


Platform accountability for gender-based cyber  
violence:  
An analysis of the Indian legal-policy landscape

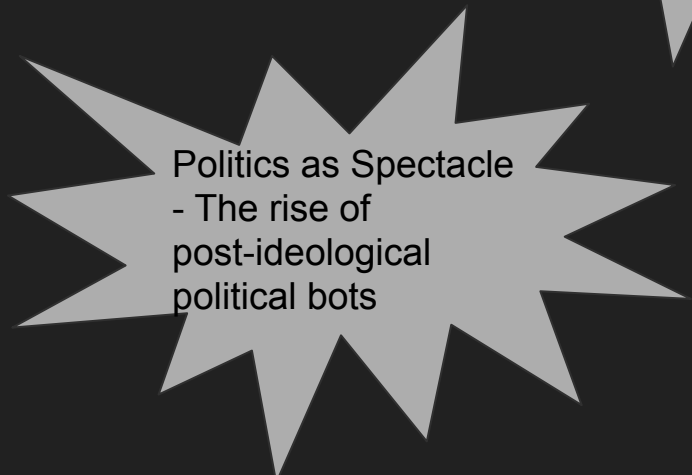
Nandini Chami, IT for Change

# The real nature of social media

Social media platforms are not passive, neutral conduits of content. Their business model is focused on active algorithmic gaming of virality to maximise user engagement with content.



Algorithmic enclaves, not just filter bubbles - QAnon



Politics as Spectacle  
- The rise of post-ideological political bots

# The real nature of social media (contd.)

The new communicative cultures of media reinforce polarized, binarised and sensationalised public discourse, intertwining with social power ideologies -- sexism and misogyny, homophobia, racism and casteism, producing hegemonic post-publics

IT for Change's 2019 survey of 881 young women aged 19-23 years across three states in India found that over 3/4th of respondents had faced gender-trolling.

31% of respondents who faced cyberviolence reported being bullied about their body shape; 30%, their weight; 27%, their looks; and 22%, their skin colour.

Though all women faced demeaning commentary about their physical attributes, women from marginal social locations faced particularly heinous forms of gender-trolling denigrating their social identity. Misogynistic vitriol faced by Dalit women is also casteist.



Immediately after [my friend] posted comments critical of the #MeToo movement in India, two-three men started piling on, saying things like, "Look at your face; you're so disgusting... nobody would even think of raping you; why are you thinking about #MeToo?"

– Female Dalit rights activist, Karnataka

# Self-regulation by social media platforms has not necessarily produced effective outcomes in terms of addressing sexist hate speech

## Facebook (April 2018 -December 2020)

- Three tier approach to classifying hate speech
- Plan for Content Oversight Board
- AI-powered detection technology to address non consensual circulation of intimate images
- Policy on targeted cursing expanded to include gender-specific terms
- Automated content moderation overhaul initiated

## Twitter (April 2018-December 2020)

- Policy on 'dehumanizing speech' introduced
- Prohibited speech categories expanded to include language on the basis of religion, caste, age, disability, disease, and also extended to cover circulation of web links

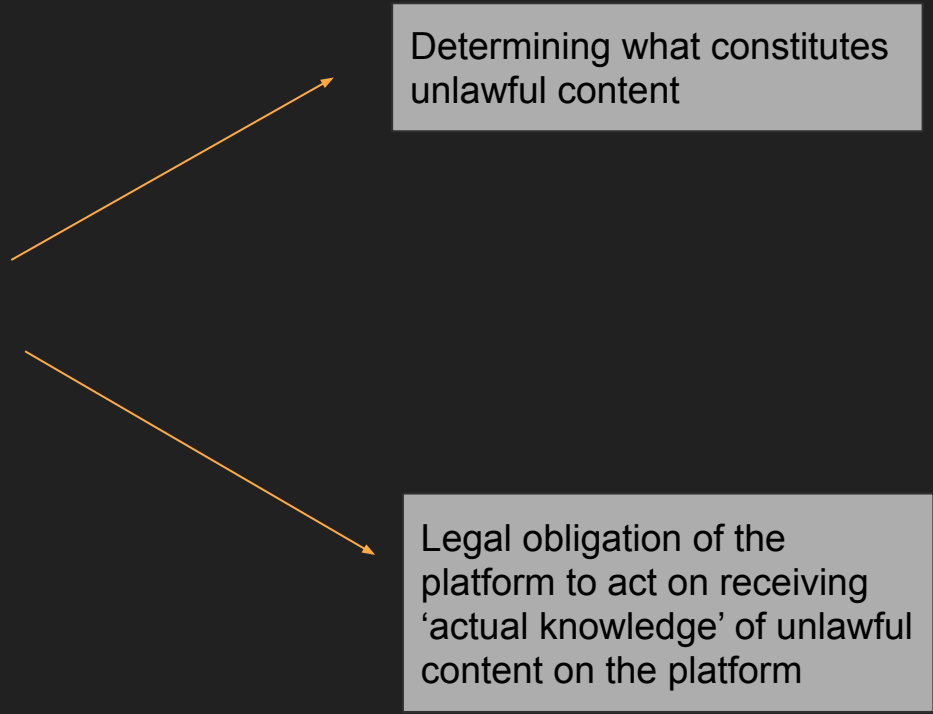
What Equality Labs' research on Facebook India (2019), one year after the updation of community standards to include a three-tier approach to hate speech, tells us:

- 93% of all hate speech posts reported to Facebook remain on Facebook. This includes content advocating violence, bullying and use of offensive slurs, and other forms of Tier 1 hate speech (violent/dehumanizing speech).

What Amnesty International's research study of gender-based hate speech (2020) on Twitter tells us:

- Despite the changes to hateful conduct rules, Twitter is not doing enough to protect women users, leading many women to silence or censor themselves on the platform. Women from ethnic or religious minorities, marginalized castes, lesbian, bisexual or transgender women, women with disabilities, as well as non-binary individuals are disproportionately impacted by abuse on the platform.

# Holding social media platforms accountable for sexist hate speech



Determining what constitutes unlawful content

Legal obligation of the platform to act on receiving 'actual knowledge' of unlawful content on the platform

# What constitutes unlawful content?

defamatory

obscene

Invasive of  
another's  
privacy

Pornographic/  
Paedophilic

Libellous

Racially or  
ethnically  
objectionable

Insulting/  
harassing on  
the basis of  
gender

IT Rules 20201 - A union of  
incompatible categories?

# Broader view of the Indian landscape on online content governance

No specific provision on gendertrolling

Criminal defamation or criminal intimidation - a very high threshold

Non-targeted attacks leading to a “death by a thousand cuts” difficult to address

Lack of proposals that are grounded in the intent to outlaw content that violates the right to equality, rather than overbroad provisions for criminalisation of ‘offensive speech’ - like the Kerala Police Act amendment proposal that was subsequently withdrawn

Whoever makes, expresses, publishes or disseminates through any kind of mode of communication, any matter or subject for threatening, abusing, humiliating or defaming a person or class of persons, knowing it to be false and that causes injury to the mind, reputation or property of such person or class of persons or any other person in whom they have interest shall on conviction, be punished with imprisonment for a term which may extend to three years or with fine which may extend to ten thousand rupees or with both.”

- Kerala Police Act amendment that was since withdrawn



# What constitutes actual knowledge of unlawful content?

How is the intermediary to act when millions of user requests are made and the intermediary is then to judge as to which of such requests are legitimate and which are not?

*Shreya Singhal v. Union of India (2015)*

The term “actual knowledge” was read down to mean that intermediaries **were only required** to act on court orders and/or a notification by the appropriate government or its agency, and not notifications by individual users (Raghavan, 2021, IT for Change series on gender-based hate speech)

# IT Rules 2021

- Clarification that acting on user complaints pertaining to the specified categories of 'unlawful content' will not expose the intermediary to legal action.
- Appointing a Grievance Redress Officer +Mandating acknowledgment of user complaints within 24 hours and disposal within 15 days
- Duty to take down content that is prima facie in the nature of intimate images that the concerned individual or their representative wants removed (24 hours window)

Attempting to address the 'actual knowledge' gap but the solution here is partial and evolving, as in the case of any legal response.

# Concluding reflections

## What works:

Clear 'duty to act' on platform intermediaries with respect to user complaints

## What needs strengthening:

A new legal provision on gender-based hate

Strengthening of the definitions of what constitutes unlawful content and vertical differentiation of obligations with respect to deplatforming

Tracking of transparency reporting

# Links for further reading: Suggested resources

IT for Change series of [papers](#) on Rethinking Legal-Institutional Responses to Sexist Hate Speech in India

IT for Change (2021). [Submission to the UN SR on FoE on gender perspectives on freedom of expression](#)

Chami, N and Kanchan, T. (2021). [A feminist social media future: How do we get there?](#) Bot Populi.

Gurumurthy, A. and Jha, B. (2020). [Articulating a feminist response to online hate speech: How do we get there?](#) Bot Populi.

IT for Change (2020). [Submission on intermediary liability rules.](#)

Write to us at [itfc@itforchange.net](mailto:itfc@itforchange.net)

Read more: [www.itforchange.net](http://www.itforchange.net)