T20 Policy Brief

G20 BRASIL 2024

T20 Brasil 2024 Let's rethink the world

Task Force 05
INCLUSIVE DIGITAL TRANSFORMATION

Gendering the G20 Information Integrity Agenda

Merrin Muhammed Ashraf, Research Assistant, IT for Change (India)

Malavika Rajkumar, Project Associate, IT for Change (India)

Julia Powles, Director, UWA Tech and Policy Lab and Tech-Crimes Coalition (Australia)

TF05

Abstract

As the UN Secretary-General (UNSG) observed in 2023, the crisis of information integrity on digital platforms cannot be addressed without effectively tackling the role of platforms in spreading "gender-based hate speech and disinformation that seek to systematically subjugate women by silencing them and pushing them out of the public sphere." The global dialogue on online gendered disinformation that UNESCO convened in January 2023 revealed the urgent need for governments and technology companies to deliver on their responsibilities to protect women's and girls' human rights online. Unfortunately, given the normalization of everyday sexism in a patriarchal society, policy responses to countering digital threats to democracy tend to sidestep the question of gender.

This paper seeks to contribute to the digital transformation agenda of G20 2024 a concrete action plan for regulating social media that creates a safer and more equitable experience for women and girls. This centers on two contributions:

- Making a case for legal recognition of gendered disinformation and gender-based hate speech based on a clear and common understanding of their key characteristics.
- Identifying the key components of a model platform regulation framework, including techno-design changes, to address the weaponization of misogyny, incitement of violence, and gendered disinformation.

The paper is informed by a legal and policy review of regulatory approaches in India, Brazil, and the European Union (EU) since these jurisdictions share democratic values and a strong commitment to building a rights-based cyberspace.

Keywords: gendered disinformation, gender-based hate speech, platform regulation, social media platforms, techno-design

Diagnosis[1]

It is widely recognized that digital platforms such as Facebook and Twitter, through their business models, architectures, and protocols, enable routinized censure and abuse against women and girls, resulting in a gendered restructuring of online communications. In 2023, the UNSG (United Nations 2023) and UNESCO (2022) highlighted gendered disinformation and gender-based hate speech targeting women and girls on digital platforms as serious threats to information integrity, requiring urgent attention from governments and technology companies. Through this policy brief, we aim to contribute to a shared strategy for gendering the G20 information integrity agenda by consolidating international expertise and analyzing existing regulatory approaches in India, Brazil, and the EU.

Platform architectures designed to perpetuate, amplify, and normalize misogyny and sexism

The lion's share of the business model of social media platforms is revenue generation through online advertising. This creates incentives to maximize 'user engagement,' even at the cost of veracity and safety. Algorithms powering information flow on these platforms are geared to amplify sensational, toxic, and harmful narratives that garner attention, which, in a patriarchal world, means that stereotyped, sexist, and misogynistic

[1] Authors would like to acknowledge the inputs of Anita Gurumurthy (Executive Director, IT for Change), Belén Luna Sanz (Policy and Advocacy Officer, HateAid), Maria Paz (Head of Legal, Policy and Research, Global Partners Digital), and Mariana G Valente (Associate Director, Internet Lab) in the writing of this policy brief.

narratives proliferate. Platforms like Facebook and Twitter enable gender-based violence (GBV) by allowing perpetrators to exploit women through inauthentic accounts, bot armies, and coordinated attacks while remaining untraceable (Gurumurthy and Dasarathy 2022).

The proliferation of gendered disinformation and gender-based hate speech comprises a systematic assault on women's human rights in digital society. Women in politics, journalism, activism, and public service are frequent targets of online GBV (Posetti and Shabbir 2022). These brutal attacks are a powerful tool of patriarchal censorship, muting women, particularly those in marginalized positions, and pushing them away from public life. Beyond personal harm, this systematic suppression of women's political presence online deeply impacts democracy. As Sarah Sobieraj (2020) notes, "digital misogyny erodes free speech and limits the diversity of speakers and ideas composing our democratic discourse." This impacts our ability, as citizens of a democracy, to make critically informed choices and reach a democratic consensus because the information we feed on may be incomplete, distorted, unreliable, and exclusionary.

Social media platforms have done very little to address the amplification of gendered disinformation and hate speech. Most have 'community guidelines' that prohibit violence, incitement, hate speech, and spreading of false information, but they suffer from lack of clarity, context-sensitivity, and proper enforcement (Gurumurthy and Dasarathy 2022). Given their business interest is attention, platforms are incentivized to delay or ignore moderating problematic posts, and often adopt different standards in different jurisdictions.

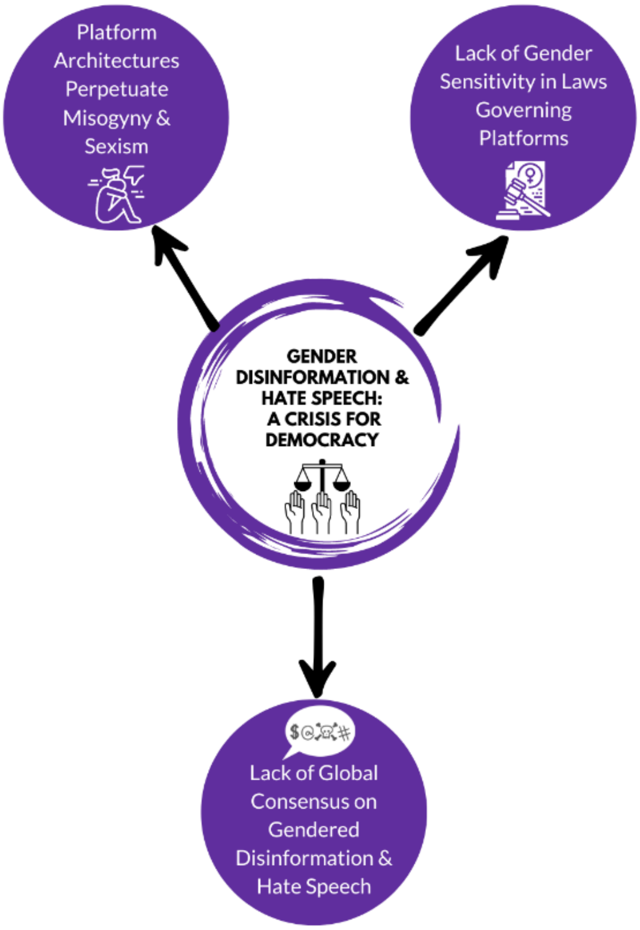
Grievance mechanisms are also tardy and ineffectual, often causing irreparable harm to the victims due to the viral nature of social media content (Amnesty International 2018).

Further, opaque platform operations vis-à-vis grievance redressal, actions taken on problematic content, and the logic of algorithms, prevent regulatory efforts to address gendered disinformation and hate speech online.

Lack of gender sensitivity in laws governing platforms

Brazil, India, and the EU have recently adopted and proposed laws and other measures to hold platforms legally responsible for moderating online content. They require platforms to remove specific items of illegal and harmful content on notice from users or a competent authority, actively moderate such content, fulfill transparency obligations, and conduct systematic risk assessments. (For an analysis of these laws, refer to Appendix.)

While the laws are differently expressed on these aspects, none effectively address gendered disinformation and hate speech. This is due to the absence of explicit gender considerations, weak platform accountability measures, and a failure to address the dominant business model and design of platform architecture and algorithms.



Platform Architectures Perpetuate Misogyny & Sexism
Lack of Gender Sensitivity in Laws Governing Platforms
GENDER DISINFORMATION & HATE SPEECH: A CRISIS FOR DEMOCRACY
Lack of Global Consensus on Gendered Disinformation & Hate Speech

FIGURE 1. Gender Disinformation and Hate Speech: A Crisis for Democracy

Absence of a shared understanding of gendered disinformation and gender-based hate speech among countries

The inefficacy of recent laws is compounded by the absence of a shared understanding of gendered disinformation and gender-based hate speech within and across jurisdictions. This has resulted in either a policy vacuum or counterproductive legislation. For instance, there is no legal definition of gendered disinformation and gender-based hate speech in India and Brazil (Gurumurthy et al. 2019; OHCHR 2023b).[2] While the recently proposed

[2] Brazil, however, in 2021 enacted Law N°14.192, which established a new type of gender violence, i.e., political violence against women, defined as "any action, conduct, or

EU Directive on Combating Violence against Women defines 'cyber incitement to violence or hatred' on grounds of sex or gender (Article 10), there is no provision addressing gendered disinformation.[3]

It is important to develop an operational definition of these terms that is intersectional and context-sensitive. The G20 must provide appropriate leadership to combat gendered disinformation and hate speech, crucial for safeguarding information integrity and democracy as a whole.

omission with the purpose of preventing, hindering, or restricting the political rights of women."

[3] For more information, see https://tinyurl.com/2smwfb9z

Recommendations

1. A globally shared language and discourse on gendered disinformation and hate speech

Urgent action is needed to recognize gendered disinformation and hate as well as the role of platforms in perpetuating such violence as critical issues for digital governance.

International and regional instruments, such as the Istanbul Convention on Combating Violence against Women, and the Convention on Elimination of All Forms of Discrimination Against Women, already recognize the interconnectedness between gender stereotypes, inequality, sexism, and violence against women. Yet, legal regimes have mostly been reluctant to recognize hate speech based on gender.

The 2021 report of the UN Special Rapporteur for Freedom of Speech recognizes ‘gendered hate speech’ as a hindrance to women's freedom of expression and argues the need to integrate ‘gender’ into the international framework of the International Covenant on Civil and Political Rights (ICCPR) for combating hate speech. Further, the 2023 report of the Special Rapporteur identifies gendered disinformation as a barrier to women’s participation in the public sphere, pointing to the often coordinated effort to gamify platform environments. Based on this report, gendered disinformation may be understood as including “false, deceptive, inaccurate or misleading narratives spread deliberately with malign intent to target, silence and discredit women and girls, and thwart their right to participation in public life” (OHCHR 2023a).

These recent international developments in the international women’s rights discourse are useful starting points for national laws and policies. States have a duty to address the systemic undermining and silencing of women and girls by the ruthless logic of digital marketplaces. Ignoring the issue delegitimizes the collective struggle and movement for gender equality everywhere.

2. Regulatory frameworks to prevent online GBV and increase accountability of platforms.

Globally, there is a push to hold social media platforms accountable for perpetuating and amplifying illegal and harmful content.[4] However, there is a distinct blindsiding of gender considerations in these efforts. This hampers efforts to tackle sexism and associated violence in the digital space. The table in Appendix compares the extant regulatory frameworks for platforms in three jurisdictions: the EU, India, and Brazil, and identifies their shortcomings in approaches to gendered disinformation and gender-based hate speech.

Below we propose some minimum standards and elements that must be adopted by governments and policymakers to strengthen platform regulatory measures to address online GBV, including gender disinformation and hate speech, and strengthen the information integrity agenda of the G20.

A. New institutional framework for platform compliance and accountability

1. A **strong platform liability framework**

- Governments should develop a strong platform liability framework grounded in human rights to hold platforms and those directly responsible for the conduct of

[4] See, UNESCO Guidelines on Regulation of Digital Platforms, Digital Services Act (European Union),and Germany's Network Enforcement Act (NetzDG),and Singapore's Online Safety (Miscellaneous Amendments) Act.

business accountable for enabling or facilitating harms, including online GBV, disinformation, hate speech, incitement to violence, and any systematic or deliberate failure to take steps to prevent or mitigate the harm.

- Platforms should be required to respond to information requests from law enforcement authorities and regulatory bodies in a time-bound manner. Government requests made to platforms and actions taken thereon should be proactively made transparent, to provide an avenue for judicial recourse.

2. An independent national regulatory body

- Governments must establish an independent regulatory body to oversee and enforce the compliance of platforms. This body should be constituted in adherence to the principles of governance systems outlined by the UNESCO guidelines (2023)—transparency, checks and balances, openness and accessibility, diverse expertise, the protection and promotion of cultural diversity, and due consideration of the local social and political context.
- The regulator should have the following legal mandates, at the minimum:
    - a) Perform investigative, inspectorial, supervisory, or other functions to ensure platforms' compliance with prescribed standards.
    - b) Establish standardized and periodic reporting mechanisms and formats for transparency reporting by platforms.
    - c) Take necessary and proportional enforcement measures, in line with international human rights law, when platforms consistently fail to implement the prescribed standards to contain the amplification of abusive and violent content.

d) Establish a database similar to the Lumen database, to collect and analyze content takedown notices along with other legal removal requests of content.[5]

e) Engage with the government, civil society, academia, platforms, and relevant public authorities and institutions to implement digital media and information capability programs to equip the public with skills and knowledge to engage with content on digital platforms critically and act in ways that are respectful of the rights of others.

B. Mandatory compliance measures for platforms

Regulation must require platforms to adopt the following measures:

1. Preventive Measures

- Conduct human rights impact assessments to identify systemic risks to the rights of women and girls, especially from marginal social locations, arising from the design/functioning of their service, including their algorithmic systems (UNESCO 2023). This should include identification of any actual or foreseeable negative effects vis-à-vis GBV.
- Provide regulators risk assessment reports for scrutiny, including for due diligence actions, before any major design changes, decisions, operational adjustments, new activities, or relationships, as well as significant events or changes within the operating environment of the platform.

[5] For more information, see https://lumendatabase.org/pages/about

- Make changes to their content moderation or recommender systems, decision-making processes, features or functioning of their services, and terms and conditions to mitigate the identified risk.
- Submit follow-up compliance reports for regulators to undertake periodic scrutiny.
- Set up human-led content moderation systems with a minimum prescribed number of locally present human moderators with expertise in local language and cultural context.

2. Documentation and transparency

- Systematically record the following information (including gender-disaggregated data) about their content moderation processes to enable independent audits of platform operations:
  a. Types of complaints received concerning content hosted, the category of rule that is violated by such content, and the data for each type, with clear identification of complaints of GBV
  b. Action taken by the platform on the complaints received and the number of links and/or extent of information removed or made inaccessible
  c. Time taken to resolve the complaint
  d. How harmful content is determined by the platform, and what actions are to be taken concerning such content
  e. Appeal procedure
  f. Number of appeals and the number of cases in which the original decision was revised

- Periodically publish transparency reports submitted to the independent regulator with details about the design, development, and deployment of content moderation systems in clear, intelligible, and unambiguous language.
- Provide access to machine-readable data, after fulfilling safeguards to protect the privacy and personal data of users, for public interest research on illegal and harmful content, including that which incites/portrays GBV.

3. Protocols for techno-design

- Institute specialized engineering teams composed of individuals of diverse genders, equipped to develop algorithmic solutions for various types of gender-based disinformation, including violent and other forms of toxic speech and harmful, stereotypical content (UNESCO 2023).
- Introduce friction through design features by adding warning labels indicating the truthfulness or falsehood of the content, and allowing users access to contextual information regarding the content or details about the user posting the same, where appropriate (Forum on Information and Democracy 2020).
- Conduct periodic friction testing to check effectiveness to avoid being too restrictive while at the same time ensuring compliance with laws.
- Implement steps like internal circuit breakers to arrest algorithmic amplification of misogynistic content as an ongoing commitment to their statutory duty of care owed to users (Abraham 2020).
- Invest in systems to predict the reshare cascades of content with reasonable accuracy, and use them to arrest the viral spread of misogynistic content.

Scenario of Outcomes

1. Shared consensus on definitions

Reaching a shared consensus and achieving international adoption of definitions for gendered disinformation and hate speech remains an ongoing challenge. Established jurisprudence from international monitoring bodies holds that restrictions on free speech should only be applied in highly exceptional circumstances. Therefore, there may arise opposition to proposing a definition specifically tailored for women and girls. The UN Special Rapporteur Irene Khan's 2021 report highlights that gendered hate speech can be prohibited under international law, citing gender equality clauses in ICCPR and a broader non-discrimination approach in human rights law. Given the urgency to ensure the full participation of women and girls in the digital society and economy, globally accepted definitions of gender-based hate and gendered disinformation can provide impetus to tackle these issues effectively in local jurisdictions, including holding platforms accountable for propagating such forms of online violence.

2. Regulatory overreach and State co-option

Effective regulation of social media platforms, with a strong accountability framework, is vital. However, the predominant regulatory approach in many jurisdictions focuses on setting out criteria and processes under which platforms are obliged to remove certain content, often on a short timeline, or face consequences, including penalties. While these measures provide immediate relief to victims and reduce harm, they are insufficient without addressing the systemic risks enabled by these platforms.

There are also two drastic consequences of such a content-focused approach in authoritarian or undemocratic regimes. First, to avoid liability, platforms may err on the side of caution and resort to overzealous censorship of content, leading to the removal of even legitimate content (Dara 2019). Second, a content-focused approach to regulation makes it vulnerable to co-option by governments seeking to suppress dissent, as they can coerce platforms to remove unfavorable content on the threat of legal sanction (United Nations General Assembly 2021).

Therefore, instead of predicating liability of platforms on single instances of failure to remove content, liability should be pinned on failure to comply with due diligence requirements and for systematic and deliberate failure to remove illegal content or arrest the spread of harmful content.

3. A caution against 'Brussels Effect'

As the EU Digital Services Act is considered "a far better law than most that have been proposed in other parts of the world" (Keller 2024) there is a possibility that other jurisdictions may directly transpose EU provisions into their laws. This is not desirable, especially concerning gender issues that demand context sensitivity. A broader, gender equality approach that recognizes women's right to public participation—one tailored for the digital—will be key to localizing efforts to tackle gendered disinformation and hate.

References

Abraham, Rohan. 2020. "Facebook working on 'virality circuit breaker' to identify fake news before it goes viral." The Economic Times. https://tinyurl.com/ye26myfn

Amnesty International. 2018. *Toxic Twitter: A Toxic Place for Women.* https://tinyurl.com/3nyc6yt3

Dara, Rishabh. (2017). "Intermediary Liability in India: Chilling Effects on Free Expression on the Internet." CIS. https://tinyurl.com/2t6knaux.

Forum on Information & Democracy. 2020. *Working Group on Infodemics*. https://tinyurl.com/2u7cs99b

Gurumurthy, Anita and Dasarathy, Amshuman. 2022. *Profitable Provocations: A Study of Abuse and Misogynistic Trolling on Twitter Directed at Indian Women in Public-political Life.* IT for Change. https://tinyurl.com/4zhkbar4

Gurumurthy, Anita et al. 2019. *Born digital, Born free? A socio-legal study on young women's experiences of online violence in South India.* IT for Change. https://tinyurl.com/5n85wawy

Keller, Daphne. 2022. "The EU's new Digital Services Act and the Rest of the World." Verfassungsblog. https://tinyurl.com/4t9zaj6f

OHCHR. 2023a. *Gendered disinformation and its implications for the right to freedom of expression – Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. https://tinyurl.com/4kt7fj6r.

OHCHR. 2023b. *In Dialogue with Brazil.* https://tinyurl.com/yf64fb6d

Posetti, Julia and Shabbir, Nabeelah. 2022. "The Chilling: A Global Study of Online Violence Against Women Journalists." International Centre for Journalists. https://tinyurl.com/bddsx6rt

Sobieraj, Sarah. 2020.*Credible Threat: Attacks against Women Online and the Future of Democracy.* New York: Oxford University Press. https://doi.org/10.1093/oso/9780190089283.001.0001

UNESCO. 2022. *Global dialogue: Online gendered disinformation.* https://tinyurl.com/yvmty7m2

UNESCO. 2023. *Guidelines for the governance of digital platforms.* https://tinyurl.com/3t5yef34

United Nations. 2023. *Our Common Agenda-Policy Brief 8: Information Integrity on Digital Platforms.* https://tinyurl.com/bdf9t43m

United Nations General Assembly. 2021. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan.* https://tinyurl.com/6ud7pa2b

Appendix:

Comparative Table of Regulatory Mechanisms - EU, Brazil, and India

G20
BRASIL 2024
BUILDING A JUST WORLD
AND A SUSTAINABLE PLANET

T20
Brasil 2024
Let's rethink the world

Let's
rethink
the world

CEBRI
Rethink Tank

FUNDAÇÃO
ALEXANDRE
DE GUSMÃO

ipea
Instituto de Pesquisa
Econômica Aplicada