative comments that it forces you to think you are somewhat wrong. Even if you are not, it is kind of crazy. It keeps going through your head.

---

[12] In his theory of recognition (*Anerkennungstheorie)*, Axel Honneth (2015) nominates legal recognition when the law and law enforcers recognize offensive conducts as a violation of dignity by enabling reparations. This also affects the victims' self-esteem, their sense of worth, and of belonging to a community.

According to Thaís, lack of confidence in the judiciary can lead to situations of re-victimization – which can manifest in new aggressions by attackers due to the victim's attempts at seeking judicial reparation and even at the hands of the law enforcers who are not sensitized enough to address violence against women. She also chose not to file a lawsuit against Alan dos Santos to stave off further attacks in retaliation from his supporters.

4. "I had to Restrict Myself": Online Violence against Women in Politics

The need to look more closely at gender-based political violence has caused the Organization of American States (OAS) to write the 'Declaration on Violence and Political Harassment Against Women'. The document calls for "a definition of political violence and harassment against women, considering regional and international discussions on the subject". It also states that "both violence and political harassment against women may include any gender-based action, conduct or omission; and may affect individuals or groups". Furthermore, such violence "has as its objective or result to belittle, annul, prevent, hinder, or restrict women's political rights, or rights to have a life free of violence, and the right to participate in political and public affairs on equal terms to men".

As mentioned earlier, while we can catalog some of the strategies used to spread offensive behavior, it is vital to keep in mind that the tactics used to carry out attacks are constantly updated. As Karla Coser, the young councilwoman of the Workers' Party of Espírito Santo, (PT-ES) observed in her interview:

> (…) the right wing is already more innovative [than the left wing] in this area. They do not bring the debate to our profiles to not generate engagement [for us], but they cut out our lines and publish them in their profiles, and there they comment and criticize us.

This strategy can reduce the ability of victims to engage in dialogue and defend themselves, however, this paradoxically also increases risks stemming from excessive exposure and participation in social networks. Another concern among candidates is the possibility of gendered disinformation. For example, Coser had to rethink her posts and appearance on social networks. As she reports:

> (…) during the pre-campaign, I polished myself, which was something I did not like to do. I have always been a cherished person, and I did not like it, anyway; I posted a bikini photo, and I did all these things. Moreover, throughout the

> pre-campaign, we talked about the concern with montages of fake news...I always shared aspects of my personal life on social networks…but the fake news industry and all this logic of being a person who has several targets on the back –daughter of the former mayor, married, and young - this made me stop and think…I decided that I would need to do some things to protect myself…I had to change my way a little because I am a very transparent person, very natural, but to preserve and preserve my family's life, I had to be careful about what to share on social networks. So, I started to rethink some things with my staff, fearing how malicious people could transform the content on the networks. Not out of shame, in any way, but out of concern of photo manipulation, etc.

The manipulation of images, whether from deep fakes or montages, becomes gender-based violence when it is triggered to expose a woman morally. As a result, the concept of 'gendered disinformation' has received attention from academics and activists to designate these practices that aim to damage the public image of politically engaged women.

Women in politics are disproportionately the targets of disinformation campaigns, with a gendered element. Often, they are subjected to humiliating and sexually charged images. The purpose of these attacks, in general, is to make them feel untrustworthy, unintelligent, or simply uncomfortable and, thereby, shame them. The narratives are profoundly sexist and seek to distort public understanding of the concerned woman's performance. A known case is of Manuela D'Ávila who has constantly been subject to altered photos that make her look like a drug user, among other things. Images are often hypersexualized to discredit these women's reputations within the conservative imagination.

The possibility of such alterations leads women politicians, and those preparing for their candidatures, to make changes in their private social media, something men rarely have to resort to. The United Nations resolution 23/2 2013 entitled 'The Role of Freedom of Expression and Opinion in Women's Empowerment' highlights the role of states in banning "discrimination, intimidation, harassment, and violence, also in public spaces, that women and girls face". Nevertheless, in Brazil, little progress has been made in creating appropriate measures for addressing gendered disinformation and offenses against women in politics, leaving victims to deal with the risks alone. Their ability to participate in the digital public sphere is thus already limited from the moment they consider entering it.

5. "Platforms are against Us": Private Censorship against Activists

Lola Aronovich, a professor at the Federal University of Ceará, is one of the pioneers of cyberfeminism on the Brazilian internet, and runs the blog, 'Escreva Lola, Escreva', launched in 2008.

> When I started my blog in January 2008, I discovered a level of misogyny that I did not know existes. Happily, in our personal lives, we mostly choose the people that we talk to. Therefore, I would not have to talk to a guy that calls women whores. I did not know the level of misogyny that I came to know online. My first indirect contact with these masculinists, whom I nicknamed 'mascús' [man-assess], was in 2008 regarding the case of Eloá. I wrote four posts about it on my blog, and one of them was about a community at Orkut that someone told me about – at the time, Orkut was the most used social network in Brazil – titled 'Eloá virou presunto' [Eloá became ham].[13] Furthermore, it was signed by the 'Supreme Order of the Men of Good.' They were the 'Homini Sanctus'. I didn't even know them at the time. I had no idea.

Aronovich became the main target of masculinists for reporting such actions on her blog. Marcelo Valle Silveira Mello, (who used pseudonyms such as Psycl0ne, Psy, and Ash), one of the arrested men who attacked her in response to her blog, was behind several masculinist groups. Silveira was also the first Brazilian internet user convicted of racism on the internet for posts on Orkut. He was prosecuted by the Public Prosecutor's Office in 2005 and sentenced to one year and two months in prison. However, Silveira's lawyers claimed "insanity" and converted his sentence into a right-restricting penalty and probation.

After this, he continued his activities on Orkut. Silveira founded the page 'Men of Good', administering several communities in the social network. Those communities often contained racist, misogynist, and homophobic content, and incited violence through slogans such as, "Haiti, we are hoping to reach 1 million dead", "The woman is inferior to the man", "Cats + Gasoline = Bonfire", "I kill children for macumba", "Teenage girls are all whores", and "Eloá turned into ham".

[13] The latter was a masculinist "tribute" to the murderer of Eloá, her ex-boyfriend, Lindemberg Alves Fernandes, who kept her kidnapped for more than three days. The case received enormous attention from the media with live coverage on national television. In the Orkut community, posts declared support for Fernandes for having put an end "to that whore".

According to information from the Federal Police, Silveira was the founder of another masculinist and neo-Nazi group on Orkut called 'Homem Sanctus', which had a linked blog called 'Homini Sanctus'. The group presented itself, in the description on its website, as:

> We are the Holy Men, we follow the National Socialist ideology of the Third Reich, and we believe in Pagan-Christianity. In addition, we have sympathy for Islam and Eastern Philosophy. We only accept Men, preferably over 18 years, who know how to take risks and remain anonymous, and white Caucasians of the Aryan race. We fight all the garbage coming from the New World Order, among them Leftists (Anarchists, Communists and Social Democrats, Feminists, LGBT, and Blacks) as well as Right-goers (Liberals, Neo-Conservatives, and Anarchocapitalists) both manipulated by the Jews-Masons (Illuminati). We are favorable to survival, and military techniques aimed at our self-defense. In addition, we have Martyrs Wellington Menezes, Anders Breivik, Ted Bundy, Elliot Rodger, Adolf Hitler, George Sodini, and Charles Manson.

As various scholars, including Whitney Phillips, Angela Nagle, Bharath Ganesh point out, conspiracy theories, such as narratives that institutions hide truths about society, and that masculinity and the "legacy of the West" would be under attack (replacement theory), are commonly found in masculinist groups on the internet. Groups of masculinists (referred to as androsphere or manosphere) are often composed of a 'coalition' of men's rights activists, bloggers who call themselves experts in attracting the 'opposite sex', gamers, and other members of nerd subcultures, as well as neo-Nazis, Islamophobics, libertarians, anarchocapitalists, conservatives, and nationalists, who also profess eugenicist perspectives concerning racial supremacy, such as the aforementioned replacement theory. Analyzing the involvement of these groups in online attacks, Ganesh categorizes them under the umbrella concept of online hate culture whose common denominator would be the ideology of the Red Pill. As noted by Debbie Ging:

> Central to the politics of the manosphere is the concept of the Red Pill, an analogy which derives from the 1999 film 'The Matrix', in which Neo must choose to take one of the two pills. Taking the Blue Pill means switching off and living a life of delusion; taking the Red Pill means becoming enlightened to the world's ugly truths. The Red Pill philosophy purports to awaken men to feminism's misandry and brainwashing.

In the same way, many of the activities led by groups of masculinists on the Brazilian internet have the *idée fixe* that the expansion of the rights of women, black people, and LGBT people is an attempt to subordinate men.

Another part of *the modus operandi* that shapes the performance of these groups is trolling.[14] Trolling can therefore provoke feelings of disgust with, for example, images and videos with obscene and disgusting content – extreme pornography, mutilation and excretion – or leave the interlocutor exposed and uncomfortable.

Another community managed by the profile 'Homini Sanctus' on Orkut was 'Eu ri do massacre de Realengo' ('I laughed at the mass shooting in Realengo'), in reference to the attack perpetrated by Wellington Menezes de Oliveira at the Tasso da Silveira public school in 2011 at Realengo – a neighborhood of the West Zone of Rio de Janeiro. Oliveira murdered 10 girls and one boy at the school. In his suicide note, he claimed to be a virgin. These facts supported the thesis that the massacre was motivated by misogyny, with Oliveira described as an 'incel' – an abbreviation for 'involuntary celibates'.

Figure 2: Image of the 'I laughed at the massacre of Realengo' page on Orkut

Source: Wikinet

The Federal Police, responsible for investigating terrorist acts, linked the incident to Marcelo Silveira and Emerson Santos in 'Operation Intolerance'. Oliveira would have committed acts influenced by the 'Homini Sanctus' group. Silveira and Santos

[14] The etymology of online trolling can be traced back to mythology, with trolls being the grotesque mythological creatures, or in reference to fishing, as trolling is tying a bait behind a fishing boat.

were charged and arrested for incitement of violence on the internet after more than 70,000 complaints regarding their blog and another site administrated by them, called 'Silvio Koerich' – a name stolen from another anti-feminist blog from which they were both banned in 2011. Silvio Koerich (the original) promoted himself in the networks with the narrative that he would "unmask feminist women" and teach men to become more "manly". The eponymous blog Silveira and Santos created in revenge posted content containing zoophilia, necrophilia, and incitement to rape, among others. The appeal court reversed the conviction and granted a pardon the following year.

In 2013, after leaving prison, Silveira created 'Dogolachan'. Despite the two previous convictions, Silveira continued spreading racist, misogynistic, and homophobic hate speech and disseminating online violence along with several other trolls. Five years later, in 2018, he was finally arrested by the Federal Police under 'Operation Bravata', and sentenced to 41 years in prison for "criminal association, dissemination of images of pedophilia, racism, coercion, incitement to commit crimes – such as rape and feminicide – and terrorism committed on the Internet". His arrest, however, did not discourage 'Dogolachan' members.

Soon after Silveira's arrest, for example, one of the moderators, André Luiz Garcia (29), known as 'Kyo, El Fuego Sanctus', shot a 27-year-old girl and then committed suicide in the interior of São Paulo after she denied his advances. In a post on the forum, Garcia warned that he would kill himself and do it by taking other people with him. He also left a message wishing Silveira well when he escaped prison.

In 2019, the forum was also linked to the killers of the Suzano mass shooting, where five boys aged 15-16, pedagogical coordinator Marilene Ferreira, and school employee Eliana Regina de Oliveira were murdered by two former students at the state school Raul Brazil, at Suzano, in the interior of São Paulo. The killers, Guilherme Taucci Monteiro (17) and Luiz Henrique de Castro (25), were both members of Dogolachan.

According to a report by Vice, the act was organized with the help of the forum and disseminated by the actors. After the attack, they posted the song 'Pumped up Kicks' by the band 'Foster the People' whose lyrics talk about massacres in schools in the United States from the perspective of a perpetrator. Forum participants criticized the act for not being significant in terms of the number of dead and injured.

Dogolachan remains active – at least until the time of writing this essay. When accessing their page, there is a link to their Discord channel – a voice social network where content moderation is poorly applied. The website can be accessed easily by searching on any service provider.

When the author last checked the website, the invitation page had the message "Guilherme Alves – national hero", referring to the killer of Sol, a young woman who was a gaming streamer. Alves murdered Sol on 22 February, 2021, because he saw her as an "enemy". The murder was disclosed in an email from him to Lola Aronovich right before the incident. He said he had counted on the 'Crazy Chan' support to prepare the act. Unfortunately, in disclosing the incident publicly, Aronovich faced, once again, several death threats by trolls and other masculinists on the internet.

In Brazil, feminicide (also known as femicide) includes murder resulting from domestic violence and discrimination on account of being a woman. However, even though Sol's murder was linked to masculinist forums, the Public Prosecutor's Office of São Paulo decided to remove the characterization of the crime as feminicide. In several cases, the law has only been applied in cases when there is a marital relationship or a previous relationship between the victim and the offending person.

The media's unwillingness in reporting such crimes as acts that reflect a supremacist ideology has been severely criticized by Aronovich. It happened in the Realengo and Suzano mass shootings – both were portrayed as arising from bullying and mental health problems. This hinders the proper diagnosis of such events and, consequently, the development of efficient actions to deal with the problem.

Aronovich's efforts in disclosing one of the most extensive masculinist networks on the Brazilian internet led her to enter the Special Protection Program for Human Defenders of the Federal Police. In her honor, the deputy Luizianne Lins from the Worker's Party drafted the Bill 13.642, known as 'Lei Lola', which was passed in 2018. The law authorizes the Federal Police to investigate online crimes against women. Prior to the Lei Lola law, the Federal Police could only investigate specific crimes stated on Bill 10.446/2002. However, although Lei Lola represented significant progress, Aronovich also stated in her interview that "little can actually be done because there is a lack of preparation regarding gender issues in such institutions", specifically referring to the constant neglect of victims' reports of violence.

Needless to say, the prominence of Aronovich's work was not enough to protect her from platform censorship. In 2017, masculinists, using coordinated flagging, took down her blog that was hosted on a Google domain. After a week without her blog, several attempts to contact Google, and a massive campaign, "Google, stop censoring Lola!", a representative of Google finally restored her access to her account. Over the phone, this representative rejected Aronovich's demand for a public apology. She was also not allowed to monetize the content of her blog. "Platforms are openly against us. They don't like feminists. They don't like activists. They are at the disposal of those who disseminate hate. No platform helped me in any case," she said.

Twitter also took too long to verify her account, which made it possible for masculinists to keep creating several fake accounts in her name, manipulating posts to degrade her image on other social media platforms.

> In 2015, I traveled to a place where my cell phone was out of service, and I didn't have a notebook. And when I arrived home, my mother was horrified, asking me, "what did you post online because there are thousands of people swearing at you". What happened was that they took advantage of my absence to create a post on Twitter with my name and photo. First, they posted it on Facebook on Dogolachan's racist and misogynist page. The tweet celebrated the death of Alckmin's son,[15] saying "what a shame that the father didn't die too".

Twitter took too long to remove her main aggressor from their platform. The several reports that Aronovich made about harmful content and harassment on the platform did not pass through the automatic moderation, which according to her always stated, "We have reviewed your report carefully and found that there was no violation of the @Twitter Rules against abusive behavior." Aronovich further said, "I like Twitter, but they only removed my main aggressor after I went to their headquarters. I reported him for over a year and a half!"

Despite becoming a target, Aronovich never stopped her activities of reporting masculinists groups, telling the victims' stories, giving classes and lectures about feminism, and posting on her blog. She says, "The people who are attacking you count on your silence. Exposing myself is a strategy for defending myself."

---

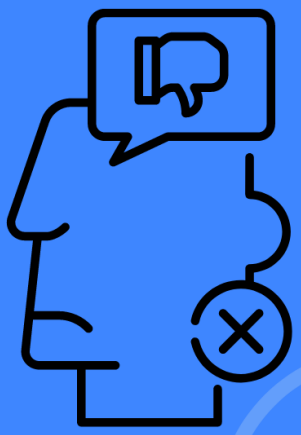
[15] Geraldo Alckmin is a well known Brazilian politician and former governor of the state of São Paulo. He belonged to the center-right party PSDB (Partido da Social Democracia Brasileira) at the time.

6. Final Considerations

OGBV in Brazil continues to be of low relevance for public institutions. The neglect of such issues may come from (1) widespread discrediting of violence against women, even in cases involving physical aggression; (2) the lack of institutional capacity in understanding the functioning of the online environment, which often makes them look at online violence as less damaging than offline violence.

On the first hypothesis, it is necessary to point out that there are already several relevant studies on gender-based violence and the difficulties in receiving adequate redress from public institutions. Nevertheless, it is still necessary to advance the research on how platforms are receiving victims' reports and what they do about them. Concerning the second, little attention is given to how regulators, judges, and legislators understand the debate on freedom of expression on the internet.

Based on the victims' impressions on the subject, it becomes clear that platforms, punitive institutions, and employers often do not take their claims seriously. The interviewees frequently stated feelings of abandonment and having to deal with cases alone. Without support, they internalize the costs of online violence. Not having anyone to count on, they often silence themselves. Nevertheless, they also often try to resist by changing their own activities to reach new goals or by continuing to talk loudly. In this regard, support from their families and workplaces is a crucial element so that they feel more secure to keep their voices resonating.



II

Contextual Content Governance

Gendered Hate in a Video World: An Analysis of Gender-based Harassment and Community Guidelines for Short-form Video

DIVYANSHA SEHGAL[1]

Abstract

Online gender-based violence takes many forms: stalking, harassment, threats of harm, and non-consensual sharing of personal information to name a few. This problem is not new. As early as 1993, Dibbel's 'A Rape in Cyberspace' described gender-based harassment and harm in a text-based virtual community. This is a problem that will follow the technology and the people using it wherever it goes next. We are already seeing different versions of this as multiple women are coming forward with their experiences of being harassed on virtual reality platforms. If companies want to keep their platforms safe for users of all gender

[1] Divyansha Sehgal is a technology researcher at The Centre for Internet and Society, India, primarily studying platforms and their politics. Her research has focussed on issues of intermediary liability, online misinformation, digital infrastructure, and platformized labor. She is an ONI-IIC Young Leaders in Technology Policy fellow and an engineer by training.

identities as most of them purport to do, they need to take a stand against gender-based violence in its various changing forms.

Community guidelines are official documents that platforms put out to establish ground rules for users, lay out acceptable uses of their technology and outline the recourse available to users when these guidelines have been violated. It is the one document that should ideally present a blueprint for interactions on and uses of, the platform. Thus, studying these guidelines is a great opportunity to understand what the platform considers to be issues and behaviors it needs to safeguard its user against, and how it specifically understands gender-based violence. The essay aims to cover the context and effects of online gender-based hate speech on users, and why community guidelines are an effective tool in the arsenal in the battle against it. It shares observations from studying community guidelines for short video-sharing apps in India (Moj, MX Takatak, etc.), and explores whether and how these apps incorporate local contexts, and what the sector can do better to make the platform safer for all users.

1. Introduction

If you were online in April/May 2022, it is highly unlikely that you would have missed the Internet's obsession with the proceedings of Johnny Depp's defamation case against Amber Heard for an op-ed she wrote in *The Washington Post*.

The details of the case are not important for this essay but are thus: Johnny Depp sued Amber Heard for defamation for an op-ed in *The Washington Post* where she referred to herself as a public figure who is a sexual abuse survivor. The proceedings of the trial became a social media phenomenon, being filtered across all of the major platforms, and eventually the mainstream press. TikTok, however, played an outsized role due to its short-video format which allows users to quickly edit and re-upload recorded videos; a black-boxed algorithm that can sense trends and rewards their propagation; and creators who have become adept at spotting and jumping on bandwagons of algorithmic virality to expand their audience.

The TikTok coverage of the trial was predominantly anti-Heard and was marked by a kind of vitriolic hate that is seldom seen on this scale. Creators manipulated trial footage to create meanspirited reaction GIFs, psychoanalyzed body language to discredit Heard, created videos mocking her, and generally encouraged audiences to participate in extremely gendered abuse. Some creators even pivoted to

covering the trial only to expand their audience by humiliating Heard. Depp, on the other hand, was seen as a charming fan favorite who could do no wrong.

Though this level of abuse and vitriol might be new, it is unhelpful to act surprised every time the digital sphere finds new ways of persecuting women for simply existing online. As early as 1993, Julian Dibbel's article, 'Rape in Cyberspace' for the *Village Voice* described gender-based harassment in a text-based virtual community. The perpetrator ran a script that allowed him to control the victims' avatars as he pleased, while their users watched helplessly, unable to access their own characters. This case study is now a touchstone for those studying gender-based harassment online because it shows how, even in a fully text-based system, without the frills of modern pictures, emojis, video or any identifying features of individuals beyond their perceived gender identity, harassment can and will take place.

This pattern has been repeated over the years as digital communication has matured. Women and other gender minorities are at a much higher risk of being targeted. Women have reported receiving a much higher rate of misogynistic abuse that features criticisms that seldom have anything to do with the content of the arguments made. Research has shown that this is a trend repeated across the world and across industries; gender-based violence is global as well as endemic.

Clearly, this is a problem that will follow the technology wherever it goes next unless extremely proactive measures are taken. Multiple women have come forward with their experiences of being harassed on virtual reality platforms. Meta has had to institute a personal distance field by default for its VR platform because its users persistently grope and physically harass people who chose to present as women. In most of the reported cases, female journalists and researchers have had to abruptly quit the VR experience to stop the harassment.

Unless a community actively invests in developing safety tools and protections against existing patriarchal systems of power, minority members will always be at risk of harassment, exclusion, and self-censorship. We need to understand how platforms work and how the communities they foster interact with their design and audience mechanics to encourage gendered harassment.

The online abuse of Amber Heard can be seen as a veritable cocktail of the various ways in which abuse shows up for outspoken women online – as a team sport with a sense of belonging for those who participate in it, driven by a network of

loosely coordinated actors, and algorithmically amplified by platform affordances that aim to capture attention above all. The rest of this essay shall explore in more detail these characteristics of online gender-based violence, which are pushed into overdrive due to easy video editing features and amplified by personalized algorithms. The essay then discusses if newer short-form video platforms in India are taking these threats more seriously through an analysis of their content guidelines.

2. Non-consensual Pornography

One of the most obvious ways in which images and videos are unique in their potential to cause harm is through the sharing of private content. Feminist activists, scholars, and civil society members recognize the non-consensual sharing of personal videos as a form of domestic abuse and intimate partner violence. Abusive partners or exes, blackmail and threaten victims to exert their control over them.

When shared publicly without consent, the images and videos created in the trust and intimacy of a relationship can have extremely adverse effects on the victim's social well-being and career prospects. The mental health toll and the social stigma experienced by victims are also extremely significant, as victims can get cut off from school, jobs, and/or other support groups, and engage in suicidal ideation. The pandemic has only made the problem worse by forcing more people into situations that they are ill-equipped to escape.

These waters are further muddied by the easy access to artificial intelligence tools that help create and manipulate media, which has given rise to deepfake non-consensual pornography. Images can be scraped from public profiles and combined with existing databases of pornography to create pictures/videos of victims. The intent here is the same as that of a hack photoshop job—to embarrass and humiliate the target. By forcing women to police their behavior or exit the online sphere, abusers rely on the popular patriarchal notions that blame victims for the actions of their abusers and enforce disparate moralities on different genders.

3. Harassment via Loosely Connected Networks

Further, the abuse and harassment of victims are amplified by the affordances of social media networks. During GamerGate, a network of abusers rallied behind the hashtag #gamergate to send death threats, rape threats, stalk and dox feminist media critic Anita Sarkeesian and game developer Zoe Quinn among other women associated with the gaming industry. Both women had their addresses leaked and their physical safety threatened. They had to make changes to when and how they appeared at public events because of threats from attackers. These abusers organized their attacks across platforms, moving from asynchronous communications of Twitter to the real-time affordances of IRC platforms like 4chan to organize attacks against their targets.

Communications researchers have termed this type of coordinated harassment against individual targets as networked harassment. These networks of harassment are extremely well-defined in India as well, especially against prominent female journalists. An investigation earlier this year by *The Wire* shows how IT cells in India automate hate and direct abuse to their targets. Women are more specifically targeted by using misogynistic and gender-based slurs to try to discredit them.

4. Harassment as a Team Sport

Another feature of networked harassment of this sort is the sense of community and belonging that is fostered by the network of abusers. Much like how Johnny Depp's US defamation case against Amber Heard was divided into Team Depp or Team Heard, harassers during GamerGate also thought they were part of the "right" team and that the women were outgroup members being unfairly given access to "their" space. Thus, harassers will escalate attacks on the targets to establish their claim as part of the group and win the respect and approval of their peers. This pattern gets repeated across the internet and can clearly be seen in India in the harassment of prominent women online.

5. Algorithmically Amplified

The issue is complicated by the platform's amplification affordances that recommend similar types of content to users according to their preferences. For example, consider YouTube's culture of 'response videos', in which creators respond to other video content on the platform by showing why the initial creator is wrong. Watching one of these will likely lead the user to other similar videos

through the recommendation algorithm, adding incentives for creators to produce these videos. Lewis et al. show how this type of response culture promotes harassment of the target since they are portrayed as wrong and thus abusing them is seen as righteous and as protecting public discourse. Further, moderation decisions are made on specific pieces of content without taking the networked nature of these call and response videos into account, and thus the networked harassment directed towards the target is not registered. There can be multiple videos on the same topic, and YouTube's algorithm will reward abusive content the same as informative content, especially if it gets watched by a large number of viewers. TikTok is no different. The content of the videos created during a 'dogpiling event' hardly matters, even if it results in harassment of the target.

Platforms and law enforcement are not adept at dealing with any sort of coordinated activity. Thus, the onus of keeping themselves safe falls on the users. While there are some protections available through the health and safety features of most platforms, these are contingent on users pre-emptively blocking abusers and generally planning for harassment whenever they use their social media accounts. Institutional responses through companies have been truly insufficient.

6. Indian Short-Video Apps: Community Guidelines

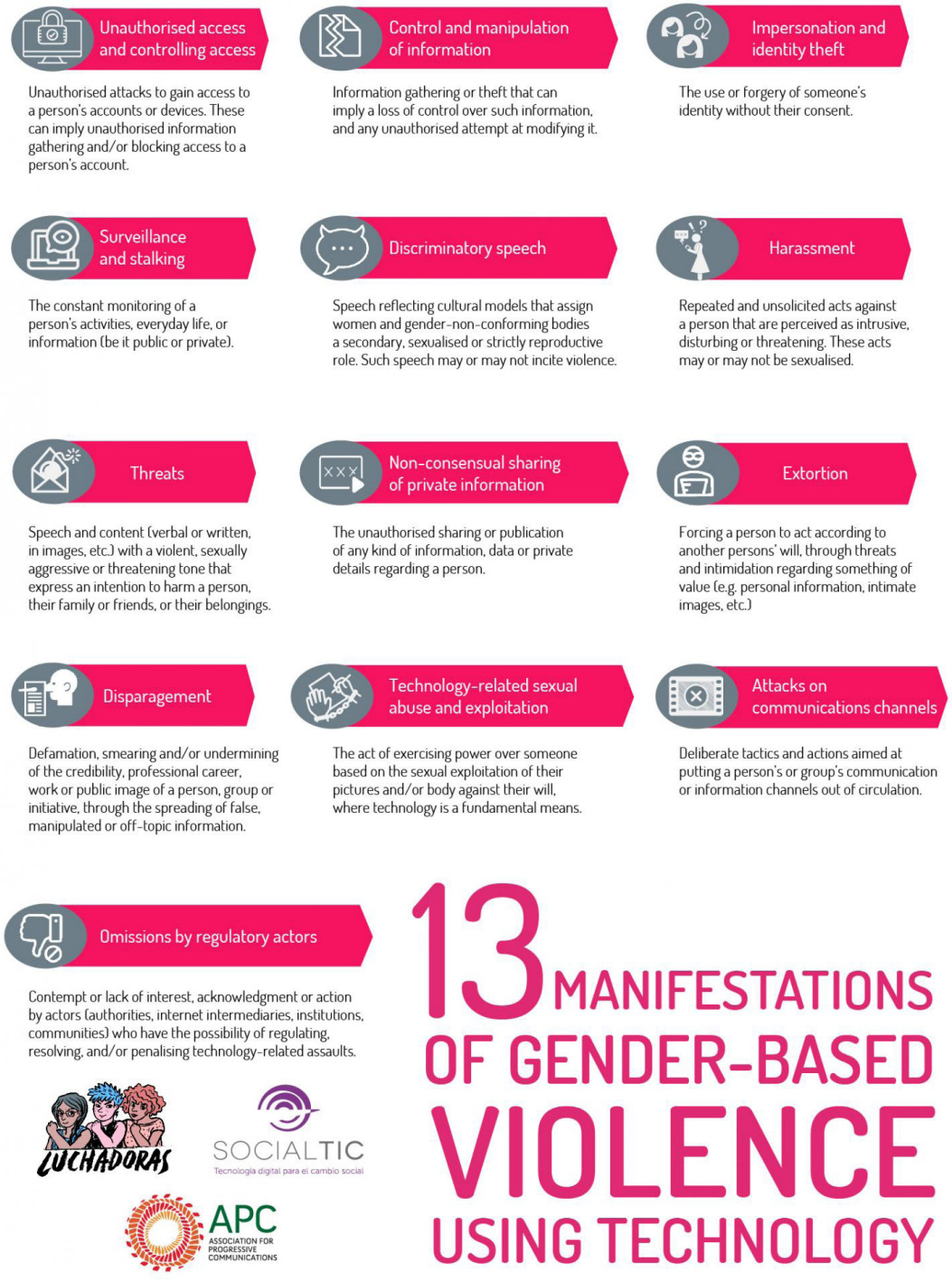
TikTok was banned in India in 2020. The void left by the short-video app was filled by homegrown companies that provide similar features. My colleagues and I analyzed the community guidelines of these apps to see how prepared these companies are for the various types of gender-based violence that occur online (paper forthcoming). Given that these problems of gender-based violence have existed online for almost as long as the internet, it is surprising how taken aback platforms appear to be whenever a new incident captures media attention. So far, I have not discussed community guidelines and what is acceptable behavior on a platform. It has been well-established that platforms are notoriously terrible at enforcing their own guidelines. However, it is important to analyze them since they reflect the public priorities of these platforms. By analyzing community guidelines and terms of service, we can find what actions the platform prioritizes and the ways in which they foresee their product being used and misused. Community guidelines set forth the rules according to which many moderation decisions can be taken.

Further, they allow us to see what a company's stated goals are and thus the nature of their future product decisions, as they shall hopefully create features to enforce these guidelines.

In our study, we used a modified version of Take back the tech! and APC's typology of gender-based violence to study how prepared short-video platforms Moj, MX Takatak, Josh, and Roposo are for various forms of gendered violence online (see Figure 1). Our findings have been mixed, at best.

- We found that these apps are exceedingly concerned about issues that arise from access to their services. The 'unauthorized access' and 'identity theft' categories directly deal with authentication and reporting and can thus have a fairly automated technical response. Companies are very serious about mismanaged credentials and give utmost importance to the authentication of their users. For the most part, account access by owners was very well-defined, and practices that violate these categories are also clearly defined by the guidelines.
- 'Non-consensual sharing of private information' is another clearly defined category, where privacy policies clearly state how companies will use the data provided to them. There is language in the policies that cover non-consensual pornography and include anti-nudity and anti-sexual content clauses.
- Platforms also have rules against 'discrimination, hate speech, and harassment' that specifically refer to gender as a category, but they rarely attempt to define what behaviors these terms can cover. Further, we did not find any evidence of the fact that companies recognize the uniqueness of videos as a medium of communication. Community guidelines also did not seem to take the prospect of networked hate very seriously and had no rules that took into account the algorithmic amplification of content that might be targeted toward women and gender minorities.
- 'Threats' are present as a category of prohibited content but often include a wide mandate like financial threats and rarely call out gender separately.
- The most surprising finding was that there were almost no guidelines against 'stalking and surveillance' or 'online domestic violence' in any of the platforms reviewed.

Figure 1: 13 Manifestations of gender-based violence using technology



Source: [Take back the tech!]

7. Recommendations for Platforms

Short-video platforms are not the first social media platforms to have existed. Neither is gendered abuse a new problem. Platforms have the opportunity and the obligation to learn from the harms of other algorithmically driven social media to ensure that more people do not fall prey to existing patterns of abuse.

Existing systems of power allow for propagators to send truly vitriolic abuse to their victims. Guidelines need clear and understandable language against hate speech, discrimination, and harassment, which includes unequivocal stands against such behaviors. This needs to be combined with robust safety and reporting features and swift action against perpetrators.

Further, platforms need to recognize that their services can and do become part of domestic violence situations. Abusers can use platform affordances to torment victims and exert control over their online lives in addition to existing offline abuse. Recognizing behaviors that make up a patchwork of intimate partner abuse and supporting victims when they request resources can go a long way towards making platforms safer. Similarly, stalking and surveillance behaviors, especially those that lead to violence against partners, or seek to curb mobility of women, need to be strongly protected against.

While platforms are designed for user engagement and virality, they need to learn to identify the larger digital context in which their apps exist and mitigate coordinated activity when it moves across platforms with the intent of harassing individuals or communities. The potential for harm of algorithmic, video-based social media is varied with abuse traversing similar pathways of virality to those that propagate dance trends, reaching millions of people across the world. In such a world, platforms need to recognize the socio-technical mechanics at play and actively protect their users. Their community guidelines can show that they understand the stakes of the problem in front of them. Strict implementation of these guidelines, targeted product features, and proactive threat monitoring can make platforms much safer.

Moderating Bodies: Reproducing Systems of Power through Platform Policies

ARJITA MITAL[1]

Abstract

Documents such as the community guidelines and terms of service remain authoritative sources for social media platforms to implement content moderation. In addition, they play a significant, behind-the-scenes role in shaping platform practices and determining the shared rules and expectations for the platform users. The guidelines cover a range of topics, such as breaking the law, avoiding

[1] Arjita Mital is a research associate at Digital Futures Lab, India. Her work looks at the intersections of digital rights, accessibility, gender, and policy. In the past, she has worked with several women's rights organizations on the issues of disability and sexuality, domestic violence interventions, violence against women in the agricultural industry, as well as women farmers' experiences around practicing agroecological farming.

punishment, threats against users, harassment, privacy, impersonation, scams and spamming, intellectual property rights, content labeling, and nudity and sexual content. While on the surface these guidelines seem gender neutral, they have strong gendered implications as the targets of gendered and sexual violence are disproportionately women and other marginalized identities. The essay analyzes secondary literature available on the taxonomy of community guidelines, terms of service, and other public-facing documents around governing policies of platforms where women users create and share content. It further argues that these governing documents, while instituted with the intent of protecting their communities from violence and harm, end up reproducing the same patriarchal conditions that allow that harm to take place.

1. Introduction

The existence of a diverse set of social networking platforms has equipped young people with the toolkit to deliberately, visibly, and creatively express several aspects of their identities. Although social media is entrenched in the larger global youth culture, young women and queer people's participation in online social networking sites with specific reference to expressions and experiences of sexuality and sexual identity is a significant part of the large pool of content being created on these sites. For young people located in parts of the world with restricted or closed civic spaces, social media allows for expressions of and negotiations with identities that are otherwise culturally and socially shunned. It is, therefore, imperative that these online spaces be safe, accessible, accommodating, and emboldening. In this regard, policy documents such as community guidelines and terms of service serve as important documents that shape platform practices and content moderation. Platform studies scholarship ascribes great significance to moderation processes, calling them "governance mechanisms" that can "structure participation" and "prevent abuse". However, reports from the past couple of years have highlighted deeply rooted algorithmic and human-centric biases in moderation mechanisms against individuals and bodies that may not fit the traditional standards of beauty. The systemic biases in these products result in a takedown of content or shadowbanning of accounts which is often attributed to the violation of platform policies. Reports have also demonstrated that content around body positivity by minority communities is at a greater risk of false flagging and subsequent removal than content created by non-minority users.

Instances of heightened scrutiny, policing, and censorship that people of color, women, LGBTQ+ communities, and religious and caste minorities experience on platforms due to regulatory efforts greatly outweigh the protection these communities receive from online harms for which the policies are instituted in the first place. In fact, then, platform policies become top-down mechanisms to exercise control by recreating and reproducing the same power structures that allow the harm to take place. The social and cultural ramifications of such practices often result in fear, self-censorship, and even offline hate crimes. A lack of attention to the reproduction of structures of power implicit in platform policies in conversations around platform accountability could, therefore, take away from the notion of platforms as fundamentally political actors shaping the global political sphere.

2. Regulating 'Obscenity'

Analysis of the appearance of rules for some of the most popular platforms shows that Facebook had the most comprehensive community guidelines for posting content. All content-sharing platforms had rules against sharing adult non-consensual sexual imagery, child exploitation imagery, and minor sexualization. Popular platforms such as Facebook, Instagram, YouTube, and Twitter also had policies pertaining to the distribution of content depicting adult nudity and sexual acts with specific emphasis on pornography, sexual intercourse, genitals, and buttocks. Across these platforms, "uncovered female nipples" was a focal point; central to this stance is an assumed shared understanding that nipples are gendered and that one is inherently more sexual. To save face, platforms have argued that they do allow certain depictions of female nipples. For instance, depictions of female nipples appear to be permitted in cases of breastfeeding or for "educational, documentary, scientific, or artistic" purposes. However, it can be argued that the decision to categorize content as "educational" or "artistic" is left up to the discretion of the platform and is often governed on a case-by-case basis. This is evident in several instances where content that complied with platform policies was also removed for violation of community guidelines. The invisibilization of discretionary decisions behind a seemingly, universally applied rule-based system also fosters a web of double standards which ensures that minorities face the brunt of unfair content moderation. Reports have shown that content by plus-sized and body-positive influencers was often flagged for "sexual solicitation" or "excessive nudity". It is also noteworthy that despite making

exceptions for female nudity in educational or documentary contexts, YouTube's sexual content policies go so far as to highlight that "out-of-context nudity in indigenous communities, during medical procedures, during childbirth, during artistic performances or during breastfeeding" may not meet their documentary exception.

Microblogging platform, Tumblr, which has historically been one of the few websites that allowed its users to share adult and sexual multimedia content, announced in 2018 that it will be banning all adult content across the site, including "photos, videos, or GIFs" displaying explicit material, as well as "illustrations that [depict] sex acts". Tumblr was a popular platform amongst young people to explore issues around sex, sexuality, identity, and body positivity. Individuals, especially those with marginalized identities, had formed communities and found safe spaces within the platform. Therefore, the announcement titled 'A better, more positive Tumblr', laden with overt moral sensibilities of a white, patriarchal society, subsequently alienated many longtime creators and users of the platform. While the platform assured that its policy strives to "strike a balance" and allows adult content that is not explicitly sexual in nature, Tumblr has also had a history of unjustly censoring primarily queer content.

Tumblr's original policy that did not ban sex-based content, from an affordance approach, meant that the platform, despite its best efforts to control it, became a hotbed for child pornography. Subsequently, Tumblr's iOS app was banned from Apple's App Store, which many speculate was the catalyst for the platform's decision to take down all sex-based content. Scholars have pointed out that platforms work not just politically, but also discursively to keep up with the ever-shifting social, cultural, and regulatory demands of other firms. It becomes important for platforms to align themselves with the legislative requirements of other services and businesses in order to secure future profits. However, for users sharing sex-based content or adult art blogs, Tumblr was an important tool to redirect communities and fanbases to their personal websites, Etsy and Patreon pages, secure commissions, and build art portfolios. Tumblr, then became a hub for successful businesses which otherwise would be housed under more niche adult websites, thus making them harder to access. With a blanket ban on all sex-based content and nebulous discretionary decisions around contextual nudity; sex workers and artists, who were mainly part of the LGBTQIA+ community, found their livelihoods under threat.

Platforms that render a service not just to their users but also to advertisers find themselves in a precarious position with regard to nudity and sexual content, jockeying between keeping loyal communities engaged by remaining neutral, and monitoring and moderating content to serve ads. YouTube has proclaimed that it is "committed to offering the best user experience and the best platform for people to share their videos around the world". However, in an endeavor to build a viable e-commerce platform for users and partners alike, virality ended up being celebrated. YouTube also received criticism for its proclivity to flag, restrict, and de-monetize content created by LGBTQIA+ and other minority creators. Compliance with YouTube's advertiser policies meant creators from minority communities would have to resolve to self-censoring to be able to continue partnering with YouTube. The platform's policy also further reinforces this idea of self-censorship by emphasizing that risque content must not be shared with the "purpose of sexual gratification" or depicted in a manner that is "intended to sexually arouse the viewer". Bonnie Ruberg's analysis of the policies of the video streaming platform, Twitch, highlights the existence of double standards in platform policies around nudity and sexual content. Twitch's policies stress on the bodies and the clothing of the streamers by stating that "attire intended to be sexually suggestive and nudity [are] prohibited". They further double down that since streaming is a "public activity", only "appropriate public attire" is acceptable. While such guidelines are posited to be gender-neutral, they do carry strongly gendered undertones. Ruberg's analysis also points out that despite their appearance and attire, women streamers are highly likely to receive sexual comments from viewers because of the heteronormative, male-dominated nature of the platform. In this regard, these guidelines aren't merely ideologically charged, but also put the onus of sexual violence on women streamers.

3. Language, Specificity, and Abuse: Who is 'Protected'?

Governing documents of digital platforms have also been noted by scholars to be imprecise and vague in their language. Policy documents across platforms disguise meanings in lengthy and complex writings all the while evading specificity. A strain on the nurturance of the community can be observed across platform policies; Instagram, for instance, repeatedly stresses that one needs to "foster and protect" this "diverse community of cultures, ages, and beliefs". Facebook's document emphasizes that it takes "great care to create standards that include different views and beliefs, especially from people and communities that might

otherwise be overlooked or marginalized". Jessica Maddox and Jennifer Malson, in a critical textual analysis of community guidelines of various platforms, note that social media platforms rely on the metaphor of the marketplace of ideas to absolve themselves and allow for the maximum amount of free expression. A key concern that emerges from platform scholarship is that the strategic usage of such metaphors tend to not just exonerate platforms from accountability, but also protect abusers and those in power who use hate speech, harassment, and abuse. Both Facebook and Twitter hate speech policies have come under scrutiny for offering protections to extreme right-wing and white supremacist politicians in various parts of the world. In its hate speech policy rationale, Facebook identifies certain protected characteristics, i.e., race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. Calls against violence, hateful, or dehumanizing speech against any of these categories are prohibited according to Facebook's policies. However, a ProPublica report noted that while Facebook offers protection to certain categories, hate speech against groups that are subsets of the aforementioned categories is often overlooked because one or more of their characteristics may not be protected. In the same policy rationale, Facebook further elucidates, "speech that might otherwise violate our standards that can be used self-referentially or in an empowering way" can also be exempted from facing consequences. It is clear that by skirting specificity and offering protections equally across groups while ignoring the nuance of hate speech against minorities, Facebook's policies make room for targeted attacks.

The political, cultural, and ideological conditions created due to the place-specific legal and social context and the simultaneously vague language in platform policies also stifle the safe online participation of women users and content creators and curb self-expression. This is especially true in cases of women who actively do political and cultural commentary online. Prominent journalist and author of the book 'Gujarat Files: Anatomy Of A Coverup', Rana Ayyub is no stranger to online violence. Online attacks against her range from ridiculing tweets/comments, doxxing, rape threats, deepfakes, etc. Activist and student leader, Shehla Rashid Shora, deactivated her Twitter in 2018 due to Twitter's refusal to take action against the targeted attacks against her citing that the attacks did not violate their policies. Twitter, often, upon not finding reported content to be abusive or objectionable according to its policies recommends that the user resort to unfollowing or blocking the account. Instagram, similarly, advises that its users

unfollow, block or delete the offending comment/s. Instagram also suggests that "misunderstandings" be resolved "directly between members of the community" as far as possible by asking the offenders to remove the offending content. However, as can be observed in the cases of Ayyub, Shora, and many other women who are targets of such violence, offending comments, posts, tweets, etc., are often in hundreds, if not thousands. Burdening communities at the receiving end of such violence allows platforms to pre-emptively disavow the harms suffered.

4. Conclusion

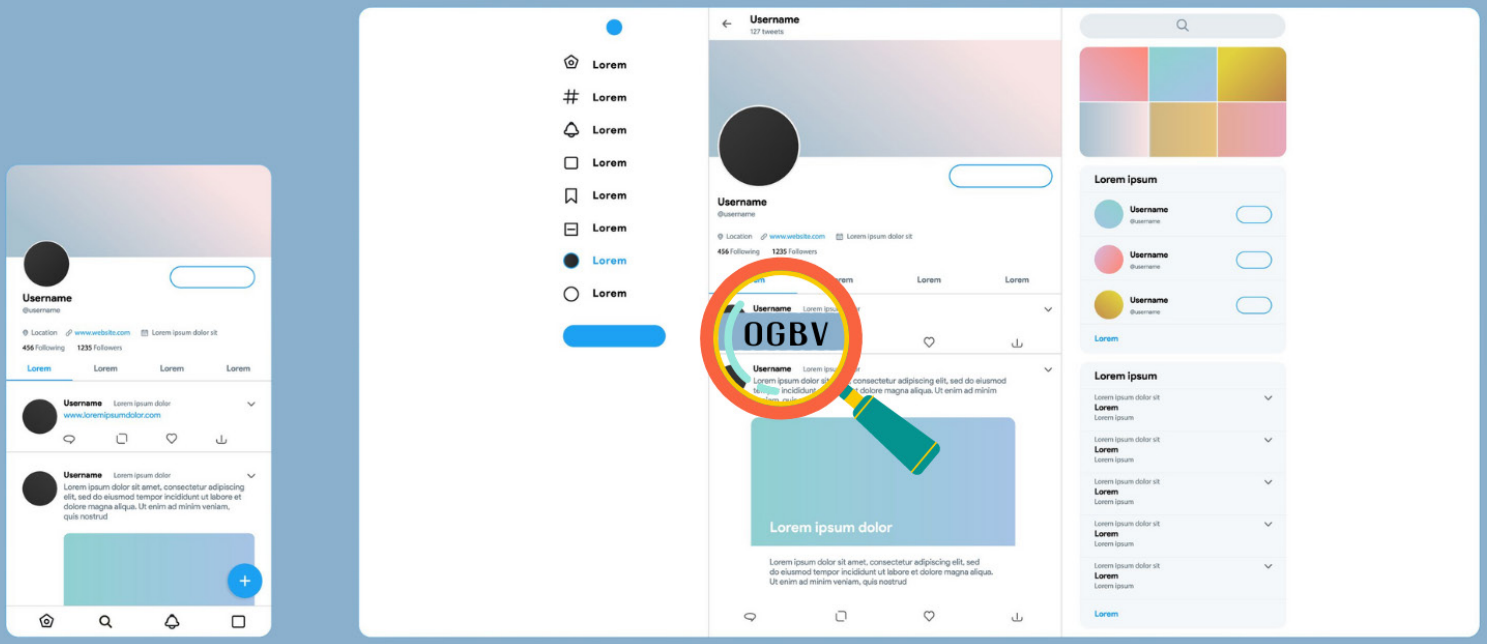
The language used to describe bodies, sexuality, harms, and abuse in these policies and guidelines reflect the structures, logics, and values of the specific cultural contexts where they were designed. In 'The Platform Society: Public Values in a Connective World', Van Dijck speaks about the non-neutrality of platforms, whose policies and infrastructures are engraved by specific norms and values of the people they were designed by. White, male hegemonic libertarian traditions of social media platforms have long dictated what the policies will consider problematic. Across platforms, the language used in policies pertaining to nudity and sexual content has established sex and sexuality as the antithesis of positivity, safety, and community. Sexist double standards are also rampant across policies where "female nipples" or "female-presenting nipples" are prohibited. Definitions of obscenity and those of pornography conceived from an overwhelmingly white, male perspective leave considerable room for bias and discrimination against marginalized communities who are already at a heightened risk of violence on these platforms. Being removed from the cultural contexts of regions where these platforms operate, it leaves a significant amount of room for platforms to not only commit unintentional errors, but also deliberately turn a blind eye to socio-political conflicts in certain regions whose user base they benefit from.

Digital spaces have been touted as important spheres, especially for participation of persons from marginalized and persecuted identities who live in closed civic spaces. However, the policies of platforms are heavily biased in the interests of corporations and orchestrated to protect people in power. Hyper-detailed guidelines that govern women's bodies and sexualities, and flimsy policies around hate speech that protect powerful abusers, coupled with communicative affordances that enable anonymity and asynchronicity allow platforms to birth new masculinities and power structures. Conceptualized as tools to protect members

of a community from harm, these guidelines end up reproducing systems of oppression that exist offline to govern women's self-expression. While on the surface these guidelines appear gender-neutral, they have strong gendered implications as the targets of gendered and sexual violence are disproportionately women and queer people. Platform policies, then, become tools that govern 'unwanted' sexual content. However, in the case of women content creators, such policies play a significant role in determining the 'legitimacy' of the content and in turn the 'legitimacy' of the bodies that create the content. Moderation mechanisms that exist in these conditions perpetuate norms of propriety dictated by the standards of the Global North that allow for the existence of 'acceptable' bodies (white, slim, non-disabled, female-presenting) and deem all other bodies overtly sexual. Community guidelines, terms of service, and content moderation policies then become protectionist documents for the 'acceptable' bodies, i.e., protecting them from abuse and harassment, while demonetizing, suspending, or removing content creators with 'unacceptable bodies'.

A 16-month inquiry into 'Disinformation and Fake News' by the United Kingdom's Digital, Media, Culture, and Sport (DCMS) committee reported that "Social media companies cannot hide behind the claim of being merely a 'platform' and maintain that they have no responsibility themselves in regulating the content of their sites". Platform studies scholarship has argued that platform politics is becoming increasingly difficult to separate from global politics. There is an urgent need to democratize these spaces and the first step needs to be to engage with the diverse kinds of scholarships and perspectives that have been emerging across the globe.

Community-led Moderation Systems



Designing for Disagreements: A Machine Learning Tool to Detect Online Gender-based Violence

ARNAV ARORA, CHESHTA ARORA, AND MAHALAKSHMI J[1]

Abstract

This essay elaborates on the authors' experience of building a user-facing browser plug-in to detect and mitigate instances of online-gender-based violence (OGBV) in three languages: Indian English, Hindi, and Tamil. The plug-in deploys machine

---

[1] Arnav Arora is a PhD student at the University of Copenhagen, investigating the impact of media bias on public opinion. He has previously worked as a research engineer in the content moderation divisions at Checkstep and Factmata.

Cheshta Arora is an independent researcher, writer, and ethnographer based in India, studying socio-technical systems.

Mahalakshmi J. is a Master's student at Oregon State University. They have previously worked as a data scientist at Tattle.

learning (ML) as well as non-machine learning features to help users detect, replace, and archive problematic content on their Twitter feed. The design of the ML model was grounded in participatory approaches to account for contestations and disagreements vis-à-vis OGBV, and how users use different parameters to decide what is problematic. To capture these diverse notions of harm, the authors invited six expert annotators for each language to annotate 8,000 posts per language. The essay reflects on this experience of building an ML model that is co-designed with members of a larger community, the challenges encountered in the process, the limitations of ML-based approaches, and what it means to design with 'human-in-the-loop'.

1. Introduction[2]

Tweet 1: <handle replaced>shows us how to balance glitter in a #RGRK metallic skirt. #RGRK #fashiondesigners #Lisahaydon

Annotator 1: This post had a picture. I was concerned regarding how the picture will be circulated without consent

Annotator 2: But this is also a picture of a model posting with consent

Tweet 2: <handle replaced>rip sir

Annotator 1: I read this as an instance of possible misgendering which is a common problem faced by our community.

Annotator 2: I read it as a harmless tweet

Given above is an excerpt of a discussion between two annotators and how they disagreed on their annotations vis-à-vis a particular post. As is evident, both annotators had different and valid reasons vis-à-vis the post and if it should be marked as online gender-based violence (OGBV) or not. Both were part of a larger team of 18 annotators who were invited to annotate a total of 24,000 posts in three languages: Tamil, Hindi, and Indian English. All annotators identified themselves as sexual and gender minorities and, as part of their professional and personal lives, had engaged deeply with the problem of online harassment and hate speech,

[2] Interspersed with the text of this report are verbatim reproductions of violent, abusive, misogynistic, offensive, and stereotyping mentions, which may be both disturbing and upsetting to read. Please consider this a trigger warning.

either as members of community-based organizations, journalists, activists, or community influencers. The annotators were engaged as part of a larger effort to build a user-facing browser-based plug-in to help mitigate instances of OGBV. The browser-based plug-in called Uli[3] (Tamil: உளி, meaning chisel) was co-designed in collaboration with a larger community that is often targeted for its gender/sexual expression and opinions. The tool offers both machine learning (ML) as well as non-machine learning features.

Non-ML features include options such as an archive button to help users build evidence. With the help of the archive button, users can store a tweet either locally or on a mail client. The tool also allows users to build a network of friends who can be easily notified if anyone in the network is at the receiving end of online violence.[4] Finally, the tool offers resources such as a simplified version of Twitter's community guidelines in Hindi[5] as these have not been translated into under-resourced languages even though Twitter supports Tamil, Hindi, and seven other Indian languages. An option to build a custom slur list, which could redact those words on a user's feed is also part of the non-ML features.

These features were incorporated into the tool after a series of consultations, and the authors believe that they will be useful in mitigating forms of OGBV that are not addressed by legal or platform governance measures. Among other things, the non-ML features allow the tool to detect hateful content that is difficult to engage with or the kind of content that can produce everyday fatigue for users, especially those who use social media for professional or political reasons, or for emotional support. It is worth noting that everyday fatigue, for instance, does not fall within the ambit of 'criminality' that grounds the Indian legal system.

The ML feature, on the other hand, could largely be a proof-of-concept to make a case for more investments in under-resourced languages and for building AI/ML models using scaled-down approaches built in consultation with communities affected by tech-mediated violence.

This essay elaborates on the experience of building this ML model and co-designing it with community members, the challenges encountered in the process,

---

[3] The project was supported by Omidyar India Network and developed in collaboration with the Centre for Internet and Society and Tattle Civic Tech. The tool was officially launched on 28 July 2022 and is available for download.

[4] This feature is yet to be released.

[5] The Tamil translations of these resources are yet to be released.

and what it means to design with 'human-in-the-loop'.

To put it simply, the ML model learns from pattern recognition and can detect future content depending on what it is trained to detect. The success of such a model depends on the quantity and quality of a dataset, how it is being annotated, and who is annotating it. To begin, the authors interviewed experts to understand the need for a tool such as this. In addition, they conducted one workshop in English, four smaller focus group discussions in Hindi, and 10 interviews in Tamil. The group discussions, interviews, as well as the involvement of 18 annotators, broadened their understanding of what constitutes 'harm', the kinds of content the tool should be able to detect and flag, and the types of content it should leave untouched.

To build a robust and diverse dataset, they crowdsourced a list of slurs and asked the group discussion participants to add as many offensive words/phrases as possible. This slur list was used to identify and scrape content off Twitter. This list also helped incorporate the tool's redacting feature that can blur these words for users without depending on the ML model. This list is not exhaustive and could grow with the number of users on the tool.

Additionally, the authors of this essay manually built a list of accounts that are often at the receiving end of online hate, as well as a list of accounts found perpetuating hate and abuse on Twitter. This was complemented by data shared by Microsoft Research (MSR)[6] and IT for Change, including a list of women who are influential or highly active on Twitter and are often subjected to online abuse and harassment, as well as annotated data for different variants of online harm. The team scraped tweets using three methods: (1) crowdsourced slurs and keywords, (2) tweets by known perpetrators, and (3) replies to highly influential women on Twitter; in total, scraping close to 1.6 million tweets from 2020 to 2021.

In order to curate a diverse dataset and select tweets to be annotated from the scraped collection, the team chose a data pooling approach, selecting 8,000 tweets per language for expert annotations. For the pooling, they trained ML models for each language on existing datasets for hate speech, toxicity, misogyny,

[6] The team used MSR's central database of politicians, influencers, and celebrities to extract replies to known influential female accounts. This data, however, was limited to female accounts and was not inclusive of other gender and sexual minorities. This bias was corrected to an extent by creating a diverse slur list as well as by curating a list of accounts that represented influential sexual and gender minorities on Twitter at the intersection of caste and religion.

and aggression detection. These models were then used to predict a score for the task they were trained on, for all tweets that were scraped. The assigned scores, which represent the confidence of the models that a particular example is worth flagging, were then averaged and tweets were picked from different mean intervals. A higher average confidence implies that most classifiers are in agreement that a tweet is toxic. A lower average confidence indicates that more classifiers disagree on the toxicity of the tweet, whereas a zero-confidence score suggests that the models consider the tweet to be not toxic. The team picked more tweets that didn't fall in either extreme of the mean intervals, thus prioritizing tweets that were mislabeled by the classifiers trained on existing datasets and ensuring good diversity of data for training.

The success of an ML model, however, also depends on an 'agreement score'. Put simply, it means that a machine learns better if the humans training it (i.e., the annotators and their annotations) agree on the nature of the content. That is, it can learn better if all annotators label a red ball as a red ball and a blue ball as a blue ball without disagreeing with each other. As the excerpt this essay started with shows, the agreement scores in this area of research are generally low. Section 2 uses examples to discuss the subjective nature of the task that leads to a low agreement score.

To maintain a high agreement score, it is important to arrive at precise definitions of the phenomenon that we want the machine to learn to identify. To this end, the researchers developed a set of intuitive guidelines. These guidelines were tricky, allowing the annotators to label from their experience while also capturing that experience under the 'correct' label designed, to improve the agreement score among different annotators. Multiple discussion sessions between annotators were organized to communicate the guidelines and learn from their rationale vis-a-vis each post. Some of the feedback from these discussions was incorporated into the guidelines.

The guidelines had three labels:

1. Is this post OGBV?
2. Is this post OGBV when directed at a person of gender and sexual minority?
3. Is this post explicit/aggressive?

As one can see, two of the three labels relied on some understanding of OGBV.

This is because, defining OGBV, the authors realized, was far more difficult than defining the redness of a red ball. The study used two labels (1 and 2) to capture OGBV in order to account for the meaning and the direction of the post. This is explained further in the next section.

2. Delineating the Problem of OGBV

OGBV manifests in various ways. A visual infographic created by Take Back the Tech!, Luchadoras, and SocialTic, identified 13 manifestations of OGBV. Even the nomenclature around OGBV is still contested as existing literature notes the use of various prefixes such as 'cyber', 'digital', 'electronic', 'internet', 'online', 'tech-mediated', etc., to note the various ways in which it manifests. The initial consultation and interviews with experts confirmed that image-based harassment is more prevalent in India. The use of Twitter and text-based harassment is heavily skewed toward the literate urban elite. Thus, the tool was designed with an understanding that it is impossible to exhaust all instances of OGBV, and the tool will only focus on the kinds of harms that affect users who are politically active on Twitter and produce fatigue among them as they use social media to engage with their communities. The diversity in the nomenclature was narrowed down to identify text-based abuse as opposed to multimedia content.

Thus, our problem in delineating OGBV was not vis-a-vis the diverse nomenclature but rather what constitutes gender-based violence, especially when the focus is on text-based abuse. Even though our funding agreement with the donor was to work on the problem of OGBV, the use of gender as the overarching explanatory variable for violence was problematic, especially in postcolonial contexts. Postcolonial as well as LGBTQ+ scholarship has already established that gender-based violence should be disassociated from binaries of man/woman and modernity/tradition. That is, gender cannot be read as 'woman' and existing gender-based violence ensues from modern relations of colonization, imperialism, and inequality, rather than static cultures stuck in patriarchal traditions. In other words, gender relations and the violence afflicting them are intersectional through and through, and are affected by multiple relations, encounters, and histories of violence. In order to remain mindful of this literature in our own research, the authors needed to navigate the following concerns:

First, in the case of text-based harassment, it was a matter of debate if the ML model should only focus on the text of the tweet to see if it evokes gender

or sexual abuse – i.e., misogyny, sexism, homophobia, transphobia – without assuming any context or the direction of the abuse. For example, in the two posts given below, misogynistic undertones are explicit in the text of the tweet itself, irrespective of who is targeted:

> – Pic 1- #नारी_शक्ति (translation: #femalepower), Pic 2- #नारी_कुत्ती (translation:#femaledog) बाकी जय श्री राम तो है ही (translation: The rest is Jai Shree Ram) #goodnight
>
> – We need tons of patience while fighting false Fabricated cases filed by misusing #GenderBiasedLaws #498A. May god give you all the strength #Repeal498A #MarriageStrike

Second, the above contention could be further complicated by pointing out that a lot of Twitter users who identify as gender and sexual minorities may simultaneously be disadvantaged on account of caste, religion, ethnicity, and other intersecting social identities. These users might find a tool that only detects misogynistic tweets quite useless. If the goal was to mitigate fatigue resulting from a constant barrage of hateful and offensive content, the tool and its detection had to be as expansive and intersectional as possible to include hateful content in general, rather than only misogynistic or sexist content. For instance, the ML model had to find ways to account for following tweets that would otherwise be seen as caste-based, religion-based hate speech and not gender-based abuse:

> – जो जिस भाषा में जवाब चाहेगा, हम उसी भाषा उससे बेहतर ढंग से जवाब देते हैं ... मुँह से और लात से भी ... भीमटों, लाल और हरे टिड्डो के लिए जनहित में जारी.... [sic] (translation: Whoever wants an answer in whatever language, we can respond to them even better … in words and in kicks … this is issued in public interest…)
>
> – <handle replaced><handle replaced>Ur own quom [sic] is not afraid of you … look what is happening in Afghanistan … chal nikal (translation: get out)

Third, a lot of the content that might appear harmless to a 'regular' user becomes offensive when it is specifically directed towards a community or an individual. For example, the following two tweets could be seen as part of the usual banter on

Twitter but can be used to target a vulnerable user:

> – <handle replaced>if u truly are inspired you would not tweet.
>
> –<handle replaced><handle replaced><handle replaced> इनका पिछवाड़ा जल रहा है । (translation: their backside is burning).

Fourth, while the guidelines were initially designed to capture the gendered aspects of a tweet as well as hate speech, the team's annotators added one more classification, of creepy posts, that the tool should be able to act upon. This added an extra layer to the definition of OGBV, and one on which annotators disagreed considerably. By including a category of creepy posts, the team moved away from merely focusing on the meaning of a post to the action of repetitive posting. Thus, a post such as, "Good morning!!!" when posted by a stranger, either once or 10 times, could also be read as creepy.

All these concerns had to be considered. Where others in the field of ML research have had the luxury to create a long list of classifications to capture the diversity of abusive content, these labels had to be as intuitive and simple as possible since annotators were mostly volunteering for the task.[7]

With the three labels mentioned above, the research team hoped to capture all these concerns vis-a-vis OGBV. For label 1, annotators only had to focus on the meaning of the post and discern if it evoked gender and sexual problematics and mark it as OGBV if it was explicitly misogynistic, transphobic, sexist, or homophobic. For example, the following post would be marked as OGBV under Label 1:

> "– We need tons of patience while fighting false Fabricated cases filed by misusing #GenderBiasedLaws #498A. May god give you all the strength #Repeal498A #MarriageStrike"

With label 2, annotators could capture all kinds of content – religion and caste-based hate speech, creepy posts, or regular comments that they think could be

[7] While annotators were compensated for the task and received Rs. 8 for every post, their engagement was considered voluntary and could be terminated whenever they wished to do so and without any contractual obligations.

abusive if they are directed at a person identifying as a sexual and gender minority. Thus, a post such as "Rice Bags!"[8] would be marked as OGBV under label 2 but not under label 1.

Finally, label 3 captured posts that were visibly offensive and aggressive irrespective of who they were directed at. For example, something like "Shut Up!" could be marked as 'explicit/aggressive' under label 3. Simultaneously, it could be 'not OGBV' under label 1 and 'OGBV' under label 2.

Given that each label was expansive, annotations were not taxonomical and wouldn't produce statistically significant knowledge about the prevalence of OGBV and its different manifestations on social media, if OGBV were to be narrowly defined. However, with this label, the hope was to capture posts that produced fatigue and shrunk the space of engagement on Twitter for sexual and gender minorities situated at various intersections of caste, religion, ethnicity, and other intersecting social identities.

3. Making Sense of Disagreements

3.1 From a social science perspective

These labels, while intuitive, led to different kinds of disagreements among Hindi and English annotators. These disagreements, however, should not be seen as a hindrance since to reach an absolute agreement was not the intention of the project. The research team invited 18 annotators precisely to capture these disagreements and strengthen the model. While some disagreement could be accounted for in the model, the presence of other unanticipated ones, that affected the agreement score substantially, made it difficult from the ML perspective. This section elaborates on the disagreements qualitatively. Section 3.2 describes how these disagreements would potentially affect the ML model, the ways in which they could be circumvented, and consequently, the limitations of a model such as this one.

The disagreements that could strengthen the models were those that emerged from the diverse experience of our annotators. For example:

– Rice bags!

– <handle replaced>RASHTRIYA SAMLAINGIK SANGH

– <handle replaced><handle replaced> भो श्री के (translation: a cunt)

[8] This is a derogatory slur, often used by Hindutva ideologues, to target Christians in India.

In the above examples, only those who have a contextual understanding of caste-based hate speech will be able to mark "rice bags" as OGBV under label 2. Similarly, a political play on the RSS,[9] "Rashtriya Samlaingik Sangh" where *samlaingik* is the Hindi word for homosexual, exploits normative heterosexuality which could be offensive for sexual and gender minorities. A wordplay on a Hindi gendered slur भोसड़ी के (translation: 'a cunt' or 'of a cunt') as भो श्री will likely be identified only if someone has commonly encountered similar wordplay and may be missed by others. Similarly, among Tamil annotators, only one person could decipher the use of the word 'nine' as targeting trans groups, whereas others read it as a regular post reporting an incidence of violence.[10]

Some disagreements, however, also emerged because of other reasons such as confusion around the three labels, cultural differences between Hindi, Indian English, and Tamil, political differences among the annotators, etc. It is important to note that these disagreements were as valid as any other and should not be seen as errors that needed to be corrected. While they did lead to considerable panic within the group vis-a-vis the efficiency of the model, these disagreements point towards the nature of the problem at hand and demand that ML models be built keeping these challenges in mind. Current ML approaches, however, are not equipped to address each of these disagreements and still be efficient for every user. This means that the team had to make difficult decisions and narrow down definitions, asking annotators to not mark something as OGBV under label 1 and label 2 even when their opinion was that a particular post should be marked as such.

The discussion that follows builds only on the disagreements among the Hindi and English annotators. This is for two reasons: first, the authors writing this essay were only familiar with Hindi and English, and one of the authors of this essay was involved in facilitating Hindi and English annotations. Second, the agreement score for Hindi and English was considerably low compared to the agreement score for Tamil, which forced the authors of this essay to step back and assess the concerns surrounding the Hindi and English posts. Outlining the discrepancies vis-a-vis the agreement scores between Hindi and English on the one hand, and Tamil on the other, would have made the arguments richer. However, this essay only focuses on two of the three languages under consideration.

---

[9] RSS, Rashtriya Swayam Sevak, is a Hindu right-wing, volunteer-based organization in India.

[10] The authors would like to thank their co-lead on the project, Brindaalakshmi K, for sharing this insight.

Confusion with labels: To begin with, the three labels led to considerable confusion among the annotators. While the team had consistently insisted that the annotators were the experts, the guidelines still had certain rules that they had to follow to capture each post under the correct label. This led to some confusion. For example, initially, a few annotators marked under label 1 extremely aggressive caste-based religious-based hate speech which should have been marked under label 2. While this confusion was, to an extent, addressed after multiple rounds of annotations, certain annotators continued to capture some of these posts under label 1, albeit unconsciously. This also reflected the limitations of the category 'gender abuse' when dealing with intersectional abuse and is elaborated further.

It was difficult for everyone to strike the right balance between the guidelines and their intuition. At times, some annotators would go against their intuition to be "true" to the guidelines, others would continue with their intuition and disregard the guidelines, and yet others would interpret certain tenets in the guidelines differently. For instance, for label 3, the team had arrived at a compromise to not capture a post as aggressive if it appeared well-intentioned. This well-intentioned clause for label 3 led to varied interpretations of aggression. For example, a post such as this could be read as both 'well-intentioned' or demeaning:

> <handle replaced> NOT many Women are accomplished as you are in this field before the age of 36

Confusion due to different linguistic backgrounds: Disagreements emerging due to different linguistic backgrounds were more specific to the Hindi annotators with regard to the Hindi dataset. The annotators for Hindi were from different states, including Delhi, Gujarat, and Maharashtra. A cursory comparison of the Hindi and English datasets suggested that the former had an overrepresentation of electoral/political/hate speech. This speech was often written either in the textbook/standardized style or in a style specific to some regions that was difficult to understand for annotators from different regions. Moreover, since Hindi has several local dialects and styles with varying political contexts in each area, it was difficult for annotators from one region to pick up on references from another. For example:

> – <handle replaced> कसाई के श्राप से गाय की आयु लम्बी होती है ।
> (translation: The butcher's curse increases the lifespan of the cow).
>
> – expected by <handle replace> under Mamta begum they can definitely do it.

Difference in the nature of Hindi and English datasets: Initially, the guidelines were created on the basis of an English dataset that was randomly sampled and didn't have enough iterations of the kind of content desired. This led to some revisions once the annotators were onboarded and the final dataset was pooled for each language. The final English dataset, however, led to another interesting problem. A lot of the content scraped using English gendered slurs was from the larger Anglophone world that has witnessed different kinds of appropriations and casualization of explicit language. For instance, "bitch", "slut", and "whore" are commonly used and appropriated by Black Americans. While some English annotators were aware of such instances of appropriation in the Anglophone world and themselves engaged in such appropriations in their everyday lives, others had strong opinions against using these words in public. This means that the English annotators had different thresholds for tolerating English gender and sexual slurs. On the other hand, among the Hindi annotators, instances of appropriation were very rare, and everyone agreed to mark a post with gendered slurs as OGBV under both labels 1 and 2. Annotators' responses to gendered slurs in three languages could not be standardized through one set of guidelines. For example, the following post could be marked either way:

> – <handle replaced>don't stress you're never mean from what I've seen unless ur [sic] secretly a vicious bitch

Differential understanding of political context and affiliations: While some annotators had strong opinions against any use of explicit language in political speech, others believed using some explicit language was okay when speaking back to power. For instance, some annotators who were also journalists were usually tolerant of posts that used explicit language to mock or criticize mainstream media or personalities.

> — <handle replaced>The most criminal and corrupt section of people feed their greed are media people, you are high on this heirarchy [sic] madam including your hubby.

Varied understandings of gender abuse and its ancillary concepts such as patriarchy: Despite our attempts to narrow down the definition of OGBV, some annotators had a very expansive understanding of OGBV, and rightly so. Some posts were marked as OGBV because the annotators understood patriarchy broadly as domination where each relation of domination crisscrossed with the other and could not be isolated which led to considerable disagreement.

This disagreement also pointed to the limitations of 'gender' as a discursive category specific to the anglophone world. Its connotations cannot be transposed into other regions where activists have to work with a very expansive notion of violence that cannot be reduced to one category. Any attempt to transpose one conceptual category to another context leads to serious problems of epistemological colonialism and misplaced framing of the problem which wreaks further havoc.

Some of these disagreements were incorporated into the guidelines. For instance, references to male sexual organs for men from marginalized communities, such as "*Katii Lulli*", or a threat of sexual violence directed at a male journalist were also marked as gender-based violence under label 1.

> — <handle replaced>Jo dono jagah se katwa le wo ashutosh kehlate hain... (translation: whosoever gets cut from both sides is called Ashutosh…)

While these disagreements were quite insightful from the social science perspective and for future research on the topic, they posed a strong challenge to the development of the ML model.

3.2 From a machine learning perspective

In a traditional supervised ML text classification pipeline, a model is trained on a particular set of *data* along with their corresponding labels. This process defines the world the model knows and teaches it to differentiate and identify specific

phenomena (e.g., positive vs negative sentiment sentences). As mentioned previously, for generating a labeled dataset, multiple human annotators (be it crowd workers or experts) are asked to label each post, and typically, a consensus for the final label is reached by taking a majority vote. This is done since, while training the algorithm, there needs to be a definitive, distinct, single label for the phenomena expressed in the text, so that it can be categorized and captured. Differing labels for the same or similar posts end up confusing the model on what it should be learning, and in turn, adversely affect its performance.

This process has its flaws. To begin with, a model is only as good as the data provided to it and it has no common sense built into it. So even if a post is intuitively or logically close to the phenomena we're trying to capture, in our case OGBV, the model won't be able to flag it if it hasn't seen similar examples during its training. Further, because the model has no reasoning capabilities of its own and learns exclusively from the textual data provided to it, it picks up on human biases and often toxic content like stereotypes present in the training data and amplifies them.

Lastly, as previously mentioned, the model needs a single label for each example it is trying to learn from. This need is fundamentally at odds with our attempt to include multiple viewpoints and accommodate different experiences and perspectives.

The authors attempted to circumvent these limitations in the following ways:

- The approach was guided, and the data was annotated, by people who are often at the receiving end of the phenomena being captured, hence minimizing the risk of including harmful stereotypes or otherwise toxic content in the data, along with getting the best estimate of what kind of content is causing fatigue when using Twitter.
- The annotations for OGBV are separated into the aforementioned two labels. A model trained on the first label will capture a generic version of OGBV that is more constrained in its sensitivity since the target of the comment is unspecified. On the other hand, a model trained on the second label, where annotations specifically consider that the comment is directed at a person belonging to a gender or sexual minority, will be more sensitive. This is because a comment that does not appear as OGBV at first glance may become so when directed at someone who is marginalized. The second label thus allows for control of sensitivity based on the user of the extension.

- Typically, disagreements between annotators are resolved using a majority vote. While this approach has proven to be successful with many classification tasks, it does run the risk of suppressing minority viewpoints. The authors later experimented with ways other than a majority vote for collapsing their existing labels into final ones. A simple method that was attempted was flagging the post, if at least one annotator marks it as OGBV instead of a majority.

In the future, based on time and funding, the authors will try the following steps to build a robust model:

- One approach being tried is Jury Learning, which is a supervised learning approach where researchers get to specify whose voices an ML model should imitate, and in what fraction. To achieve this, ML models are trained on each annotator's annotations. This model predicts what label each individual annotator might assign to a particular item. The algorithm then picks annotators from different groups to form a jury based on the specifications. On seeing an unlabeled example, the algorithm predicts the labels assigned by each annotator in the jury pool. These labels are then aggregated to arrive at the final classification label. This approach would let us experiment with different jury compositions as well as observe the effect of different compositions on the final classification label.
- Finally, the authors will experiment with launching a custom model for each user using the extension. As discussed earlier, each individual's exposure, tolerance, and definition of OGBV varies. To account for each user's preferred sensitivity level, using active learning to continue training and updating the model based on a user's preference will allow it to be updated based on the user's feedback to the extension.

This essay approaches the predictions of the ML model with a healthy grain of salt. ML models, while being incredibly good at narrowly defined tasks, tend to behave erroneously when used on data dissimilar to the ones used in training. Additionally, these models are black-box, and it is an active area of research to establish a clear reasoning path within the model, which makes it incredibly hard to interpret the predictions of a model. To this end, there is work being done on a thorough error analysis and keeping the ML model in beta to get user feedback and update or remove the model as necessary.

While being aware of the limitations and inefficacies of using ML approaches, the authors of this essay decided to build the model as a proof-of-concept advocating scaled-down AI/ML interventions in under-resourced languages that are co-designed by and with the larger community.

4. Conclusion: Designing with Human-in-the-Loop?

While it's still unknown how the model will behave as the development of the tool is still underway, the essay has pointed toward some of the challenges of designing with human-in-the-loop and the limitations of existing approaches. The rationale behind developing such a model was to illustrate the efficacy of scaled-down moderation models designed with communities, individuals, and targeted groups. The challenges that the team encountered point toward the need to look at the algorithmic designs more critically.

While the authors began by co-designing this tool with all the stakeholders who would benefit from a tool such as this, ML approaches are not yet ready to fully take into account the contingencies and disagreements emerging from these participatory approaches. Existing ML approaches are still not able to code these healthy disagreements among the collaborators as meaningful information, which is important, especially for projects that profess ML/AI for social development. Nonetheless, as it bears repeating, the non-ML features will keep the tool alive till the team figures out how to code disagreements into the model.

Feminist Communities of Care and a Transnational Response to Redesigning Social Media Governance

QUITO TSUI[1]

Abstract

Social media governance, in its current iteration, relies upon a feminized chain of labor that stretches across the globe. By redistributing digital care to hidden content moderation centers, Big Tech companies have evaded their responsibility of creating more purposeful online spaces. Instead, they have repeatedly chosen to transfer harm to precarious workers; turning towards technological solutions which are poorly placed to deal with the complexities of moderating online

[1] Quito Tsui is a research manager at The Engine Room where she works on technology in the context of conflict and humanitarian organizations, and the development of a technological environment rooted in human rights. Her other research work includes thinking about transitional justice, memory, and how to go about the work of care.

violence; actively refusing to take the kind of measures that could reduce harms occurring in the first place. The roots of poor social media governance and online violence are connected by threads of carelessness perpetuated by Big Tech companies. This essay argues that centering care within our approach to social media governance is vital to constructing a tenable alternative. Drawing on feminist ideas around community care-building, and transnational feminist strategies of creating globalized webs of understanding, this intervention proposes potential entry points into rethinking social media governance.

1. Introduction

How do we tell ourselves the story of a problem? When we think of social media governance there are many threads to follow – which one you choose to pull on, in many ways, shapes the solutions one might proffer. But if we follow the thread of harm, one story becomes overwhelmingly clear: women and members of minority groups are repeatedly targeted online, their everyday digital lives marred by cyberviolence harassment. This story repeats narratives of old and new iterations of centuries of violence against women and minoritized groups.

The narrative of solutions, however, is complex; what seems necessary to combat online violence oscillates, caught in questions around which characters should be involved and to what extent: More or less tech? More or less scale? Most of this discussion, however, is happening within the status quo – these questions ask how we can make small changes to a system that increasingly appears not fit for purpose. Indeed, when we look at the deeply flawed system of social media governance it is easy to focus on post-incidental reactions – Who is being harmed in this space?; What legal mechanisms are there for accountability?; How can we remove or punish those engaging in social media's purposefully misogynistic designs? In doing so, we often limit our ability to be emboldened in the kind of interventions we seek, making it difficult for us to think of social media governance that does more than just reduce the harm we have already witnessed. We need a model of social media governance that seeks to reduce the chances of harm happening in the first place. To do so, we must therefore ask ourselves, what is required to build something new? What is it going to take for us to be able to account for the varied ways in which misogyny and platformized hate arise?

For me, the story starts with this question: to what end are we undertaking social media governance and for whom? Currently, the search for answers to

this question is being undertaken within the confines of the system itself. But in a system where sensitivity, kindness, and thoughtfulness are not inherent to the design, we must ask ourselves, "How much social media governance can actually achieve?" Instead, we rely on established responses that are limited in nature: vacillating between legal responses that are inaccessible and lengthy, and content moderation that is almost impossible to scale and eminently traumatizing, as is the case with human content moderators. The limits of introducing even more technology and artificial intelligence (AI) content moderation, which lack the ability to understand tone and nuance of online harm, have been widely noted. These mechanisms are not ahead of the curve, rather they respond to previously observed harms, making them unable to account for new forms of harms that may arise.

This essay seeks to explore the ways in which feminist care practices and understandings of communities of care can offer us a holistic lens for redesigning social media governance. It looks towards transnational feminists and their work thinking through international solidarity to consider what might a global response to the vagaries of social media governance look like. The great chain of social media governance frequently sees the business of enacting it externalized: precarious, invisible work now relocated to the Global South where the cost of a supposedly more "caring" social media landscape can be forgotten. This feminization of labor – regardless of the identity of those undertaking it – embeds misogyny further within platform governance, protecting some women at the cost of imperiling countless others. This intervention considers how feminist conceptions of care that operate across different identities and modalities can offer us a new imaginary in which to sincerely ground the rethinking of social media governance efforts.

2. What does Social Media Governance Currently Look Like?

Even as they have accumulated immense power while embedding themselves within our day-to-day lives, social media giants have been slow to confront the ways in which their platforms have been weaponized. Despite the rapidly expanding list of issues – from misinformation to targeted content and the political manipulation of social media – governments have also been loath to wade into the quagmire of platform governance, preferring, instead, to rely on social media companies to find a way out of a mess they created. In practice, this has meant that the governance of the online world has largely fallen to platforms who still insist on the neutrality of their services, despite all evidence to the contrary.

As the swell of online violence in recent years has threatened to upend social media platforms, we have seen the proliferation of top-down responses – such as codes of conduct that outline how users are expected to interact with both the platform and one another – as these platforms attempt to mitigate hate speech, disinformation, and aggression. The benevolent paternalism of these male-dominated tech companies, nearly all of which are located in the Global North, has thus far failed to ameliorate the harms experienced by women and members of minority groups.

When the tradeoff has been between freedom of speech and safety, between too many constraints or too few, decisions have clearly trended towards minimizing regulations. There are some notable exceptions to this: the banning of women's nipples on Instagram, for instance, demonstrates the willingness of tech companies to police certain speech over others. Compare this to wider reticence over limiting hate speech and the blind spots of social media governance, and those responsible for setting policies, is blindingly clear.

Technological responses to platform governance run into many of the issues facing AI and machine learning broadly, i.e., questions around the precedent-setting nature of early decisions and the need for consensus around rules and their exceptions. This requires local knowledge to interpret globally set standards and translate their relevance to different spaces; a task that demands immense real-time capacity to interpret conversation for the nuances of intent.

Inherently, these responses are reactive, taking place after the harm itself has already been enacted. Given the directionality of harm, it often means that women and marginalized individuals have to endure hurt before any action is taken, making these spaces unsafe even if some semblance of safety may be returned post hoc.

2.1 The feminization of content moderation

As social media companies have chosen to focus on reacting instead of preventing harm, they have created a model of content moderation that merely transfers harm under the auspices of preventing it. There is no way to describe the current content moderation model without flinching. Severely underpaid individuals, who sit at banks of screens scrolling through flagged posts containing some of the most visceral imagery or phrases fathomable, are repeatedly traumatized by what they are viewing. They lack almost any means of psychological support, and frequently

work in spaces that are themselves unhealthy and pressurized. In pledging to protect online spaces, tech companies have chosen a form of platform governance that merely transfers harms from one part of the system to another.

What is striking about this transference is how, in attempting to tackle the pointedly feminized experience of harm, social media companies have feminized the labor of content moderation itself. Feminized labor refers to both the kind of labor being undertaken – of which irregular conditions of precarity and insecurity, and uncompensated invisible emotional labor are the hallmarks – and the demographic disproportionately employed in such work, namely, women. It is not only women, however, who can experience this labor: under the conditions of globalization such working conditions have increasingly become the norm, especially in recent years as gig work has proliferated. Content moderation has also followed the arc of globalization, with companies like Facebook employing moderators across the globe as they try to deal with the firehose of content spewing forth. The teleology of this process has resulted in content moderation being offshored to the Global South, where moderators take home as little as USD 1.50 an hour in exploitative, sweatshop conditions. Far away from Silicon Valley and Big Tech execs, content moderators in Kenya, for instance, fear the legal consequences of sharing their stories, and are prevented from unionizing – all whilst performing the labor that allows social media platforms like Facebook to claim they are protecting users from harm. Even as this work upholds the continuing functioning of the platforms themselves, it is chronically undervalued – the realities of the human costs of such content moderation left out of discussions on social media governance.

Contextualizing social media governance as part of a global feminization of labor gives us critical insight into the links between informal, feminized labor and the social media governance models that are underpinned by a drive for consumption over care. In many ways, the way content moderation takes place mimics exactly the kind of informal labor that platformization and many Big Tech companies have themselves incentivized and monetized. The layers of harm baked into social media governance are indefensible and cannot be part of a feminist future.

2.2 Why center care?

Concerns around online platforms and the wider state of the internet, have long been topics of discussions amongst feminists. Discourse around building a feminist internet has steadily gained traction as feminists have sought to excavate new potentials for the internet. This intervention reaches for care as a framework and grounding principle that can situate these discussions in parallel with earlier observations. Thinking through and about care is not only useful, it is also necessary given the acute carelessness that threads its way through how social media platforms are both used and governed.

The tentacles of social media governance stretch far, recent discussion of content moderation and its offshoring to countries within the Global South demonstrate how embedded the feminization of labor is within the responses social media companies currently present. This invisibilized, underpaid, and precarious work externalizes the cost of protecting those who can hold or even attempt to hold social media giants accountable. In many ways, it is akin to the globalized chains of care we have observed with domestic work, for instance, where women in the Global North rely on the precarious labor of women from the Global South. Care work has long relied on the disempowerment of women from the Global South in order for career women of the Global North to claim some sense of empowerment under a misogynistic system. Content moderation digitizes this care work; operating on the same logic of relieving some women from patriarchal harm by employing others from the Global South to fill the gap by doing care work in dangerous conditions.

We are stuck in a vicious cycle where we embed misogyny further within platform governance, protecting some women at the cost of imperiling countless others, all whilst social media platforms make claims to undertaking care. Other mechanisms are also insufficient – reliance on legal mechanisms requires that national laws adequately protect women, else they risk replicating the very same misogyny they are meant to be preventing.

3. Feminist Practices of Care

Often diagnosing problems and critiquing a system is the easy part: infinitely trickier is figuring out what it might take to build something new. In dissecting the different ways harm emanates from social media platforms, this essay has sought to identify the ways in which our current platform governance models fall short,

and the manner in which feminist interventions may address these shortcomings. Feminists have long advocated for the power and potency of care, advancing a feminist ethics of care focused on resisting the 'injustice inherent in patriarchy'. More recently, the discussion has tilted towards an intersectional feminist care ethic that properly accounts for racialized care, and returns in part to initial ideas put forth by the Combahee River Collective on community-rooted love. Centering care forces us to confront the great chain of platform governance, and ensures a holistic effort to rethinking how the business of social media governance ought to happen — within this recalibration the outsourcing of digital care work becomes untenable.

Care as a stubbornly gendered concept is uniquely placed to help us think through both how sex-based hate needs to be tackled, and also confront the repeated sexism and gendered nature of current responses. Introducing feminist conceptions of care can help bridge the technologically-minded concerns of digital rights thinking on platform governance, together with the social justice focus of research on harms in the digital space. Further, if we do see current models of platform governance as some form of benevolent paternalism, then it is only right we seek solace in feminist conceptualizations to avoid our thinking being subsumed or dictated by other patriarchal institutions.

Feminist practices of care offer us new potentials and allow us to set the stall out for what is truly necessary, whilst equipping us with the knowledge of how to collectively organize towards such goals. With this in hand, we can open up our horizons to consider alternative approaches to social media governance and bring together different feminist critiques and considerations of social media platforms.

3.1 Communities of care

Caretaking is an act of revolution. In the face of a deeply uncaring world, insisting on giving care to one another is a refusal of the status quo. Feminists, especially feminists of color, have long known that care is vital to sustaining the movement: without solace feminists wage a lonely and exhausting fight. Many social media platforms claim to be about the creation of community, but their ethos, especially when contrasted with a vision of mutually beneficial caring relationships, rings hollow. Social media platforms do not incentivize caring interactions even as they may sometimes punish those who cause harm. Indeed, the bar for removal from these so-called online communities is often very high and at times unenforceable

given how easy they are to circumvent. This is neglectful.

Big Tech companies which curate online spaces ignore the ways in which the online world is intrinsically bound up in our offline world. The border between these two worlds is highly permeable: harm and hatred diffuse with ease across them. It is precisely because of this that we should draw inspiration from communities that have found ways to shelter one another from harm, even in the face of immense difficulty. Feminist communities of care are built on the sharing of both lives and politics, when living aligns with our values and ethics. Though it is not feasible to ask that everyone on social media platforms share the very same principles, it is time to ask that we look towards collective norms and establish expectations around the ways in which we wish to hold each other accountable. In so many ways, social media is built, at least, on the notion of sharing our lives — even if it is posed or calculated, but the sharing of politics is almost entirely absent. To be clear this does not mean being politically aligned but rather a common understanding of ways of being on the internet. In order for harms to be prevented, individuals must be incentivized to be better in the first place rather than, at most, met with lackluster punishment after the fact.

What we have seen in the failures of social media governance are the pitfalls of sharing our lives whilst sharing little else. Clearly, this is not sustainable: the implosion of social media platforms and the resulting struggle to moderate content is a direct result of communities that are not built on care. The surprise at this reality tells us more about the creators than the products they created. If we want something different than the status quo we need to know what kind of community is tenable, and ask instead for that.

4. Towards a New Vision: Transnational Feminist Movements

Knowing what we want is one thing, envisaging the shape it will take is another entirely. Given the reach of social media platforms, social media governance must come with a degree of flexibility. Setting out the desire for a social media governance rooted in care requires room for contextual considerations. How then might we have a shared vision that allows us to organize across space without removing the agency of difference?

Transnational feminist movements have been asking this question since the movement's nascence. In the process of grappling with globalization, transnational feminists have strived to identify the key spaces of overlap between different kinds

of feminisms. Often, they have done so in the absence of clear leadership, seeking instead to create free autonomous spaces in which impetus comes from collective motivation. This method has leant the movement flexibility and adaptability across a varying landscape of need and progress. Crucially, it has also allowed the movement to reject or work in contrast against "colonial and colonizing, liberal and neoliberal, Western paradigms, narratives, processes, methodologies, practices, and application" that so often infiltrate attempts to push back against the status quo. In practice, this translates to a sharing of space between the global and the local. Transnational feminism works to bring together all the different strands of conversation – bridging local issues with the global landscape in which many of these digital spheres operate. Social media companies have tried to soothe concerns in the Global North by relocating content moderation further afield, distancing the human costs of content moderation from the spaces in which Big Tech is most likely to be held to account. Within a transnational feminist approach, the specious nature of such smoke and mirrors becomes a collective concern: a localized harm contextualized as part of an international issue.

With issues as complex as social media governance, any viable alternative must operate across internal social, cultural, and structural differences. Navigating such waters is treacherous, and transnational feminism also contains important warnings about how to plot a course between the global nature of social media governance and its localized impacts. In particular, it warns us against the turn towards nation states that appears especially tempting as the lesser of two evils. Platforms have tried as much as possible to shift the onus of harm onto those who perpetrate it – ignoring their abetting of such harm. Attempts to shift away from this dynamic have resulted in putting too much power into the hands of governments in the hopes they will be more motivated to act. This reification of the nation seems the natural refuge and response to desires to constrain multinational tech companies, but it fails to reach beyond the limitations that create the conditions for harm. Transnational feminism offers genealogies of organization that emphasize instead the common cause, inviting us to consider how we might reject all tyrannies as we mobilize against specific ones.

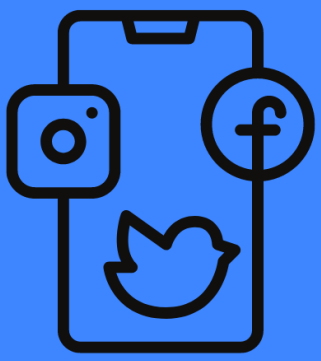
Perhaps, most important is the way transnational feminism empowers us to reject the false choices we are so often presented with under the status quo; to push back against the decision-making calculus that says nation or social media company, harm to you or harm to others. These are not decisions: they are ultimatums.

Transnational feminism exposes the way such choices are merely constructions of power that will only reenforce existing power dynamics.

4.1 New visions of social media governance

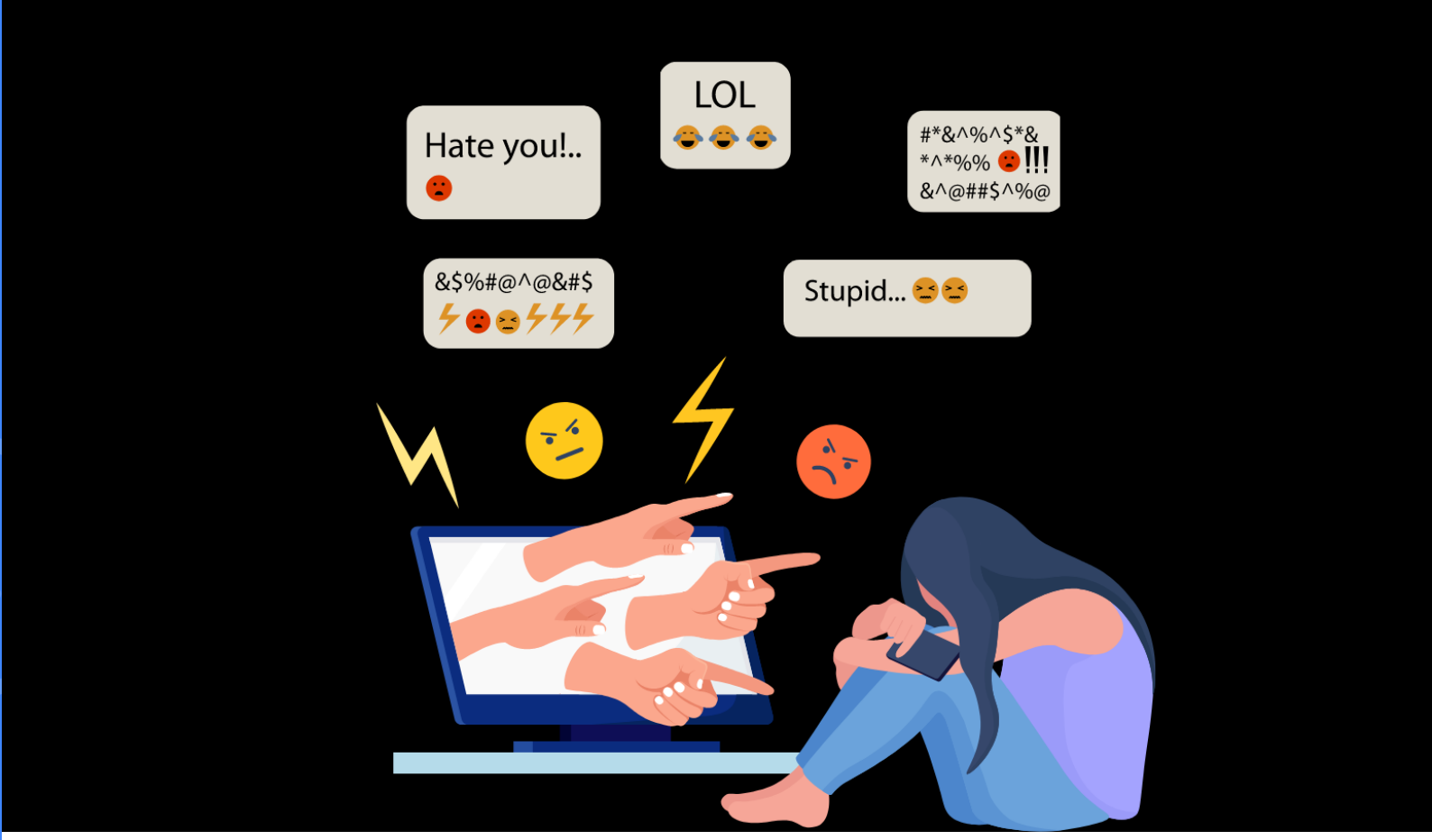
We live in a world that reels from the consequences of uncaring decisions: harm in the digital world stems from this negligent attitude to harm. So, what story should we tell ourselves instead? If feminist care practices tell us we must center harm prevention, if it problematizes the myopia that dominates social media governance, then transnational feminist discussions tell us how we can act on this knowledge and awareness to produce outcomes that make the system anew.

In some ways, these ideas and asks may seem almost impossible: perhaps, they are. But only when we put forth what we want can we then confront the vast sea of indifference that must be traversed to get to our visions. We have to know how little those holding the levers of power are willing to meet us where we wish to be, whilst we are instead expected to meet them where they are. Rejecting the primacy of harm in the offline world is bound up in the rejection of harm in the digital world. Big Tech companies and governments alike have refused to ensure this outcome, but within our communities of care we already possess the tools necessary to fashion a social media governance capable of compassion. Now, more than ever, it is time for us to use them.



IV

Legal Responses to Online Gendered Violence



The Responsibility Gap and Online Misogyny: Beyond Misogyny, through Violence, Towards (Legal) Safety?

KIM BARKER[1]

Abstract

Online misogyny is one of the many forms of online violence against women (OVAW). It is a pernicious, silencing phenomenon that is designed to subjugate women online, silence them, and limit their participatory rights. Legal systems fail to define it, and frequently exclude gender (and/or misogyny) from their hate crime frameworks, rendering women in the online sphere the lowest priority for protection. Responsibility gaps persist beyond the law too, and the fragmented

[1] Kim Barker is a senior lecturer in Law at Open University, UK, and Director of Europe's first Observatory on Online Violence Against Women. She is an expert in online violence against women, including online hate, online misogyny, and the legal regulation of social media. Her wider expertise focuses on the regulation of social and interactive media platforms, including online safety, & online policies.

approaches of policy bodies, international organizations, online platforms, and the law compound this problem. This essay will discuss these issues, advocating for a holistic, 'joined up' approach to tackling misogyny, its impact and harms, and suggest approaches to tackle online safety and the current responsibility gap.

1. Introduction: Online Misogyny – Deliberate, Not Accidental

Online misogyny, defined here as, "the manifestation of hostility towards women because they are women…communicated through online platforms, particularly social media, and participatory environments" is no longer an emerging issue. Rather, it is a pressing issue, affecting significant proportions of women and girls online, every day in their interactions. The scale and breadth of the issue is stark, with some recent studies suggesting that over 85% of women have experienced some form of harassment or online violence (without accounting for the known underreporting and the incidences that are not recorded). Other indicators are equally dismal, with only 15% (out of a total of 1002 surveyed) of women and girls believing social media is a safe space for them in 2022. Given the reliance – increased through the Covid-19 pandemic – placed on online interactions and engagement, this is a severe indicator that being a woman online is not safe, nor without harm.

Alongside the harassment, abuse, gender-based violence, and hostility faced by all women, some smaller groups of women report even more shocking experiences. Almost 73% of women journalists, for instance, report being subjected to online violence including physical and sexual threats of violence. This is closely mirrored by the experiences of other women who are prominent in public life, including online publics, but particularly in politics. In social media – for instance, Twitter is widely recognized as being a toxic space for women, especially women in politics Similarly, Facebook is regarded as one of the least safe social media sites for women.

It is far from surprising that the phenomena of online misogyny and online violence against women (OVAW) have developed alongside online interactivity. These challenges are long-standing, albeit now prominent and recognized, they are still left under-addressed. For instance, the UN's report of the Special Rapporteur on violence against women, its causes, and consequences in 2018 highlighted the benefits and the drawbacks of information communications technology (ICT) for women – commenting specifically on the misogynistic and violent content women encounter online with regularity. Despite this, little has changed in the

ways to address, and reactions to online misogynistic abuse as a form of OVAW. The behavior and acts of online misogyny, and the broader actions amounting to OVAW are not accidental, but they are established methods of silencing women online. They are also recognized mechanisms of encouraging women to 'stay away' or 'stay silent' online. Depressingly, while methods and mechanisms of online misogyny and OVAW are well-recognized and developed, the responses to them are not (yet).

While the silencing of women, and the recognized chilling effects of OVAW and especially misogynistic abuse serve to make online environments increasingly hostile to women, this is not their only impact. There are much broader, cumulative impacts on women, and also women's expression and participatory rights. These are all severely affected where OVAW and misogynistic abuse are allowed to permeate.

This essay therefore explores the phenomenon of online misogynistic abuse and its impacts before introducing a typology of online harms caused by such forms of online violence. The discussion then turns to the challenges posed by fragmented responses to online violence and misogynistic abuse, and the resulting responsibility gap. Finally, this essay outlines some potential resolutions and mechanisms that may mark a step change towards addressing this phenomenon and making the internet a safer place for all women. In seeking to highlight the deliberate (rather than accidental) nature of OVAW and online misogynistic abuse, it is not only important to identify, recognize, and respond to the harms caused, but also outline potential resolutions and responses to make meaningful change. It is to this discussion that the essay now turns.

2. The Harm is in the Harm

To capture text-based abuse specifically, and separately, is an important step towards shifting the narrative around online harms. There is an emerging awareness (though perhaps not yet fully understood) on the variety and widespread impact of harms caused by online misogyny and OVAW. It is therefore particularly important that in calling for action to address online misogyny and OVAW and make the internet a safer place for women, the harms are captured together in their breadth and severity. Without a clear and well-developed typology for harms, meaningful mechanisms to address the phenomenon of OVAW will be difficult to implement.

As such, through case studies, first-hand accounts, bystander revelations, judicial commentary, and political commentary, the following typology of harms developed by Barker and Jurasz has identified and categorized twelve types of harms caused by online misogyny and OVAW.

Barker and Jurasz's Typology of Online Harms

Type of Harm	Impact on Victim
Residential	Additional security measures, expense incurred in sourcing alternative accommodation for family members
Economic	Time and/or money spent on removing public presence and / or information available online
Emotional	Feelings of fear, panic, anxiety, possible personality changes
Physical	Online abuse, violent threats of physical and / or sexual harm
Social	"Always connected" invasiveness in family and non-work environments
Reputational	Frequent exposure to abuse reinforcing demeaning ideas of sexuality
Intersectional	Mixed abusive content, commenting on multiple characteristics of identity e.g., race and religion, or race and gender
Democratic / Participatory	Wide-ranging impacts of women removing themselves from public life and / or politics due to abuse
Disablist	Abuse based on an identifiable disability of victim; death and / or sexual violence threats
Ageist	Abuse based on identifiable age profile of victim; mockery of too old and / or too young, mixed with remarks on competence
Psychological	NB: pervasive harm arising because of the breadth of impact of online misogyny and OVAW e.g., other categories of harm brought together leading to severe, repeated, long-standing psychological impacts
Emotional	NB: (distinct from above) pervasive harm arising because of the breadth of impact of online misogyny and OVAW e.g., other categories of harm brought together leading to severe, repeated, long-standing emotional impacts

All of these 12 harms have been identified from first-hand online misogynistic abuse, including harms to society, to participation, and to democracy. Barker and Jurasz's typology of online harms is by no means exhaustive, but it does capture the emotional, psychological, economic, and physical harms imposed on the victims by these forms of abuse. This typology was created by capturing and categorizing these harms from victim statements, reported cases, and emerging literature in the hopes that courts, policing bodies, and legal systems can conceptualize the impacts more widely and respond, if not appropriately, then symbolically to the phenomenon. Above all else, categorizing the harms caused by behaviors which are not (always) recognized as problematic by enforcement and judicial bodies, or by other stakeholders (including platforms offers an opportunity to trigger proactive responses in tackling responsibilities for online misogyny and OVAW.

3. From the 'Fragmentation Problem' to the 'Responsibility Gap'?

Much like the confusion surrounding terminology and its interchangeable use to discuss online misogyny and/or OVAW, there is an equivalent lack of clarity on the effective approaches to tackle these behaviors but also to mitigate the impact and severity of the resulting harms. The typology of harms (discussed above), together with the statistical evidence outlining the spread of the phenomena, indicates that the time for action is long overdue. The behaviors and problems of online misogyny within OVAW are relatively well-documented and are no longer issues affecting only the minority groups. Quite the opposite, there are large-scale, empirical studies documenting the scale and breadth of the issues online. This is yet to translate into effective responses, and has left a stark gap. Amongst discussions of definitions, categorizations, and labeling of the problem, this 'gap' is what I have termed the 'responsibility gap'. Online misogyny, and OVAW, fall often at the intersection of law, policy, and regulatory response. This renders the phenomenon a problem for everyone, but a problem for women in particular.

The textual aspects of online misogyny are often overshadowed by their image-based equivalents, by differing branches of law e.g., communications law, internet laws, hate crime, and human rights (in the context of speech protections). Interestingly, the image-based elements of online misogyny tend to have much faster legislative responses to capture – criminally – these behaviors. In England for example, image-based abuse was captured in legislation back in 2015. No

such developments have been forthcoming for the equivalent text-based sexual abuse – despite ongoing discussions in the UK surrounding the Online Safety Bill, these text-based behaviors and forms of OVAW still do not feature in the draft provisions at the time of writing. Despite the hype therefore, frequent discussions with broad brush terms like 'online safety' and 'online harms' forget about the gendered dimensions and overlook the behaviors which they really ought to capture within their provisions to protect women online. There is little joined-up (legal) thinking on display here, but it extends far beyond the legal silos of specific branches of law. There is a lack of cohesion evident across different policy, treaty, and legal levels too – something Barker and Jurasz described in 2020 as «the fragmentation proble"'.

This responsibility gap has become increasingly evident in recent years, amid discussions of legal reform in different regions and countries. Discussions of, for example: e-Safety and the online safety laws in Australia, online hate speech in Germany, online harmful digital communications in New Zealand, online harms (and later safety) in the United Kingdom, communications decency legislation in the USA, sexist hate speech in Japan, online harmful content in Canada, online falsehoods in Singapore, and computer misuse in Kenya are just a handful of examples focusing on national initiatives to address elements of online safety, online harm, computer misuse and online hate speech. These are not the only initiatives triggered by the rise in online hostility and vitriol, regional measures have also come to the fore such as the Digital Services Act 2020 at the European level. While there has been activity – and that is laudable – to reform the law to capture some of the behaviors encapsulated within online misogyny and OVAW, very few of these legislative proposals specifically include OVAW. In the UK's much-discussed Online Safety Bill for example, there are 215 pages of proposed legislation and not a single mention of women.

Despite the flurry of law reform activity across a spread of countries, the predominant inaction on developing responses aimed at closing the responsibility gap for tackling online misogyny and OVAW seems to have rested on oft-claimed 'forthcoming' law reforms – as though enacting legislation can operate as a silver-bullet and solve these problems alone. This is optimistic at best, and wildly unfeasible when grounded in reality. Disappointingly, the lack of effective response(s) is not limited to one sector, nor to one legal jurisdiction. It is the result of a distinct lack of 'joined up' thinking when it comes to online misogyny and OVAW.

Frequently the responsibility gap (flowing from the fragmentation problem) arises because of the disconnections that are prevalent at domestic levels, with criminal law regimes often not speaking to the communications law regimes, or to the victims' support services. There is all too often a further disconnect between the 'legal' expectations and responsibilities, and the 'practical' responsibilities that can (but are not) asked of platforms and providers. For instance, while there is a symbolic benefit in having the law 'recognize' a particular societal challenge or behavior by enacting legislation to, for example, criminalize the behavior and / or harm that sets expectations that the law alone will 'address' the acts.

4. Meaningful Responses?

While there are documented responses at international and regional levels, there is still much to be done to address online misogyny. It is not a problem solely resolvable by capturing online hatred, online misogyny and OVAW in policy and legal documents. Similarly, simply capturing these behaviors in new legislation is not a solution. Making law is not a catch-all response that will result in meaningful change; it must be understood, applied, enforced, scrutinized, and examined for effectiveness. Measures well beyond legislation – despite the symbolic potential of law in itself – are required. There must be less siloed thought, and much greater holistic consideration given to the issues. This must necessarily involve a range of stakeholders – including the platforms themselves – to move away from ideas and well-versed statements that platforms are mere conduits[2] and simply "host" the content. This shirking of responsibility by committed actors cannot be allowed to continue.

The responsibility for addressing online misogyny and OVAW no longer falls solely on one entity or one group, and there is much consideration to be given to the broader issues. Law alone cannot be left to tackle the responsibility gap. It is time the onus shifted from legal responses to the acts of the platforms themselves. Platforms have to be actively involved rather than paying lip-service to issues affecting women on their sites. This will require not only willing on their behalf, but also resourcing and engagement in the broader discourses. It may best be described as the need to act to retain half of their customer base, not least because women are not a minority, and online misogyny / OVAW is not a

[2] Although the Digital Services Act 2020 will retain this status, it will impose obligations to ensure platforms are acting to preserve the safety of an online environment.

minority issue. Platforms can no longer assume their status as 'hosts' of content will protect them, and allow the responsibility gap to fester. They have to be proactive partners (initially) and leaders (eventually) to tackle online misogyny and OVAW; protecting their users, and ensuring the safety of their online communities of women in order to benefit from the protection they have historically enjoyed as platforms. Platforms need better understanding of methods used by perpetrators of online misogyny and OVAW to avoid detection by content moderation systems. For example, deleting letters from commonly used abusive words is a simple but effective method used to facilitate abusive posts but avoid removals; action on this is required to address the significant volumes of content which are harmful, but not necessarily illegal. Platforms also need to ensure that their acceptable use policies align to relevant local and jurisdictional legal standards. Too frequently, terms and conditions are not enforceable because there is a difference between platform standards, and legal standards in a particular country.

The flaw in these calls for greater (and better) responsibility and action from platforms is that the business models of such platforms incentivizes not tackling online misogyny and OVAW because fostering hostility feeds further content. So even though online misogyny and OVAW encourage women to make themselves less visible, platforms still simultaneously promote the algorithmic control that prioritizes content which violates safety, and human rights standards. The harms to women are caused by valuable content, creating a self-defeating level of self-interest for platforms based on their business models. Online platforms have, for too long, enjoyed a comfortable status where they have placed themselves behind liability (and responsibility) shields, and placed the onus on users to change their behaviors and control what they post.

Unlike other facets of the interactive entertainment industry, such as copyright and intellectual property law in the gaming industry, social media platforms have not sought to introduce effective measures to self-regulate themselves. They have much to learn from other, long-standing sectors, especially the gaming industry which – voluntarily – sought to self-regulate its content (and stave off legal regulation) through initiatives such as PEGI. Why have social media platforms not acted in similar ways? Arguably, there has been no incentive to do so. It is the hope that proposed legal reforms and the global onslaught at local legal levels can – at the very least – act as a trigger to spark social media platforms to protect their self-interest and act, if for no other reason than to protect their liability-shielded status.

Finally, when considering how 'best' to tackle online misogyny and OVAW, input is required from a wide range of stakeholders, including platforms, and enforcement bodies. It is a multifaceted problem and requires holistic, joined-up, multifaceted responses. For instance, public health, and educational approaches are required in addition to enforcement, monitoring, oversight, and legal consequences. Triggers for societal change to challenge the underlying attitudes leading to misogyny are also required both online and offline. The multifaceted harms must all trigger action by different sectors. Tackling one area and one harm in isolation is unlikely to be enough given the scale of the challenge. But the symbolic power of the law (and international human rights standards) cannot be overlooked either. A collective and comprehensive set of responses is required. To do otherwise is to move away from, rather than towards online safety. Working in combination to close the responsibility gap is what is required to tackle the problem, to redress the multiple and pernicious harms, and to make the online sphere a safe(r) place for women. Women deserve better. It is time for platforms to tackle the responsibility gap, and protect women online.

Internet Safety, for Whom? South Korea's Post-Nth Room Platform Governance

ESTHER LEE[1]

Abstract

The case of the nth room supposedly marked a point of reckoning for the South Korean digital culture; yet just two years later, the post-nth room provisions for internet safety face the risk of scale-back. This piece reflects upon the debates against platform regulation, against the backdrop of the historical and systematic dismissal of gendered violence, post-colonial anxieties, and the limits of prevention-cum-criminalization – and asks what kinds of safety, and for whom, these debates defend.

---

[1] Based in Seoul, Esther Lee's work focuses on the queer, feminist reimagining of digital spaces, against the epidemic-scale proliferations of non-consensual intimate images (NCII) across East Asia. She is currently a Program Officer at the Open Society Foundations, where she oversees the Asia Pacific grants portfolios on youth and LGBTQI+.

1. Introduction

The May 2020 revisions to South Korea's Telecommunication Business Act attempted to shatter the pervasive assumption of platforms as passive intermediaries and uninhibited open spaces for content and information exchange. As part of a wider set of reactionary legislations to the 'Nth Rooms' – a term coined for both the encrypted telegram chatrooms that housed an archive of extorted content; and the national uncovering of a cartelized infrastructure for digital gendered and sexual violence – the amended Act offered a belated recognition to the epidemic consumption and circulation of non-consensual intimate images (NCII) by the networked publics. In tandem, the revisions expanded the regulatory obligations of domestic social media and open community forums to directly filter, report, and takedown exploitative materials as defined by the expanded provisions around sexual crimes.

Two years later, this imperfect, yet landmark, revisions face an imminent threat of scale-back under the 2022-elected Conservative administration. Embedded as part of his campaign pledge to dissolve the Ministry of Gender Equality and Family, and the People Power Party's (PPP) election strategy to pander to populist misogyny, the President rejected the expanded liability for social media governance under the framework of anti-censorship and a crusade for data privacy. In the context of South Korea, prescriptive changes are not solely hinged on the power of the Executive. However, this pretense for open internet – in which the parameters of a South Korean digital citizenship are defined by the safeguarding of a hegemonic masculinist playground – continues to push the fallacy of gender-inclusive online publics as being mutually exclusive from the aspirations for internet freedom.

And the roll-back of the limited protections established post-Nth Rooms has already begun. In May 2022, 17 civil society representatives of the 22-member Digital Sex Crimes Task Force submitted their resignations in protest against the PPP-appointed Ministry of Justice's reshuffle-cum-termination of the Task Force chair, prosecutor Seo Ji-hyun. As one of the few semi-independent research and oversight committees to elevate the lived experiences of women, with a focus on implementing survivor-centered protections within the judiciary, the Task Force had made a series of recommendations for clear statutory provisions that counter the current vagueness in defining NCII, and through this framework, shift the burden of content moderation directly to investigative agencies and platforms. With the de-facto dissolution of the Task Force – compounded with the general dismissal of

systematic online and offline gendered violence by this administration – the case of the Nth Rooms will, perhaps, remain as a heinous but episodic incident in the South Korean national consciousness.

2. The Kitten Takedown: The Privileging of Misogyny

In December 2021, the amendments to the Telecommunications Business Act went into effect. Perhaps, the most contentiously challenged of the series of revisions coined under the 'Anti-Nth Room Bill', Article 22-5 mandates that providers of cloud storage and/or value-added services (this being, any information-communication outside of the standard telephone network) with net annual sales of 1 billion won and more, or those with a daily traffic of 100,000 users or more, to take active measures to "…prevent the circulation of (illegally) filmed materials…" as defined by Article 14 of the Act on Special Cases concerning the Punishment of Sexual Crimes.

Shortly, domestic providers of mobile messaging apps, live streaming services, and user-generated content platforms, amongst others, announced the implementation of technological and administrative measures to identify NCII; and prevent distribution on both search engines and public community chatrooms. Naver Corporation, a South Korean search engine conglomerate with APAC and global subsidiaries in content development, deep-tech, and data infrastructure research and development, reaffirmed their commitment to their upgraded artificial intelligence (AI) image monitoring technology, X-eye 2.0, in the pursuit of a safer internet. According to Naver, X-eye 2.0 references its categorized archive of target content to filter and restrict access to still and moving images, compressed files, and videos, the latter through the processes of analyzing extracted frames into categories of 'normal, obscene, adult, or pornographic/indecent'. In tandem, Naver committed to more integrated user-response services, to ensure that reporting for NCII directly results in monitoring and takedowns, and the so-called DNA of the reported content can be used to cast a wider net for filtering. This was expected to remove the burden of self-identification and tracking, and making multiple reports for the same or similar content, from the oversight of victims of NCII, and more generally, the platform's users.

These expanded boundaries for regulation drew two distinct critiques. First, online communities with dominant male user bases – and the sympathetic media – denounced the accuracy of the filtering technologies. Screenshots of

allegedly flagged and blurred out images of kittens by the AI filters of open community chatrooms rapidly spread; and fueled fear for the regulation to extend blanket criminalization on any and all online activities. Second, the framework of intermediary liability and filtering soon became akin to state-sanctioned censorship. This argued that the Anti-Nth Room Bill, through the mandate of enforcing corporation's data reporting, and cooperation with the Korea Communications Standards Commission as deemed necessary along the provisions for prevention of sexual crime, is a direct breach of people's rights to privacy in terms of communication, freedom of expression, and the rights to information. The same argument warned that revisions would easily and inevitably widen its jurisprudence to apply the same standards of reporting in private chatrooms.

At the height of the presidential election campaigns of December 2021, these critiques were coopted as campaign promises. Citing Article 18 of the Constitution ("The privacy of correspondence of no citizen shall be infringed"), the then-PPP candidate Yoon Seok Yeol labeled the revisions as an extremist response that sought to illegalize the digital citizenship of the innocent majority. In the same Facebook post, Yoon asked how a country could be considered free when photos of kittens and family members were all subject to surveillance.

Contrary to this outcry, there are hard limits to the revisions' enforcement, and the extent to which platforms are held liable. Within South Korean legal precedence, based on a separate 2020 Supreme Court ruling, open chatrooms fall under the public forum doctrine, which allows regulation of content in the collective best interests; and the revisions are limited to application on these public spheres. Meaning, the mandates cannot enforce filtering to contents distributed through private forums or exchanges, including those via emails, closed or membership-only communities and blogs, and direct user-to-user chatrooms.

Furthermore, the database from which the filtering AI references the coding parameters of its flagging, is said to be built from information on illicit materials compiled by joint cybercrime investigative units and victim-support agencies, as monitored by the police, individual survivors, and/or independent women and children's rights oversight committees. Evidently, the PPP and the conservative media have pandered, intentionally or otherwise, to the campaign of disinformation, that has conflated algorithms in greater protection for women's digital – and, the full exercise of fundamental rights – with the façade of privacy

violations of the Anti-Nth Room Bill. The filtering, however, has implemented an approximately 10-second delay within which all contents are analyzed upon the initial uploading, with a notification banner: "Per the Business Telecommunication Act, the Korea Communication Standards Commissions is reviewing whether the (uploaded content) is classified as an (illegally) filmed material."

On the other hand, the constitutional complaints filed against the revisions to the Telecommunication Business Act, by a wide spectrum of far-right pundits to civil society-led internet watchdogs, warn of possible revocation from the intermediary liability principles: mainly, that robust and democratic online public fora are made specifically possible through the sheltering of platforms from direct monitoring obligations, and assumptions of their lack of agency in the development and distribution of information. Under the principles, as the revisions can impose limited dues, demands, and collections of penalties on providers should they fail to take measures and commit to cooperation with agencies in the prevention of NCII, the intermediaries will likely resort to expanded and overly-broad vetting. And where the burden of monitoring third-party generated contents outweighs the profits for operations, platforms will become heavily restrictive, or face an inevitable closure. The prescriptive rationale behind these principles also believe that liability-protected platforms encourage economic growth – and the removal of these protections will ultimately hamper innovation, and fundamentally stunt the internet's transformative potentials in providing equal voice regardless of gender, class, age, and status.

These debates around regulation must be nuanced against the democratic governance capacities of states – and in the historical context of South Korea, they must recognize and redress the traumas of colonial and military regimes, in which state-sanctioned censorship was systematically deployed to crackdown on civic dissent. But the internet is not autonomous to, or resistant to, the offline gendered socio-economic and political organizing of South Korea. The conflation in equating protection-cum-regulation as censorship invisibilizes the post-colonial legacies of South Korea's gendered nationalism, and in doing so, seeks to normalize platforms as a safe harbor for misogyny. Rather, its digital infrastructure and culture are both shaped by legal affordances, in which gendered and sexual violence are understood within the bounds of the private – and thus, beyond the scope of institutional or structural obligations of the state. While this trajectory continues to change, albeit slowly, the legal acceptability discourse of NCII still very much

hinges on the victims' efforts to be self-aware and self-censor, and whose proof of injury is determined by the historically skeptical and male-dominant enforcement agencies. In this context, legal and algorithmic governance of the homosocial digital culture – constructed specifically in the absence of women's unburdened consent, desire, and participation – are easily dismissed as excessive interruptions to a normalized collective behavior, and as interventions that specifically disenfranchise men.

Likewise, injecting constitutional provisions on the freedom of expression and rights to privacy against the Anti-Nth Room Bill is paradoxical. It attempts to delegitimize the rights to safeguard against commodified and non-consensual surveillance on bodies, sex, and sexualities, as inevitable side effects of a free internet. This also dismisses the grievous and continuous harms of NCII on its victims, especially where the online-offline distinctions are difficult to define in South Korea's techno-capitalist present and future. With barriers to comprehensive safety nets, victims – disproportionately, those with gendered and queered bodies – bear the burden of limiting their online presence, and suffer the loss of socio-political, economic, and cultural capital from self- or community-driven othering. As Mariana Valente notes, misogyny is a form of censorship, and NCII, regardless of enforcing intermediary liability, already gatekeeps platform-generated economic growth, and civic participation to the beneficiaries of misogyny.

3. Post-Nth Room Platform Governance

The operationalizing of the Nth Rooms was neither episodic, nor amateur. It systemized production, circulation, and consumption of NCII through leveraging communities of networked platforms, weak jurisprudence against domestic and foreign-registered servers, the affordances of encryption, crypto-currency, and both private and public user-to-user interactions. More insidious was the almost automatized re-branding and replication of similar NCII-dedicated spaces – and its operational support mechanisms – to exist in perpetuity, despite perpetrator-centered crackdowns.

As documented closely by Human Rights Watch and the Korean Sexual Violence Relief Centre, the ambiguous legislations around digital gendered and sexual violence, and a lack of comprehensive access to legal and civil support or takedown services, make it impossible for victims to forge a path forward – and reclaim the fullest exercise of their rights to expression, privacy, and participation.

Then, what national frameworks can support feminist conceptions of platform standard-building, for both free and gender-inclusive internet?

3.I Protection of the anti-Nth Room Bill

The 2020 revisions have expanded the legal vocabularies around consumption, distribution, and production of NCII within the boundaries of punishable sexual crimes – and allowed for the increasing complexities on the ambit of NCII, including deepfakes and "distribution of false video products". However, technology-mediated gendered and sexual violence is still not framed under the provisions of human rights. The revisions must complement the Anti-discrimination Bill (first of its drafts conceived in 1997, yet whose passing still remains in limbo), within which gender is recognized as a marginalized demographic, and sexual harassment as a form of discrimination; and continue to coalesce around the intersectional expansion of, and civic mobilizing on, constitutional and other legal protections on the rights to privacy and rights to digital access. Furthermore, the revisions must continue to resolve the ambiguities around NCII with clear scope for defining consent, including its revocability and duress.

In tandem, platforms should be incentivized to expand a systematic duty of care approach: A legal standard for implementing a comprehensive plan for proactive monitoring and reactive takedowns, in collaboration with, and under the joint guidance of, an independent regulator. This approach can co-exist with liability principles, and supplement protections of censorship and privacy. However, a push for platforms to adopt such an approach, without the full recognition of NCII as violence and violation of (in the context of South Korea) Constitutional rights, will strengthen socio-cultural and legal ambiguities on: what is considered as necessarily harmful, and who determines the degrees of harm in the context of NCII and platform-mediated misogyny. Feminists note that the Anti-Nth Room Bill fulfills the minimum requirements in gender-inclusive platform governance, with incentives for holding the NCII market accountable. A scale-back of the bill as proposed by the PPP, without any alternative strategies for monitoring and removal, would reaffirm the failed deterrence policy that seeks to solely criminalize the individual perpetrator, which has allowed for the cases of the Nth Rooms to fester: that digital gendered and sexual violence is illegal in technicality, but tolerated due to the high cost of enforcing the law. Realistically, where expansion of the scope of the bill, along with transformative legal reforms on gender justice

are not possible, the defense of the current bill by the National Assembly seems to be the current viable strategy.

3.2 Integrated victim support and independent oversight committee

Statutory support for victims of NCII is legally embedded through the acts in prevention and punishment of sexual violence, and legal aid; and entitles them to legal, economic, medical, and security (for data protection, personal safety, and residential relocation) support. However, the ministries authorized to provide such support – including the Ministry of Justice, the National Police Agency, and the Ministry of Gender Equality and Family – operate with their own distinct processes for registering victim status and assistance applications.

The Digital Sex Crimes Task Force's Expert Committee documents that with no one-stop window for support, individuals must, for example, contact "…the Victim Support Office of the Public Prosecutor's Office for State for (legal and attorney representation), the Crime Victim Support Center for financial support, the Smile Center for psychological support, and the Korean Legal Aid Corporation for legal advice…then go through separate application processes for each purpose". This fragmented system causes bureaucratic barriers to access and an incoherent information flow – in addition to the prolonged procedural delays in criminal proceedings, the lack of civil remedies, and the overall minimization of harm by the police and the judiciary. Victims are stuck in a cycle of abuse, bearing the onerous and full burden of proving their injury, and claiming recognition and their legally protected statutory support.

The Committee recommends establishing a unified application window, where victims can access comprehensive information for support at the initial reporting of harm. This will require training for the police and the in-lieu civil society hotlines or crisis centers to be equipped with mechanisms for immediate and longer-term assistance, external to the timelines established through criminal or civil proceedings. This integrated window overseen by an inter-ministerial body can provide direct and comprehensive channels of information to victims, while advocating on their behalf for relevant services. This is an active secondary prevention measure.

Accountability checks on integrated victim support and the incentivization of platforms to adopt a duty of care plan can be enforced through an independent regulatory body, which oversees investigative and platform compliance for

monitoring and deletion of NCII, and victim redress. This body should establish the parameters of obligations (and prevent loopholes) in, first, expanding the clear language and tools within the judiciary on NCII; and concurrently, working with platforms on the designing of transparent filtering algorithms against NCII, built on digital gendered experiences and open to civil society audit; and deploying of protective first-step actions, with immediate or expeditious removals.

4. Queering Platform Governance

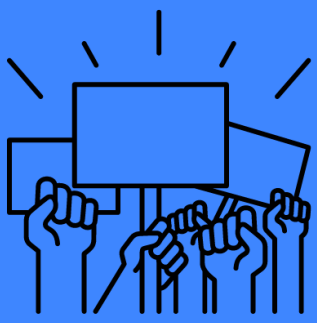
Regulations displaced from a rights framework can be weaponized. A recent example of South Korea's pandemic governance – including surveillance of Covid-19 outbreak patterns through geo-tracking, and social media and financial data mapping – under the guise of transparency and the public's rights to information, shows how easily communities in the margins can be scapegoated. In May 2020, exacerbated by the media and the networked publics, contact tracing regulations triggered a series of doxxing, mass outings, and the normalization of homophobia.

Furthermore, where sex work continues to be surveilled and criminalized, and gender is mapped on a militarized binary of male/female, feminists must reckon with the existing limits of legal recognition for gendered experiences in the digital present and future. And whether platform governance can account for the internet as a primary and queered arena for exploration, community-building, and bodily and sexual liberation. This requires a careful, and experience-based, exmaination of socio-cultural and legal vocabularies that conflate NCII with digitized sex work, labor, and queerness.

At this juncture of impasse, feminist reimagination of the digital public sphere and standard-building must continue to debunk the cooptation of rights to privacy and protection against surveillance, as being exclusive from platform accountability for gender-based hate. Combined efforts of civil society, legal agencies, and independent oversight mechanisms must normalize the expectations for greater duty of care from both domestic and international information-communication corporations. The discourse for legislating the internet is complex, but this complexity cannot justify the current laissez-faire approach to governance, and with it, the digitized allowances for misogyny.

V

Racialized Techno-political Organization of the Internet



A Digital Patriarchal Bargain: Violence against Women in Kenya

ANNE NJATHI and REBECCAH WAMBUI[1]

Abstract

A deep and nuanced account of the amplification of gender-based violence (GBV) on online platforms in Kenya has caught the attention of individuals, platform entities, and scholars. The aggressive scaling-up of online GBV (OGBV) has been reported as mainly coming from male bloggers and self-proclaimed content creators egregiously monopolizing online spaces on digital media platforms to silence, shame, embarrass, and belittle women. Our analysis of OGBV in Kenya reveals that women content creators are the victims of shadow banning, trolling,

[1] Anne W. Njathi is a PhD candidate and instructor at North Carolina State University's Interdisciplinary Communication, Rhetoric, and Digital Media program, where she conducts research on platform studies from an emerging markets perspective.

Rebeccah Wambui is an educator and public interest tech practitioner. She is an African School of Internet Governance Fellow and a Multistakeholder Advisory Group member for the Kenya Internet Governance Forum.

and cyberbullying. The three case studies we present demonstrate that while violence against women and queer folks continues unabated on social media platforms, the regulatory policies put in place to purportedly correct these harms reinforce gender and sexual hierarchies as well as class- and race-based inequalities. This, in turn, reifies the differential and unjust treatment of people belonging to marginalized sections. Our essay offers constructive criticism of existing policies such as the Kenyan Computer Misuse and Cybercrimes Act. It also envisions governance policies that dismantle systemic gender inequalities unfolding within digital platforms in the context of developing African countries.

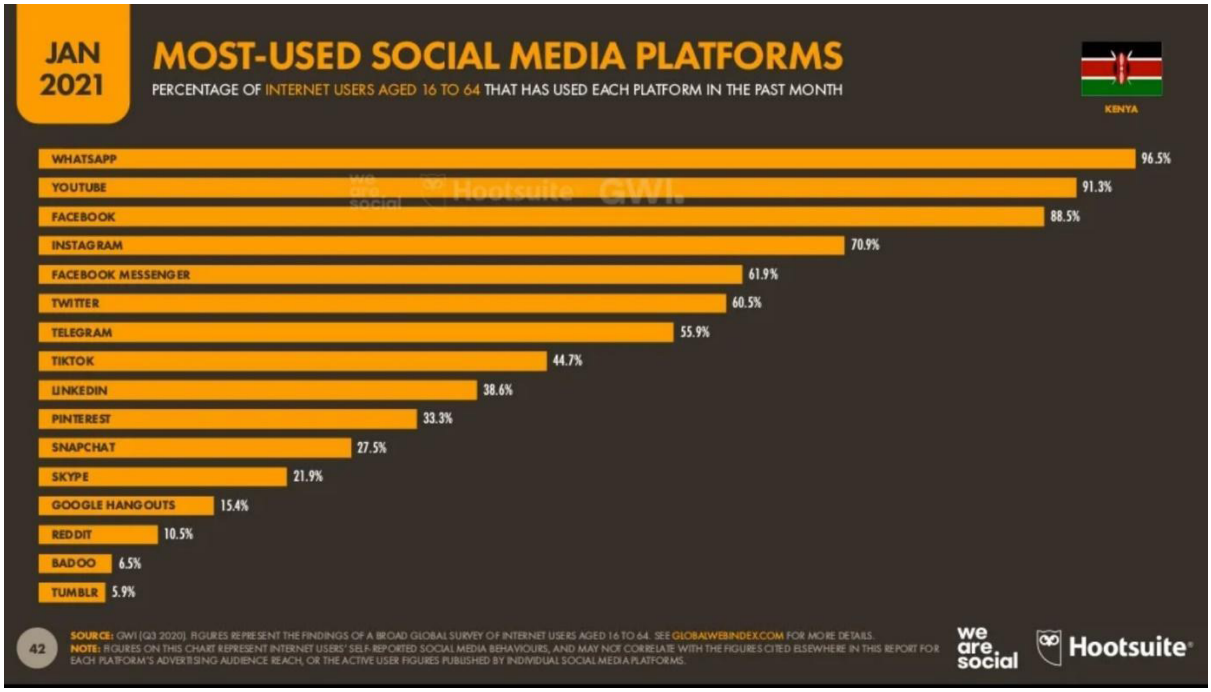
1. Introduction

A deep and nuanced account of the amplification of gender-based violence (GBV) on online platforms in Kenya has caught the attention of individuals, platform entities, and scholars. The aggressive scaling-up of online GBV (OGBV) has been reported as mainly coming from male bloggers and self-proclaimed content creators who have egregiously created monopolizing online spaces on platforms like Facebook, Instagram, YouTube, Twitter, TikTok, and Telegram to silence, shame, embarrass, and belittle women. These bloggers define and dictate womanhood according to what they define as African cultural expectations and beliefs that lay down that women are to be seen and not heard. The male bloggers have amassed cult followings on social media platforms where they personify the offline-based gender violence against women. They vehemently dissect and comment on women's online activities, as well as on topics that lie at the intersections of gender, beauty, lifestyle, relationships, sexuality, finances, and politics. These bloggers and their followers exhibit a propensity to attack female public figures, influencers, and digital creatives, but occasionally also direct their scrutiny at 'ordinary' women whose personal stories of misfortune, accomplishment, or arbitrary fame have thrust them into the public light. To advance their agenda, the male bloggers create, distribute, and monetize controversial content. Such endeavors mirror what scholars regard as emerging trends in the social media entertainment (SME) industry, where entrepreneurial content creators are harnessing digital platforms to build their brands in pursuit of gaining visibility whilst earning a living. At the same time, they attempt to balance their content to maintain authenticity and relatability with their followers.

As in other parts of the world, Kenya has seen a substantial increase in the uptake of social media platforms over the past few years. Of the 21.75 million internet users in the country, about 11 million people spend at least three hours on

average on social media platforms. WhatsApp (96.5%) is the most used platform, followed by YouTube (91.3%), Facebook (88.5%), Instagram (70.9%), and Facebook Messenger (61.9%) (see Figure 1).

Figure 1: Top five social media platforms in Kenya



Source: Digital 2021 - Global Overview Report

Kenya's overall internet usage stands at 42%, with more men than women using the internet on their mobile phones. However, reports have shown that women take the lead in the budding digital creator economy. Ironically, women also receive inordinate, subtle-to-extreme online harassment, thus eliciting interest in this juxtaposition. To provide a more nuanced analysis of OGBV in Kenya, we scrutinize three popular male digital content creators, Amerix, Andrew Kibe, and Edgar Obare, examine their digital positioning, and interrogate the content they generated between 2020 and 2022. This period is significant to this study because of the unprecedented rise in the use of digital platforms during this time due to the social distancing norms and lockdowns prompted by the Covid-19 pandemic, and the subsequent massive production of digital content. To guide our analysis, we pivot on randomly sampled content based on misogynistic focal points such as online hate, body shaming, trolling, cyberbullying, stereotypes, discrimination, and gossip. We begin by describing each of the three bloggers and providing an account of their heightened visibility strategies.

2. Blogger Case Studies

2.1 Amerix

Amerix is a verified Twitter account with over 500,000 followers. Eric, the blogger's actual name, ostensibly offers advice on men's health and wellness under the '#MasculinitySaturday' hashtag on Twitter (see Figure 2). He also posts content from the 'Masculinity Sato' account. A glance through the #MasculinitySaturday hashtag reveals a barrage of misogynistic propaganda interspersed with a minuscule dose of reasonable financial and health advice. An analysis of Amerix's posts and interactions with followers indicates that his end goal is to eradicate "weak masculinity" by offering controversial advice on "proper masculinity", and replacing "simp behavior". Simp behavior is seen as sympathy, support, or attention to women and women-related issues. Amerix's followers dedicate themselves to gathering evidence of other men simping both online and offline. They announce their presence in the online space by posting the hashtag "#simp" in the comments section of the posts they disapprove of. In their view, the range of simp behavior is vast, with the gravest offenses being posting on women's issues publicly and posing as an ally. Amerix, who uses the alias 'The President', once lost his highly coveted Twitter verification badge due to allegations of contravening the platform's rules regarding inciting hate against a particular group. He has, however, denied those allegations and attributed the temporary loss to Twitter's technical glitches.

Figure 2: Amerix Twitter profile

Source: Twitter

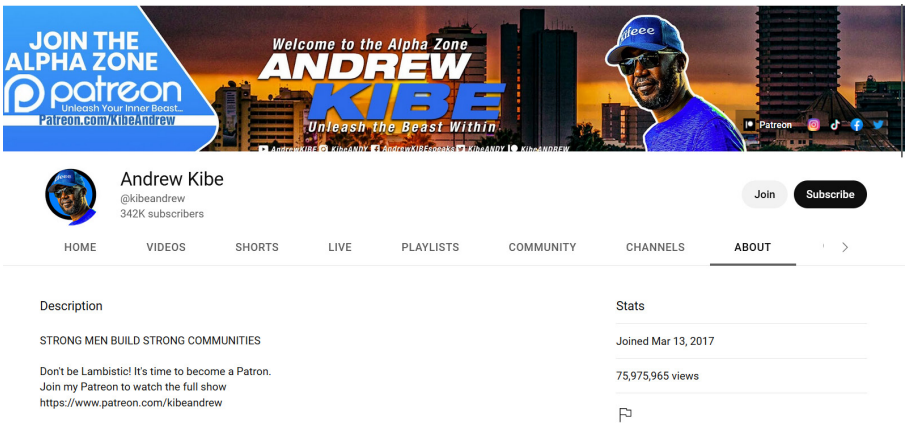
Figure 3: Amerix's tweet on #MasculinitySaturday

Source: Twitter

2.2 Andrew Kibe

Andrew Kibe's verified Instagram profile identifies him as "content creator extraordinaire", and his Twitter profile describes him as a "truth seeker". His YouTube channel advocates for male empowerment, arguing that "strong men build strong communities" (Figure 4). Kibe's Patreon account states that "all creation is groaning for the men to wake up".

Figure 4: Andrew Kibe's YouTube channel



In early 2022, Kibe announced his YouTube content would be moving to the subscription service Patreon, where users can continue to access his content for as little as USD 5.00/month.

Kibe riveted his popularity to social media from mainstream media. He has since created a career in profiling, commentating, and humiliating women on his platforms. With a following of over 1 million across YouTube (260,000), Instagram (230,000), Twitter (200,000), Facebook (432,000), Telegram, Patreon (148), and his podcast, his strategy is to hold daily live sessions on controversial but compelling content. Subsequently, Kibe's platforms have experienced exponential growth. At the live sessions, Kibe's fans and followers seek life and relationship advice. He has distinguished himself by peppering his colloquial lingo with words like "Lambistic" and "Kinuthia" that many of his followers regularly integrate into their online conversations. Across all his platforms, he riles his male followers up against women using blatantly misogynistic terminology such as "bitter feminists" and "feminazis". An article titled 'Andrew Kibe: A Troll or Kenya's Top Social Commentator?' published on a popular blog questioned his content creation tact even as it ironically endorsed him as an authority that most Kenyans, including women, are looking up to. This, despite the fact that his content is mostly targeted at empowering men. Although Kibe has received much public and local media attention, there are very few instances of people calling him out, thus normalizing online hate toward women. For example, in May 2022, the Kenyan Somali community published a YouTube video titled 'Andrew Kibe Gets a Warning for Insulting Somali Ladies'. The warning came right after Kibe's YouTube channel had been suspended in April 2022 due to offensive comments about the LGBTQ community in a previous live video.

2.3 Edgar Obare

Edgar Obare's content creation and monetization tactics have been hailed as game-changing by the Africa Digital Leadership Webinar. With 371,000 followers on Instagram as of June 2020, and 134,000 on YouTube, his popularity hinges on what is colloquially known in Kenya as "serving tea", a stand-in term for an act that involves exposing, critiquing, and initiating gossip about famous individuals. This is made possible by Obare's loyal, dedicated fans who self-identify as "Edgar's students". On occasion, they are prompted by Obare to report on high-profile public figures, but they mostly act on their own accord, surveilling online and offline spaces for evidence, also known as "receipts", of any perceived misdemeanors such as locations they visit, company kept by these individuals, their spendings, and romantic interests. Receipts come in different forms but are mostly screenshots of private text messages, recorded phone calls, photos, or

secretly recorded videos. Sometimes, these are actual financial spending receipts. These "receipts" are then delivered to Obare through direct messages (DMs) to be considered for broadcast and/or to be "served as tea", for the rest of his followers to deliberate on, dissect, and derive entertainment from. On several occasions, such sensational blogging techniques earned Obare threats and arrests on charges of violating Section 72 of the Data Protection Act.

In addition, Obare occasionally permits his followers to table their personal agendas and woes in the form of appeals for blood donation or calls for help in the search of lost kin (see Figure 5), which are relatively smaller issues and are unlikely to get much mainstream media coverage. It is for this reason that his followers consider him trustworthy and exceptional. Some of his popular exposés in the past leaned toward advocacy for social and political change. Simultaneously, Obare positions himself as an agent of change by offering positive entertainment needed as a release or escape for an already stressed Kenyan population.

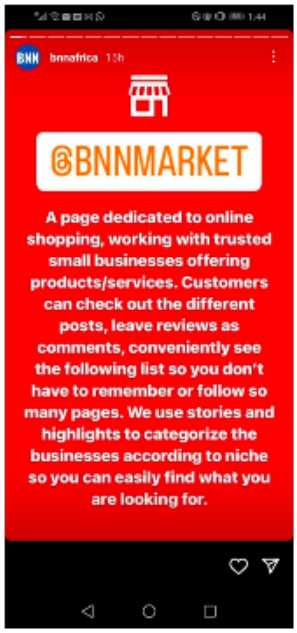
Figure 5: A follower's post on Edgar Obare's Instagram account

Source: Instagram

Obare has a sizeable and diverse following across the East African region on Instagram, Twitter, and Telegram. Among his followers are business owners and potential clients seeking the Market Day service on his BNN Market page (see Figure 6). Market Day is a paid-for service that runs twice a week, and advertises and markets local and small businesses on BNN's Instagram and Telegram pages. The USD 50 per ad maxed out the allocated 100 Instagram stories per day and

garnered over 50,000 views during prime days. This "easy and fast cash" service set a precedent for other influential figures. Since his reappearance after the Instagram ban, Obare has modified this model to include subscription models for his Telegram channel, establishing himself as a one-man tabloid.

Figure 6: Edgar Obare's BNN Market page



Source: Instagram

3. Algorithmic Mastery

Although these three male bloggers have established themselves on different social media platforms, they portray a mastery of the algorithmic design that promotes engagement with divisive content. An algorithm is defined as a computer-based program that filters what users can read. We observed that each blogger receives a fair share of internet fame made evident through platform metrics such as likes, comments, shares, saves, and reposts, making them highly visible on social media and earning them top trends in the country as well as on media coverage. The bloggers' content elicits an equal measure of admiration and disapproval from all genders, further illuminating the ingrained misogyny that consciously and unconsciously promotes and normalizes harmful gender biases, attitudes, and behaviors. The bloggers perceive their work to be informative, empowering, and borderline messianic, restoring societal gender roles and gender order. Their cultic followers, high on their leaders' missions, leverage

the anonymity, virality, and scale afforded by social media to amplify these online activities, misconstruing this as making real impact. To be sure, there are occasions when social media outcry results in swift changes in the formulation and implementation of laws. However, we submit that the arguable success of these bloggers is not of inherent ingenuity but a combination of skills, access, and the need to build an income by monetizing content on digital platforms.

4. Online Gender-based Violence on Social Media

Having profiled the three male bloggers, we highlight three case studies of victims to showcase how instances of OGBV are manifesting in Kenya, amplified by content creators at the forefront of the oppression of women. The first example is of Elsa Majimbo, a young Kenyan comedian who gained international fame at the height of the Covid-19 crisis after sharing short monologues on life during the pandemic on TikTok. Soon after, a Twitter account that called itself 'Kenyan On Twitter' began to troll her for her physicality disguised as feedback on her comedy style. Few people came out in support of Majimbo at the time and she held colorism among Kenyans responsible for the one-sided attacks against her. This further catalyzed an unending barrage of online hatred against the comedian, and eventually led Majimbo to virtually denounce her patriotism and leave the country.

Similarly, Natali Tewa, a popular Nairobi influencer, was subjected to online public scrutiny over an alleged relationship scandal with a famous politician. Obare leaked Tewa's travel passport data on Instagram as "receipts" that would prove her alleged romantic relations with the said politician. This was followed by aggravated emotional abuse directed at her by Obare's Instagram followers, which later spilled over to other social media platforms, attracting the attention of mainstream media where it eventually became a topic of national interest.

Obare was charged with violating Section 72 of the Data Protection Act for leaking Tewa's personal information and faced threats and arrest several times. Meanwhile, Tewa decided to quit her content creation work as well as her social media engagements. Her tweet (see Figure 7) captures her frustrations with gossip blogs.

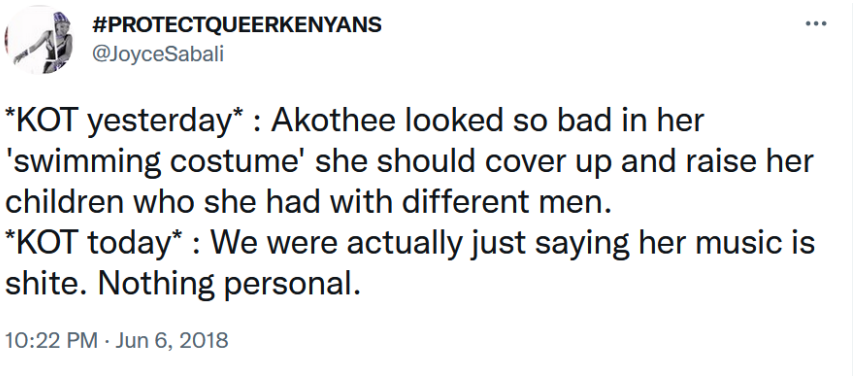
Akothee, musician and entrepreneur, also faced unending scrutiny and attacks for her social media activities. Of particular interest is the online attack against her for a costume she had donned while performing at an international event in London, which was perceived as undesirable. She was not only body shamed but the dress also became an excuse to criticize her musical abilities. Some commentators linked

Figure 7: Tewa's tweet

Source: Twitter

the magnitude of the trolling she experienced to the double-standards faced by darker-skinned women in the global entertainment industry, which are rooted in colorism. Other criticism focused on Akothee's morality and single motherhood status, with some social media users urging her to focus on motherhood (see Figure 8).

Figure 8: Criticism of Akothee by Twitter users



Source: Twitter

Akothee stated in interviews that no amount of trolling will curtail her freedom of expression on social media. Her tactic of retaliation, which is similar in nature to her aggressors, is in stark contrast to other women who were attacked, including in the

cases we studied. Interestingly, her retaliatory responses have lessened the online harassment against her. This scenario foregrounds the 'strong woman' narrative that demands that women overlook or downplay harm in order to survive. In the social media landscape, the expectation is that women should opt for silence or withdrawal from engagement, as seen in the scenarios of Tewa and Majimbo.

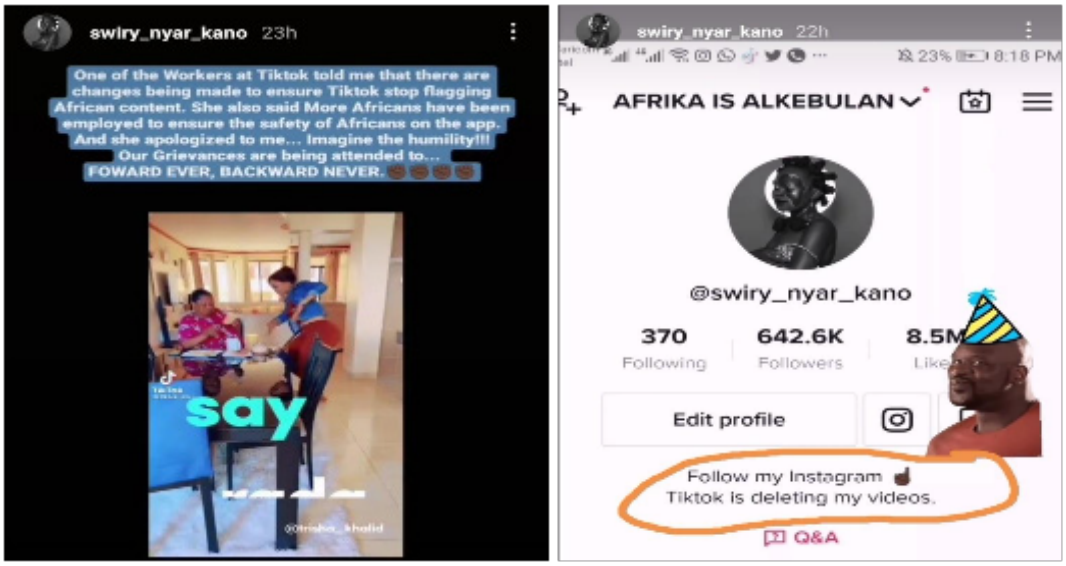
Together, these incidents highlight the grave consequences of online abuse, such as body shaming and trolling, on Kenyan women. When OGBV negatively affects women's mental health, the power of perpetrators becomes further entrenched. The attention garnered by the bloggers discussed in this essay enables them to build a network effect which they then exploit as currency to advance their financial and social capital. Overall, misogyny is reinforced, and social justice and fairness are obscured through digital harms, power imbalances, and misconceptions, complicating the safe use of the internet. Inevitably, OGBV risks amplifying existing inequalities, potentially leading to new forms of discrimination in which women are more likely to suffer disproportionally.

5. Critical Areas of Concern

5.1 Regulating content to safeguard users' interests

Our analysis of OGBV in Kenya points to two critical areas of concern. The first is on the role of online platforms in regulating content to safeguard users' interests. In the context of patriarchal societies, online self-expression is emerging as a preserve for men to enjoy and capitalize on, either as harmless online banter and/or curated commentary. Indeed, in our study, we did not come across women bloggers with content equivalent to the three male bloggers we examined. If anything, the experiences of Majimbo, Tewa, and Akothee demonstrate that women's freedom of expression is constantly policed and curtailed. Interestingly, women's self-expression on these platforms also faces algorithmic discrimination in the form of shadow banning, among other things. Shadow banning is a deliberate act to hide, restrict, or delete users' content without their knowledge or consent. For instance, Swiry Nyar Kano, a popular TikToker with a following of over 900,000, was shadow banned for her precolonial African history content, which then disappeared in perpetuity. With her efforts to reach the TikTok management proving futile, Swiry Nyar Kano resorted to reposting the content on her Instagram account (see Figure 9).

Figure 9: Shadow banning of Swiry Nyar Kano on TikTok



Source: TikTok

Other accounts such as @thespreadpod, a queer-led sex-ed platform, and @nowhitesaviors, a black racial justice advocacy and education platform on Instagram, were also shadow banned or had their content deleted. The temporal and sometimes permanent content removal at the discretion of mostly automated moderation, which lack human compassion, raises concerns about insensitivity to the Global South's cultural contexts while inadvertently upholding and promoting harmful sexist and racist content. The double standards and inconsistencies in the enforcement of Instagram's guidelines further harm and marginalize those who find themselves disadvantaged by the intersections of race, gender, economic status, and geographic location. In addition, scholars have acknowledged that platforms are convoluted global spaces that produce a wealth of problems regarding privacy, data protection, jurisdiction, labor, crime, and policing, among other ethical quandaries. Our case studies and examples reflect how platforms are "hotbeds of misogynistic activism" and demonstrate how platform design, algorithm, and politics implicitly coax the problems alluded to earlier.

The unfair treatment meted out to Instagram accounts such as Swiry Nyar Kano, @thespreadpod, and @nowhitesaviors showcase platform-based algorithmic politics. Platforms like YouTube, Instagram, Twitter, Facebook, and TikTok that filter and moderate content can cause harm to marginalized groups. Once content has been taken down, it is difficult to contest a platform's decision as shown in Figure 10 which is a screenshot from @nowhitesaviors' account.

Figure 10: Navigating the algorithm

Source: Instagram

A study conducted by Columbia University researchers found that women are more likely to be obscured on Instagram than men, a phenomenon they described as "algorithmic glass ceiling". Figure 11 provides an example of such a glass ceiling. While Meta's equity and inclusion investigation teams have attempted to respond to algorithmic discrimination claims, our findings continue to show that not much

Figure 11: Algorithmic glass ceiling

Source: Instagram

progress has been made on regulating and filtering inappropriate content, such as hate speech, to help prevent harmful interactions and foster a healthy space for women on these platforms.

Platform algorithmic design is grounded upon what Nick Srnicek calls "platform capitalism", whereby the activities of users become the natural source of raw materials for platforms to grow their user base and expand into new markets. Platform metrics imply value is being generated. The online activities and practices of the three bloggers in this study become "an economic actor within a capitalist mode of production". Data is the new form of capitalist accumulation used by platforms as a business model to maximize profits. Hence, platform regulation of content conflicts with platforms' commercial interests.

5.2 Regulating content to safeguard women's interests

The second concern is on how existing laws and regulations safeguard women's interests on digital platforms. The Global South, in particular, suffers from nonexistent or incomplete policies. We argue that there is an urgent need for social media governance to address these emerging online gender discrepancies. In May 2018, Kenya introduced the Computer Misuse and Cybercrimes Act to provide for the detection and prosecution of offenses related to computer systems, as well as to protect the confidentiality, integrity, and availability of computer systems, programs, and data, among other things. The Act was contested by an association of Kenyan bloggers and content creators at the High Court for seemingly violating the constitutionally guaranteed freedom of expression. What remains unclear is why GBV cases are still on the rise and remain unchecked despite the provisions of the Computer Misuse and Cybercrimes Act, 2018. We contend that the lack of civic knowledge, the lack of trust in the judicial institutions, and poor implementation inhibit adaptation and goodwill by the general public. Moreover, as we have highlighted in this study, the content posted by bloggers are seemingly targeted at individuals appealing to their moral, emotional, and personal sensibilities, posing difficulties in developing unanimous societal guidelines. It is on these grounds that bloggers take advantage of the loopholes to propagate their agendas. At the same time, platforms are constantly changing at an unprecedented speed, posing a challenge for policy formulation and implementation. As scholars Evelyn Douek and Genevieve Lakier state, regulating in the dark is always a bad idea and is bound to have unintended consequences. Lawmakers cannot solve problems

they do not understand. As shown in this study, OGBV is entering a new paradigm that challenges existing digital policies in unprecedented ways, and formulating effective legislation would require staying abreast of these changes.

6. Conclusion

It appears the rise in OGBV cases in Kenya is by design. It is in platforms' interest to let content remain unregulated and stir virality, even if it comes at the cost of ignoring and enabling digital harms, because private social media companies that own digital public spaces are rooted in a capitalist system. Unregulated content by bloggers stirs virality. The case studies reviewed in this essay indicate that women are caught in a catch-22 capitalist patriarchal bargain. The regulators, in this case lawmakers and implementers, are outpaced by the swiftly moving technological advancements that propel digital platforms. The overarching effect is a watering down of the milestones in digital gender equality achieved so far, leaving Kenyan women vulnerable, suppressed, and without a voice in both offline and online venues. Our analysis raises the question, "From where do platforms get their governing values?" Content regulation in Kenya appears to be prioritized during certain events, such as the general elections, but lacks the same scrutiny in policies that safeguard individuals' freedom of expression. We envision governance policies that dismantle systemic gender inequalities unfolding within digital platforms in the context of developing African countries.

Against Internet Coloniality: Notes on Misogyno-Racist Violence on the Internet

GRACIELA NATANSOHN[1]

Abstract

This essay proposes that in order to address gender-based violence in digital environments it is necessary to understand its linkages with racism, neoliberal capitalism, and commodified forms of the internet. Situated in the context of Brazil, it contends that the current techno-political configuration of the internet, characterized by datafication, platformization, surveillance, and data extractivism generates a political economy that sustains patriarchal violence. Therefore, violence

[1] Graciela Natansohn is a professor and researcher at PósCom, Universidade Federal da Bahia, Brazil. She researches digital communication with a feminist perspective, cyber-hack-transfeminisms and technofeminisms. She coordinates the research group 'Gender, Digital Technologies and Culture'.

cannot be interpreted as deviations, misuses, or defects of digital communication processes; rather, it is the outcome of a racist, neoliberal, and datafied patriarchy. This essay uses the concept of 'pedagogies of cruelty' to explain the practices that normalize violence, that teach, habituate, and program subjects into a datafied-commodified life such that they are desensitized to regimes of cruelty. It deploys the lens of pedagogies of cruelty to make sense of the everyday violence that is systematically reproduced in the current phase of capitalism — neoliberalism — and helps to understand the etiology of digital violence.

1. Introduction

Situating the essay in the context of Brazil, this essay argues that it is necessary to highlight and insist that gendered relations are always racialized, and that gender, in and by itself, is not a determinate category defining the scope, nature, and forms of violence. Individuals who identify as indigenous, Black, or Latin, and are located far from urban centers are gendered and racialized, besides facing political and economic disadvantages. The inequalities, injustices, and harms that such individuals, their communities and organizations face are exaggerated in the digital environment. It is not a coincidence that the most frequent victims of gender-based violence in Brazil are Black women.

To understand misogyno-racist violence in digital environments, we follow two stages of exploration. First, We trace the long history of patriarchy through the brief history of capitalism in its contemporary neoliberal avatar using the writings of Argentine-Brazilian feminist anthropologist, Rita Segato. The second stage relates the ideas of Segato to the media practices that guide platforms and the algorithms of public relevance. This exploration allows us to develop a framework to study the structure of the internet and how it relates to the algorithms that orient the political economy of platforms and shape their users' behaviors.

The study demonstrates how racial patriarchy, capitalism, and techno-political reorganization of social relations are not mutually exclusive but semi-autonomous spheres that co-constitute the social. More specifically, the essay intends to find and establish the interconnections between the following three spheres:

1. Patriarchy and racism which have a long history that precedes capitalism but which have been reinforced, according to authors like Segato, through colonization and the conquest of América/Abya Yala.

2. Capitalism in its most recent version – neoliberalism.
3. The current techno-political organization of the internet characterized by dataism, platformization, vigilance, and data extractivism that develops into forms known as data colonialism data capitalism, and techno-feudalism.

This autonomy may be recognized, for example, on the temporal dimension of the transformations of these spheres. While patriarchy and racialized power relations have existed throughout the centuries, its implication with digital technologies has caused certain shifts in its orientation. The classic neoliberalism of the beginning of the twentieth century has mutated into neoconservative forms from the 1980s onwards, data capitalism has introduced radical changes in economic relations, putting at stake even the role of states in the last 20 years.

The three spheres outlined above don't operate independently, meaning racism patriarchy, neoliberalism, and digital technologies are mutually influenced, transformed, and co-produced even as they produce the social together. This means that the technological does not operate only *over* the social or only *over* gender and class; they are mutually determined and co-produced, generating different effects. This essay examines gender-based violence in online environments not merely as an extension of the long-term patriarchal violence that marks the history of gender relations. Rather, it focuses on the specific digital means of perpetuating such violence while remaining cognizant of the fact that gender and race never operate separately. As a recent report from the United Nations on cyberviolence and cyber harassment against women and girls observes, the "fast development of technology and of digital spaces [...] inevitably give way to new and different manifestations of violence against women". Therefore, we can state that misogynistic-racist violence in digital environments enables the political economy of the internet in the framework of neoliberal capitalism, and is an effect of it.

White, urban, middle-class, male or masculinized persons must also be analyzed as gendered and racialized individuals. Given that racism and whiteness act in concert, white, urban, middle class, male or masculinized people hold wide subjective, symbolic, and material advantages over Black and Indigenous people. From the dynamics between gender, race, and class (this essay does not engage with other related hierarchies such as territory, religion, ethnicity, etc.) arise tensions that manifest in aggression and violence against female or feminized bodies in digital environments. This essay proposes naming these kinds of violence

in the Brazilian context 'misogyno-racism'. It is crucial to remember that misogyny is directed not only at women but also at individuals who display characteristics that could be associated with femininity; whether physical, cultural, related to their subjectivity, sexual orientation, or sensibility.

Historically, feminist activism against gender-based violence has engaged with the categories of class and gender. Race, on the other hand, has been neglected in such analyses despite the battle waged by Black and Latin feminists for its recognition. The exclusion of race from analysis of gender dynamics does only not render such an inquiry incomplete, but also precludes a full understanding of the subject, leading to what some scholars have termed epistemicide,[2] producing theories and knowledge that are blind to the complexity and richness of the analyzed experiences.

In Brazil, Black, Indigenous, young, activists, disabled, and LGBTIQIA+ women are exposed to these kinds of online gender-based violence most frequently, thus underscoring the importance of incorporating an intersectional perspective into public policies and analytical documents. As referred by CEVI,

> The convergence of multiple forms of discrimination increases the risk for some women to be victimized by a specific, composite or structural discrimination", which is evident in the cyberspace in which gender violence is perpetrated with the intention of disciplining, controlling, and/or silencing women who define themselves in multiple forms.

We do not deny that the discussion about race, racism, and violence has historical and circumstantial particularity and specificity, just as gender and sexuality also have their own history and characteristics specific to each circumstance. We advocate for analyses that do not lose sight of intersectionality or ignore the fact that the various social markers of difference, inequality, and hierarchy articulate and interconnect within these tense sociocultural dynamics and produce different effects. For example, any analyses of domestic violence on indigenous women must understand how the category of woman operates in one's local ethnic culture, and what strategies native women's collectives have put in place to prevent white supremacy from revictimizing their racialized male partners in the name of feminism.

[2] Epistemicide is a term used by Santos and Menezes in their 2009 book 'Southern Epistemologies' to describe the destruction, invisibilization, and/or belittling (or underestimation) of knowledge and cultures dominated by white European colonization. It is the result of white supremacy's influence over the construction of scientific and popular knowledge.

It is also important to highlight, as many reports and documents produced by international agencies have been doing, that the gender violence manifestations that are part of this continuum (online-offline) have a common element:

> All of them are forms of coercion, abuse and/or aggression performed with the intent of controlling, limiting or constraining the lives, bodies, movements, conditions, and opportunities of women and girls, and of maintaining, reproducing and perpetuating – online and offline – an inequality system and patriarchal structures of coercion in which women and girls are placed on a subordinate position in relation to men, based on detrimental gender stereotypes. […] Therefore, we emphasize that cyberviolence should be understood as an expression of a *continuum* of multiple, interrelated and recurrent forms of gender violence that affect women and girls since birth, and that are present in every sphere of their lives (public, private, physical and, now, digital).

This means that these forms of violence must not be interpreted as deviations, misuses or flaws in the process of digital communication. On the contrary, they are outcomes or effects of the intersection of the above-mentioned three spheres. They are a structural outcome of the patriarchal, capitalist, colonial, racist, and datafied order. It is, also, an effect of the behavior modulation induced by the use of technological artifacts. Some authors have named this the "psychic economy of algorithms", which is a mechanism of data capitalism that favors massive data collection about emotional states of platform users. Such information is highly valuable for the composition of user profiles for marketing purposes. By psychic economy of algorithms, we refer to:

> The contemporary investment – techno-scientific, economical and social – in algorithmic processes to capture, analyze, and use psychic and emotional information extracted from our data and actions on digital platforms (social media, applications, streaming services, sharing platforms and/or consumption of audiovisual content, etc.). The pieces of information that matter to the fast data capitalism are no longer mere traces of our actions and interactions (clicks, likes, shares, posts, visualizations), but also their psychic and emotional 'tonality'. This is the psychic and emotional economy that feeds the current strategies for behavioral prediction and induction on digital platforms (and, at times, outside them).

If it is true that "a complex and increasing psychic and emotional economy feeds algorithms that intend to know us better than ourselves, besides making predictions and interventions over our emotions and actions", we must not fall prey to false technological determinisms that obscure the roots of the problem. A contemporary and postmodern Luddism may be a resistance strategy, but given the omnipresence and ubiquity of technologies, its success is unlikely.

2. A Few Explanatory Hypotheses for Violence

Far from the technological solutionism nurtured in Silicon Valley – or its opposite, technophobia – we must acknowledge that the general reification and commodification of contemporary life (at least, in the western world) seem to be the most radical forms of violence; The exploitation of women and children for sexual work, labor precariousness, semi-slave labor, predatory resource exploitation, femicides, the systematic murder of Black people by the state, the expulsion of native peoples from their territories, all require 'pedagogies of cruelty' that teach, accustom, and program individuals to transform the flow of life into things, and train them for insensibility.

Data mining companies that produce commodities in the global market also convert life itself into something measurable, sellable, purchasable, usable, and disposable. This model of accumulation demands political justification and requires censorship of and control over symbolic exchanges on their platforms. The networks are the territory from which information supplies are obtained, and with which marketing and cognitive manipulation in the interest of multilateral corporations occurs.

> From the parameters of American corporations, everything that entails a trait of sovereignty – or that promotes alternative versions to reality – is liable to be cataloged as a violating infraction of the schemes imposed by those intended to hegemonize them. Based on this reasoning, the rules of interactions cannot be debated by nation states. They can only be taken as univocal and indisputable. That is the core of unilateralism: the rules are demanded as an acceptance of the global domination outside the International Law.

As Segato affirms, the paradigm of neoliberal exploitation requires a great variety of forms of abandonment and precarization of life which needs a principle of cruelty that is able to reduce empathy between individuals, and make injustice viable at both subjective and symbolic levels. It is in this scenario that technologies

are built, developed, and transformed. When seen from this perspective, the cruelty of violence; normalized and poorly understood, would gain intelligibility. Violence would be a logical correlate of the reification of neoliberal life in which technology is fused and developed.

The intense processes of abstraction, translation, and fragmentation of the individual into digital data that we have seen in the last decades have led individuals themselves to become the commodities of a new economy. The Internet and digital technologies have opened the door to a mercantilization of our idle time, our social relations, our emotional dimension and our desire. Therefore, new forms of exploitation of our bodies of flesh and blood emerge, that are not directly, but surreptitiously exerted through our data.

Life as a whole; emotions, feelings, and bonds have been commodified in a new and radical way: in the form of data. The bodies, that are now also bodies of data, are mere things, purchasable, sellable, interchangeable, disposable, 'instagrammable'. The moment we submit information about our sick bodies to health apps, for example, we generate data that can be used for medical advertising, health insurance, and perhaps determine our chances of getting a job. Hate speech is also bought and sold for political propaganda, or for 'click farms' that collect followers, or 'likes' on social media platforms, that are then converted into money. From this perspective, the 'violent' (whether bot or human) would not be an anomaly. Online violence is not perpetuated only by a deviant individual. There is an affinity between the neoliberal rationale and authoritarian subjectivities, and there is functionality between these and hegemonic ways of inhabiting the internet, promoted by the predominating actors: the platforms.[3] [4]

To make historical gender-based violence intelligible, Segato develops a complex theoretical framework from which this essay takes some ideas. She argues that

[3] While platforms are the predominant actors, they are not the only ones. In another article, we discuss how the internet can be the site for two kinds of misogyno-racist violence: the one manifested in the sphere of online interactions (the most visible one), and the one hidden in the networks' architecture. The latter allows systematic tracking and analyses of personal data to feed the internet's business model through applications and digital products targeted specifically at women; applications to track menstruation, eating habits, and vital cycles, into which women record many aspects of their daily lives.

[4] In this sense, UN Women recognizes that, as a consequence of the gender-related prejudices (this essay suggests adding race to this) incorporated in the design of artificial intelligence and machine learning algorithms, those gendered (and raced) prejudices and biases are multiplied by the new technologies, generating algorithmic racism and misogyny.

there is a tension in modern colonial societies between the status system and the contract system: that is, between an ancient stratified rationale common to patriarchal and stratified societies, and a contractual rationale common to modern societies consolidated by the modern State organization and its laws. Both rationales coexist – one sustained by traditional customs and moral values, and the other by the rationality of laws – and are not completely apart. Carol Pateman points out the patriarchal elements of the social contract in Rousseau and Locke in her book 'The Sexual Contract', in which she analyzes how women's subordination is organized within the liberal political order which presents itself as neutral in relation to gender. As a result, laws are not able to avoid gender crimes that, from the point of view of status, would be moral offenses. Segato presents, as an example, a survey with sex offenders carried out in Brasília. In one of the interviews conducted during the investigation, the offender sees himself as a moralizer, discipliner, and avenger against a generically addressed woman who affronts the place tradition has assigned her (by wearing "provocative clothes" for example). "Someone whose moral judgment falls over the woman with total severity is the same who commits what, according to Brazilian legislation, is a heinous crime," Segato affirms. This apparent paradox is a product of the tension between the status rationale and the social contract rationale that shelters the modern state and its legal and institutional framework.

> The affront from this generic woman, a modern individual, an autonomous citizen, castrates the sex offender, who restores male power and his virile moral to the system by putting her in her relative place through the criminal act he commits. This is the symbolic economy of the violation as a moralizing crime, though illegal.

One of the elementary structures of gender-based violence analyzed by Segato resides in the constitutive tension between status and contract. The status system that persists to this day is one that equates women with domesticity,[5]("beautiful modest housewife") morality, submission, and honor, as a tribute to man who exerts his dominance and exhibits it to his peers. What we call dominant masculinity is an intersubjective formation with the ability to dominate (vertically) and display prestige to peers (horizontally). It is what gives men a sense of identity and belonging as a brotherhood or a mafia group. Challenging women's aspirations

[5] As an example, a newspaper complimented the wife of Brazilian president Michel Temer by referring to her as a "beautiful modest housewife".

for autonomy, "this violent effect results from the moral and moralizing mandate to reduce and imprison women in their subordinate position, by all possible means, resorting to sexual, psychological and physical violence". That is the ethical core of the "moral and common decency" of the "goodwill citizens",[6] which is reproduced and embedded in the modern contractual state, challenging its laws and generating cycles of repeated violence which is aggravated in conservative political scenarios.

Segato, thus, states that these violent acts have an expressive, communicative dimension which manifests in two interlocutory axes: a vertical axis, when a man exerts cruelty and dominance over women in violent acts, and a horizontal axis, because the violent man also performs to his peers, other men who take part in what Segato calls a "masculinist brotherhood".

According to Segato, the racist colonial capitalist masculinity operates as a brotherhood, a mafia, a confraternity. Belonging to this brotherhood demands allegiance rituals as well as fidelity and proof of masculinity (which is why men are its greatest victims – data about violence towards Black men in Brazil is overwhelming). She interprets violence as a message exchange among fellows, that is, violence is not related to pleasure or male sexual desire. Female or feminized bodies are merely a passing point for a masculinist conversation. If there is any enjoyment in this violence, it is related to the male co-operative pact, not with the victim's body; it is about power, not desire. Or, in any case, about desire for power.[7]

We are obviously speaking of men not as an essence or a biological body, but as a historical subject that projects and introjects a hegemonic form of masculinity

---

[6] "Goodwill citizen" is a figure of speech used by Jair Bolsonaro's supporters to refer to people who advocate for the extensive use of firearms, who condemn what they call 'gender ideology' as well as broad sexual education in schools, and, in short, who reclaim a neoconservative agenda.

[7] The messages that usually circulate on WhatsApp groups in which men are present are an example of the normalization of female reification and of the operation of violent language on two axes. Whatever the aim of the group, and even if men are less numerous, we commonly see the circulation of pictures, memes, or images of naked women, obscene or derogatory comments, and even reports of supposed sexual deeds, at times in a humorous tone that generates complicity, broken only by the irruption, if any, of women that do not feel challenged by these images and that question them, causing embarrassment to the group. These images are never of companions, friends, or daughters, but of other women, those who take part in public life (models, actresses, journalists, politicians, or those whose sexual or gender identity is deemed dissident), women who stand apart from the moralizing arm, from the controlling cane, from the domesticating whiplash and, for that, are subjected to and considered deserving of symbolic punishment.

in which whiteness, dominant heterosexuality, cisgenderness, economic power, geographic location, social class, and territory intervene. It was this perspicuity that reportedly led the black North American leader Malcolm X to advocate:

> The black man is going around saying he wants respect: well, the black man never will get anybody's respect until he first learns to respect his own women! The black man needs today to stand up and throw off the weaknesses imposed upon him by the slave master white man! The black man needs to start today to shelter and protect and respect his black women!

Following this trail of thought, gender-based attacks in digital environments may be read as statements or communicative acts that act both on the victims and on other men, and have wide resonance on digital platforms. One may observe that many of the attacks on women on social media come from people who are unknown to them. Social media platforms are an agora, a public space in which men can perform their masculinity as well as show off, connect, and legitimize themselves before their peers.

Another example is the recording and sharing of group rapes on social media. The authors are aware of the risks of being tracked and legally charged, but the surplus gain of this action is on its publication and interlocution with their peers. This is part of the ritual of proving masculinity, amplified and spread on social media, and essential to fulfilling one of the main clauses of the masculinist pact.

3. Conclusion

Attacks on the human rights of women as well as LGBTQ+ and black people in Brazil are historical and endemic, result from colonization and slavery – the bases of the modern Brazilian state – and are perpetuated through a racist, masculinist, transphobic, and homophobic culture. Since 2018, a surge in institutional violence started just as Jair Messias Bolsonaro took on the presidency of the Republic, lending a conservative orientation to the government, and promoting hateful, racist, anti-rights, and anti-feminist discourses and practices, supported by evangelical pentecostal sectors, of which he is part. As such, a 'pedagogy of cruelty is underway in Brazil, characterized by a moralistic, conservative, exclusionary and authoritarian state policy which is presided over by its leader.

The ancient statist and patriarchal logics that live on in modern societies combine with contemporary neoliberal logics, and are complementary even if they seem

contradictory. Faced with millenarian customs such as female subordination, modern laws to promote equality prove ineffective because both logics coexist with tensions in the modern colonial state that has never de-patriarchalized itself. Added to this panorama is the corporations that dominate the internet and produce commodities for the global market in the form of data and advertising, and that have contributed to turning all aspects of life into commodities. The networks and platforms are the space from which marketing, cognitive manipulation, and functional accounts for racist and patriarchal corporations are generated.

To give intelligibility to the seemingly inexplicable misogyno-racist violence, it is necessary to understand how racist patriarchy, neoliberal capitalism, and contemporary technopolitical organization combine to create a hostile environment that furthers the accumulation model.

Segato incites us to create counter-pedagogies of cruelty, which encompass recovering sensibility, bonding, and feminist practices of care — of reciprocity, of sharing, and of communication. Such pedagogies also involve trying to establish dialogues with men willing to abandon the nefarious pact of masculinity which also victimizes them.

To be sure, there is a need to develop a local technological intelligence that is auditable and open, and that promotes algorithmic sovereignty and technological knowledge as a free common good. But that will never suffice because the solution is not technological. Without decolonization and de-patriarchalization, we cannot transform the economy of desire. Along the same lines, Segato states: "In [a] gender symbolic economy, a position is seen as feminine because from it circulates a tribute towards the masculine position that extorts it, and nourishes from it".

It is necessary to produce scenarios of other forms of life, to act micro-politically, to appropriate the strength of creation and cooperation, to produce other forms of subjectivization; as repeated throughout the Social World Forum in Salvador de Bahia, Brazil in 2018, to think that another internet and another world are possible.