ONLINE GENDER BASED VIOLENCE ON SHORT FORM VIDEO PLATFORMS

An inquiry into platform policies and safeguards

Online Gender Based Violence on Short Form Video Platforms: An inquiry into platform policies and safeguards

Published on 10 April 2024

Attributions:

Co-authors: Divyansha Sehgal and Lakshmi T. Nambiar
Conceptualisation: Ambika Tandon, Torsha Sarkar
Review: Amrita Sengupta and Divyank Katira
Research Assistance: Cheshta Arora
Design: Anagha Musalgaonkar

About CIS:

The Centre for Internet and Society (CIS) is a non-profit organisation that undertakes interdisciplinary research on internet and digital technologies from policy and academic perspectives. The areas of focus include digital accessibility for persons with disabilities, access to knowledge, intellectual property rights, openness (including open data, free and open source software, open standards, open access, open educational resources, and open video), internet governance, telecommunication reform, digital privacy, and cyber-security. The research at CIS seeks to understand the reconfiguration of social processes and structures through the internet and digital media technologies, and vice versa.

Through its diverse initiatives, CIS explores, intervenes in, and advances contemporary discourse and regulatory practices around internet, technology, and society in India, and elsewhere.

THE CENTRE FOR internet & society

Contents

Executive Summary

Executive summary

Being a woman or from a gender minority online is a harrowing experience. From early instances of sexual harassment in text-based internet communities in the 1990s, to apps such as Bulli Bai, and harassment in the Metaverse more recently, online gender-based violence (oGBV) is a pervasive problem, affecting 23 per cent of women globally. In India, nearly half of the women surveyed reported facing online harassment, leading to reduced online participation. Other consequences of oGBV include mental health issues, withdrawal from online spaces, and, offline violence.

In 2018, the UN Special Rapporteur on violence against women & girls, and its causes and consequences recognised online violence against women and the need to counter it, defining it as "any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of ICT, such as mobile phones and smartphones, the Internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately."

This report explores how short-form video platforms in India address oGBV by analysing their terms of service, community guidelines (CG), and reporting workflows. Recognising the role of intermediaries is crucial in understanding challenges and developing effective strategies to combat oGBV. We selected three Indian video-sharing platforms based on their download numbers, as well as Instagram reels (given their popularity in India).

The CG and terms of use of these platforms were measures against a typology of oGBV we put together based on a literature review.

The guidelines of the platforms included in the study demonstrated minimal recognition of the gendered effects of potential behaviours related to oGBV. None of the platforms had a separate policy or section dedicated to oGBV, and the policies were found to be ambiguous at several points, leaving them open to interpretation by moderators. Josh was particularly noted to have extremely poor coverage overall. Certain forms of oGBV, such as harassment, non-consensual information sharing, and extortion, were addressed to a slightly higher degree in the guidelines of Instagram, Moj, and Roposo. Some exemplary aspects are highlighted in our findings section. However, other forms, such as attacks on communication channels, omissions by regulatory actors, surveillance and stalking, and online domestic violence found little to no mention across policies, despite the likelihood of these issues manifesting offline as well. Further, policy provisions failed to address the needs of gender minorities. Reporting mechanisms were found to be lacking or inconsistent, and failed to consider the networked nature of harassment.

The harms of gendered violence are well-known and documented. The lack of clarity on implementation and policy is no longer an oversight but an active choice to disregard users.

Introduction

Introduction

In January 2022, an app claiming to "auction" Muslim women was brought to light by users on Twitter. Although it lacked payment features, the Bulli Bai app aimed to humiliate, harass, and punish the women it featured, including Muslim journalists, activists, and actress Shabana Azmi. Hosted on the Microsoft-owned open-source software platform, GitHub, the app was subsequently taken down after it attracted massive online outrage.

Being a woman or gender minority online is a harrowing experience. Bulli Bai app was not the first auction-themed app targeting the humiliation of Muslim women online. In a painfully similar and highly publicised incident in June 2021, another app with a similar auction-like user interface was taken offline by GitHub following widespread online outrage.

These are just one of a long list of online behaviours aimed at discouraging the participation of women and gender minorities in digital spaces, a pattern that has been documented since the early days of the internet. Evidence of sexual harassment dates back to 1993, when it was first seen in a text-based early internet community.[1] Additionally, cases of harassment have been reported in newer technologies like Facebook's virtual reality, Metaverse, which is currently in development and testing.[2] Harassment can only be avoided by leaving these digital spaces altogether. However, online gender-based violence (oGBV) does not always remain online. There have been documented instances of violence moving from the online to the offline, resulting in offline harm. For example, in Kerala, a trans-man took their own life after fake news about their personal life was spread on social media, resulting in severe cyber-bullying and attacks.[3] More recently, a queer teenager committed suicide after an Instagram video of them switching from a shirt to a sari attracted hateful comments and bullying; unfortunately, no action was taken by Meta or Instagram.[4]

Online spaces do not exist in isolation or outside of culture, and emerging digital norms do not counter the patriarchal systems in which digital technologies exist. Women and gender minorities often manage their online presence and under-report gendered violence to fit within what Gurumurthi and Jha term as the *"hyper-visible lakshman-rekhas (lines of propriety in women's conduct that must not be crossed) of performative online modesty"*.[5]

1. Julian Dibbel, "A Rape in Cyberspace", The Village Voice, 23 December 1993, http://www.juliandibbell.com/texts/bungle_vv.html.

2. Tanya Basu, "The Metaverse Has a Groping Problem Already", MIT Technology Review, 16 December 2021, https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/.

3. "Kerala's First Trans Man Bodybuilder Praveen Nath Dies by Suicide", *The Quint*, 5 May 2023, https://www.thequint.com/gender/kerala-first-trans-man-bodybuilder-praveen-nath-dies-by-suicide.

4. Tanishka Sodhi, "'Even after Pranshu Died, People are Writing Hateful Comments,' says Pranshu's Mother", *Newslaundry*, 30 November 2023, https://www.newslaundry.com/2023/11/30/even-after-pranshu-died-people-are-writing-hateful-comments-says-pranshus-mother.

5. Anita Gurumurthy and Bhavna Jha, "Articulating a Feminist Response to Online Hate Speech: First Steps", Botpopuli, 9 October 2020, https://botpopuli.net/articulating-a-feminist-response-to-online-hate-speech-first-steps/; Anita Gurumurthy, Amrita Vasudevan, and Nandini Chami, "Born Digital, Born Free?", IT for Change, 2019, https://itforchange.net/sites/default/files/1618/Born-Digital_Born-Free_SynthesisReport.pdf; "Forging a Survivor-Centric Approach to Online Gender-Based Violence: A Judicial Resource Guide", IT for Change, 2023, https://projects.itforchange.net/online-violence-gender-and-law-guide/about/.

There are many dimensions to the kinds of abuse women and gender minorities face online, including through auction apps, which insult the victims, mock their existence, and shame them for their views and identities. Suhana Udupa's concept of ***gali*** captures many of these dimensions, emphasising the gendered underpinnings of online abuse. Such abuse often seeks to use the logics of shame and patriarchal morality to discourage minorities from being in the public eye and to influence their participation in digital spaces, especially when they challenge existing power structures.[6]

There is an urgent need for stakeholders to recognise the negative effect of oGBV on users and the overall quality of public discourse. This report attempts to understand the perspectives of intermediaries on oGBV by analysing their terms of service, CG, and reporting workflows. Policy documents, though insufficient on their own, provide a valuable lens to understand platforms' priorities and can offer insights into how moderation decisions are made once the need arises. We seek to discover how seriously short-form video platforms in India take gender-based violence on their platforms and how they can protect their users.

6. Sahana Udupa, "*Gaali* Cultures: The Politics of Abusive Exchange on Social Media", *New Media & Society* 20, 4 (2017): 1506–1522, https://journals.sagepub.com/doi/10.1177/1461444817698776.

Background

Background

Defining gender-based violence online

In 2018, the UN Special Rapporteur on violence against women & girls, and its causes and consequences recognised online violence against women[7] and the need to counter it, defining it as *"any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of ICT, such as mobile phones and smartphones, the Internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately."*

The Special Rapporteur further emphasised that people have a right to live free from gender-based violence, the right to freedom of expression and access to information, and the right to privacy and data protection; and intermediaries can play a much bigger role in upholding their users' rights since they already *"play a central role in providing digital spaces for interaction."*

The UN Special Rapporteur used the term 'woman' expansively and inclusively to include transgender women; however, it is unclear whether non-binary and gender minorities were also included. In this paper, we extend this definition to include women, transgender persons, non-binary people, and other gender minorities.

Even before the various UN reports, civil society groups have long been advocating for better protections against oGBV, and it has shown up in key collaborative initiatives urging tech companies to do better.

The Santa Clara Principles[8] on content moderation, which many global technology companies have endorsed, advocate for companies to *"ensure that human rights and due process considerations are integrated at all stages of the content moderation process, and should publish information outlining how this integration is made."* The Human Rights and Due Process principle encourages companies to ensure that a rights-based approach is incorporated into every step of their moderation policy and its implementation. Occurrences of oGBV are a clear violation of these duties.

The Association for Progressive Communication, in collaboration with gender and sexuality activists, has created principles for a feminist internet, calling for the recognition of and cessation of all forms of oGBV so that all people can access and use the internet equitably without the threat of gendered abuse.[9]

7. "Report of the Special Rapporteur on Violence Against Women, its Causes and Consequences on Online Violence Against Women and Girls from a Human Rights Perspective", United Nations, 18 June 2018, https://www.ohchr.org/en/documents/thematic-reports/ahrc3847-report-special-rapporteur-violence-against-women-its-causes-and.

8. "Santa Clara Principles on Transparency and Accountability in Content Moderation", 2018, https://santaclaraprinciples.org

9. "Violence", Feminist Principles of the Internet, accessed 29 November 2023, https://feministinternet.org/en/principle/violence.

The violence principle states:

> *"We call on all internet stakeholders, including internet users, policymakers and the private sector, to address the issue of online harassment and technology-related violence. The attacks, threats, intimidation and policing experienced by women and queers are real, harmful and alarming, and are part of the broader issue of gender-based violence. It is our collective responsibility to address and end this."*

How does oGBV show up?

Gendered violence online is an international and pervasive problem. An Amnesty International survey of women in seven countries revealed that 23 per cent had experienced harassment online.[10] In India, according to a 2016 survey, nearly half of the women surveyed reported facing harassment online, and 28 per cent mentioned having reduced their participation in online spaces as a result.[11]

There are many ways in which oGBV appears to affect women and other gender minorities online. In her 2018 report, the Special Rapporteur on violence against women, its causes and consequences, outlined the various forms that oGBV can take, including doxing, bullying, abuse, harassment, non-consensual sharing of personal information, and incitement to physical and sexual violence.[12] These can have wide-ranging consequences for the victims, ranging from mental health issues to withdrawal from online public spaces due to threats to personal safety.

Online gender-based violence within interpersonal relationships is also extremely common, with the non-consensual sharing of private images being a prominent form of intimate partner abuse.[13]

The toll it takes on one's mental health and the social stigma experienced by victims are significant, as they may become isolated in their schools, jobs, and/or other support groups, and may engage in suicidal ideation if such images are made public. The pandemic only exacerbated the problem by forcing more people into situations they are ill-equipped to escape from.[14] Societal expectations of women lead to shaming and humiliation of women for making personal choices that compromise trust and privacy in their relationships.

Further, recognition of other intimate harms like domestic violence online can help ensure better reporting of such incidents and a more comprehensive collection of evidence in case the victim wishes to take legal action. Given that domestic violence often ends up isolating women from their in-person communities, online services and helplines may be the only resource they can turn to.

---

10. "Amnesty Reveals Alarming Impact of Online Abuse Against Women", Amnesty International, 20 November 2017, https://www.amnesty.org/en/latest/press-release/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/.

11. "'Violence' Online in India: Cybercrimes Against Women & Minorities in Social Media", Feminism in India, https://feminisminindia.com/wp-content/uploads/2021/08/Cyberviolence-Against-Women-in-India_Research-Report_FII.pdf.

12. "Report of the Special Rapporteur", United Nations.

13. Jane Anderson and Kaofeng Lee, "The Internet & Intimate Partner Violence: Technology Changes, Abuse Doesn't", Strategies, January 2017, https://aequitasresource.org/wp-content/uploads/2018/09/The-Internet-and-Intimate-Partner-Violence-Technology-Changes-Abuse-Does-Not-Issue16.pdf.

14. Jessica M Goldstein, "'Revenge Porn' was Already Commonplace. The Pandemic Has Made Things Even Worse", The Washington Post, 29 October 2020, https://www.washingtonpost.com/lifestyle/style/revenge-porn-nonconsensual-porn/2020/10/28/603b88f4-dbf1-11ea-b205-ff838e15a9a6_story.html.

However, infringements on women's safety & privacy online reduces trust in such services.[15] The UNHRC has recognised the prevalence of online forms of domestic violence and reports in other countries have found a large number of oGBV cases to link to DV.[16] Recognising these forms of oGBV in community guidelines is vital to ensuring that women are able to report these harms, and also to ensure that the forms of redress provided are nuanced and contextual to the needs of the person facing it.

Women in the public eye, including politicians, journalists, activists, and celebrities, frequently face persistent and unrestricted hate, primarily targeting their identities and seldom in response to the contents of their work. A study examining the Twitter mentions of 20 women in the political domain in India – including politicians and political commentators – revealed that every single one of them received misogynistic feedback, irrespective of their ideological leanings.[17] This recurring pattern is observed globally for journalists and activists across industries as well, such that it is often seen as the cost of doing business as a woman online.[18] [19] [20]

Research has also repeatedly shown that people with marginal caste, religion, and sexual orientation identities that intersect with gender often face far more cyberviolence for expressing their opinions.[21] [22] [23] As per a study by Amnesty International on tweets mentioning Indian women politicians, it was found that one in every seven tweets was either problematic or abusive. The study highlighted that Muslim women, those belonging to Dalit or Bahujan castes, and single women were targeted more.[25] Further, research indicates that people who identify as queer endure more cyberbullying than their heterosexual peers; within this group, the experiences of lesbian, bisexual, and transgender women are worse.[25] In India, reported instances of online violence have included the harassment and bullying of a trans bodybuilder, leading to suicide,[26] multiple instances of trans people being trolled and bullied, and misgendering.[27]

---

15. Ingrid Burdvig, Chenai Chair and Adriane van der Wilk, "COVID-19 and the increase of domestic violence against women: The pandemic of online gender-based violence", World Wide Web Foundation, https://webfoundation.org/docs/2020/07/WWWF-Submission-COVID-19-and-the-increase-of-domestic-violence-against-women-1.pdf.

16. Hannah Price, "Coronavirus: 'Revenge porn' surge hits helpline", BBC, 25 April 2020, https://www.bbc.com/news/stories-52413994; "Privacy and technology from a gender perspective: Report of the Special Rapporteur on privacy" UNHRC, 27 February 2019, https://www.ohchr.org/en/documents/thematic-reports/ahrc4063-privacy-and-technology-gender-perspective-report.

17. "Profitable Provocations: A Study of Abuse and Misogynistic Trolling on Twitter Directed at Indian Women in Public-political Life", IT for Change, July 2022, https://itforchange.net/sites/default/files/2132/ITfC-Twitter-Report-Profitable-Provocations.pdf.

18. "Amnesty Reveals the Alarming Impact of Online Abuse Against Women".

19. Gina Masullo Chen, Paromita Pain, Victoria Y Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger, "'You Really Have to Have a Thick Skin': A Cross-Cultural Perspective on how Online Harassment Influences Female Journalists", Journalism 00, 0 (2018): 1–19, https://journals.sagepub.com/doi/abs/10.1177/1464884918768500.

20. Dunja Antunovic, "'We Wouldn't Say It to Their Faces": Online Harassment, Women Sports Journalists, and Feminism", Feminist Media Studies 19, 3 (2019): 428–442, https://www.tandfonline.com/doi/abs/10.1080/14680777.2018.1446454.

21. "Facebook India: Towards the Tipping Point of Violence: Caste and Religious Hate Speech", Equality Labs, 2019, https://equalitylabs.wpengine.com/wp-content/uploads/2023/10/Facebook_India_Report_Equality_Labs.pdf

22. Mariya Salim, "Rethinking Legal-Institutional Approaches to Sexist Hate Speech in India: How Women from Marginalised Communities Navigate Online Gendered Hate and Violence", IT for Change, February 2021, https://itforchange.net/sites/default/files/1883/Mariya-Salim-Rethinking-Legal-Institutional-Approaches-To-Sexist-Hate-Speech-ITfC-IT-for-Change.pdf.

23. "Online Caste-Hate Speech: Pervasive Discrimination and Humiliation on Social Media", Centre for Internet & Society, 15 December 2021, https://cis-india.org/internet-governance/blog/online_caste-hate_speech.pdf.

24. "Troll Patrol India: Exposing Online Abuse Faced by Women Politicians in India", Amnesty International, 2020, https://decoders.blob.core.windows.net/troll-patrol-india-findings/Amnesty_International_India_Troll_Patrol_India_Findings_2020.pdf.

25. "Forging a Survivor-Centric Approach" IT for Change; Nyx McLean and Thurlo Cicero, "The Left Out Project: The Case for an Online Gender-based Violence Framework Inclusive of Transgender, Non-Binary and Gender-diverse Experiences", Gender IT, 24 August 2023, https://genderit.org/articles/left-out-project-case-online-gender-based-violence-framework-inclusive-transgender-non; Anastasia Powell, Adrian J Scott and Nicola Henry, "Digital Harassment and Abuse: Experiences of Sexuality and Gender Minority Adults", European Journal of Criminology 17, 2 (2018): 199–223, https://journals.sagepub.com/doi/full/10.1177/1477370818788006.

26. "Kerala's First Trans Man Bodybuilder Praveen Nath Dies by Suicide"

27. Ananya Desai, "Trans Rights Activist Misgendered, Trolled After Starting Online Fundraiser", The Wire, 14 June 2021, https://thewire.in/lgbtqia/trans-rights-activist-misgendered-trolled-after-starting-online-fundraiser.

Policies failing to recognise misgendering, harassment, and other harms rooted in gender identity result in difficulties in reporting – especially given the stigma attached to gender minority communities in India.[28] Second, it leaves more room for moderators to interpret policies according to their personal judgement, thus potentially resulting in lesser recognition once they have been reported and reduced chances of action being taken.

Online gender-based violence is also highly contextual depending on language variations, as well as the social and cultural nuances of violence and patriarchal standards in a particular society. Therefore, language or behaviours that may not amount to violence or harms in some contexts may do so in others. For example, slurs are often language- and region-specific. Most of this often goes unaddressed by the platforms that facilitate these communications, which succeeds in reducing the diversity of public discourse by pushing people out of public spaces.

In 2021, the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression[29] outlined how oGBV hinders women's fundamental freedoms of expression. Gendered disinformation is used to discredit women online, and state and non-state actors "weaponise public morality" arguments to censor women's expression of their experiences. Further, as more online platforms play a more important role in mediating communication, the inconsistent application of vague policies hinders the freedom of women.

> *"Vaguely worded community standards and a punitive, conservative and inconsistent approach to content moderation has led to the disproportionate censoring of female artists and artwork on themes of women's rights, which has in turn caused female artists to self-censor."*
>
> – Special Rapporteur, 2021.

Content moderation

Due to the various ways in which harmful content shows up online, content moderation becomes an important feature for a social media platform. Every social media platform needs to make decisions about what it will allow on the platform and what it will remove, regardless of ideology. Arguably, it is the defining feature of a social platform that people want to be on and interact with.[30]

Sarah Roberts defines content moderation as -

> *"the organised practice of screening user-generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction. The process can result in UGC being removed by a moderator, acting as an agent of the platform or site in question."* [31]

28. "Forging a Survivor-centric Approach to Online Gender-Based Violence", IT for Change.

29. "Gender Justice and Freedom of Expression – Report of Special Rapporteur on the Promotion and Protection of Freedom of Opinion and Expression", United Nations, 30 July 2021, https://www.ohchr.org/en/documents/thematic-reports/a76258-gender-justice-and-freedom-expression-report-special-rapporteur.

30. Nilay Patel, "Welcome to hell, Elon", The Verge, 28 October 2022, https://www.theverge.com/2022/10/28/23428132/elon-musk-twitter-acquisition-problems-speech-moderation/.

31. Sarah T. Roberts, "Content Moderation", Encyclopedia of Big Data, 2017, retrieved from https://escholarship.org/uc/item/7371c1hf

Companies will often create community guidelines and terms of service that govern how conversations can unfold on the platforms they run, and provide guardrails against the wider gambit of possible, often harmful content including explicit violence and Child Sexual Abuse Material (CSAM) that may show up.

Content moderation is inherently a subjective decision: companies need to decide what their baseline of acceptable content is and moderators need to implement the company positions even when they are vague or intentionally left unclear.

Methodology

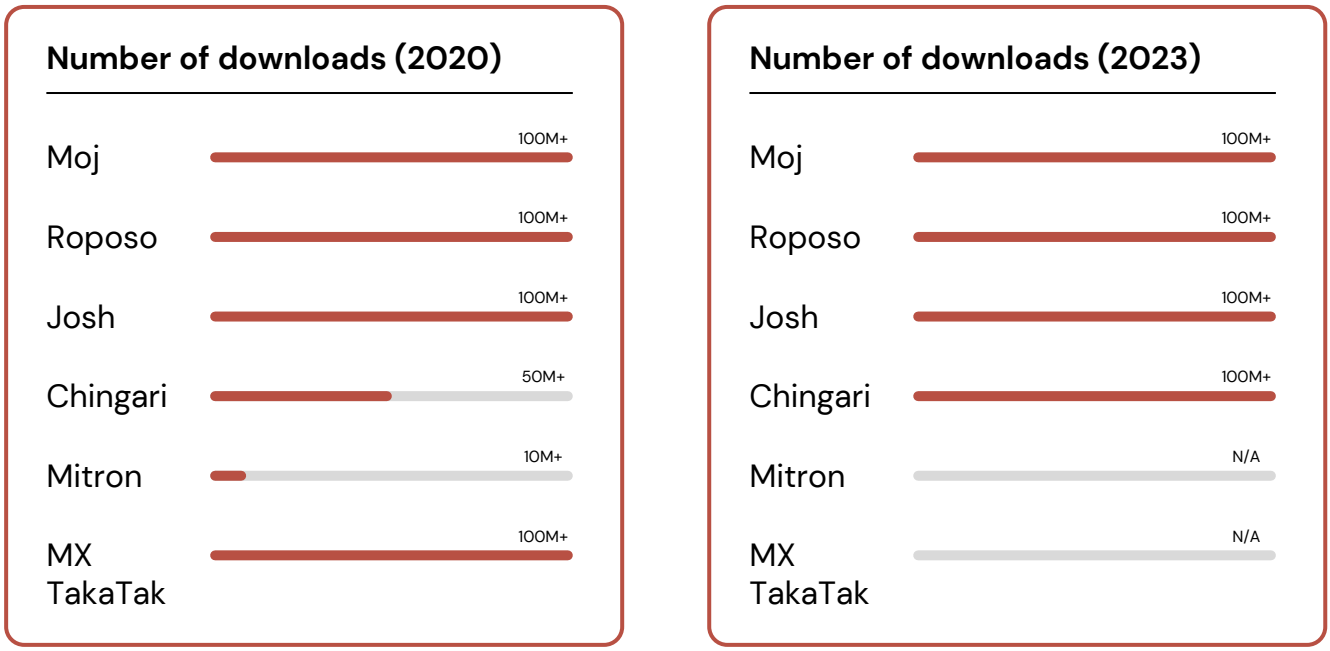
Methodology

For the purpose of our research, we sought to focus on video platforms with large Indian user bases and a substantial amount of content curated in Indian languages. By 2020, TikTok, a Chinese-made video-sharing and social networking site, had amassed more than 200 million users in India,[32] with India being its largest overseas market.[33] However, in June, in the face of escalating geopolitical tensions with China, the Indian government banned nearly 59 Chinese apps, including TikTok and several other popular and similar apps, such as Vigo Video and Likee.[34]

Subsequently, several Indian video-sharing applications have come into prominence as replacements, including Josh, Roposo, Chingari, Mitron, Moj, and MX TakaTak.[35] Table 1 provides the number of total downloads these apps have as per the data provided by the Google Play Store.

Figure 1:

Approximate number of downloads as seen on 22/12/20 and 8/12/23 on Google Play Store



32. Manish Singh, "TikTok Makes Education Push in India" *TechCrunch*, 17 October 2019, https://techcrunch.com/2019/10/17/tiktok-education-edutok-india/.

33. Manish Singh, "TikTok Goes Down in India, its Biggest Overseas Market", TechCrunch, 30 June 2020, https://techcrunch.com/2020/06/30/tiktok-goes-down-in-india-its-biggest-overseas-market/.

34. Divya Bhati, "Full List of Chinese Apps Banned in India so far: PUBG Mobile, Garena Free Fire, TikTok and Hundreds More", India Today, 21 August 2022, https://www.indiatoday.in/technology/news/story/bgmi-garena-free-fire-tiktok-and-more-banned-in-india-check-the-full-list-1990048-2022-08-19.

35. Ananya Bhattacharya, "TikTok Rip-offs Fail to Gain Traction in India as Users Still Hope that Ban will be Lifted", Scroll, 21 August 2020, https://scroll.in/article/970927/tiktok-rip-offs-fail-to-gain-traction-in-india-as-users-still-hope-that-ban-will-be-lifted.

For our analysis, we analysed four short-form video applications to study their privacy policies and community standards. These include Moj, Josh, Roposo, and Instagram Reels.

Initially, we aimed to analyse MX Takatak as well, but in Feb 2022, the app was bought by Moj's parent company, ShareChat, and was eventually merged with Moj, creating the largest short-form video platform in India.[36] At this stage, we also included Instagram Reels in our analysis to account for its growing popularity in the country and the extensive reach that the platform has.[37]

To identify the various forms that oGBV can take, we adapted the typology created by Take Back the Tech!, Luchadoras, and SocialTic, which attempts to comprehensively cover all manifestations.[38] We expanded the existing 13 categories to 15 to include 'domestic violence' and 'hate speech' as additional categories to account for the intent of the perpetrator to cause deliberate harm and their effects on victims. The final typology of oGBV harms is present in Table 2.

Figure 2:

Forms of online gender-based violence

Unauthorised access and controlling access	Unauthorised attacks to gain access to personal devices/accounts. This may be to collect data or to restrict access to a person's account. **Including:** Electronic sabotage in the form of spam or malignant viruses
Control and manipulation of information	The gathering of information or theft of information, resulting in the loss of control over such information or any unauthorised modification.
Impersonation and identity theft	Use or forgery of someone's identity without their consent. **Including:** Forging someone's identity online with the intent to defame, disparage, or spread false information about them.
Surveillance and stalking	Constant monitoring of activities, everyday life, or information (whether public or private). **Including:** 1. stalking/surveillance by family and/or intimate partners 2. stalking/surveillance with the intention to curb or control a woman's mobility
Discriminatory speech	Speech reflecting cultural models that assign women a secondary, sexualised, or strictly reproductive role (may or may not incite violence). **Including:** 1. discriminatory speech based on caste, religion, race, disability, sex, sexuality, etc. 2. speech directed at outspoken women or famous women 3. cyberbullying

36. "'MX Takatak to Merge with Sharechat' Moj to Create India's Largest Short Video Platform", *The Economic Times*, 19 February 2022, https://economictimes.indiatimes.com/tech/technology/mx-takatak-to-merge-with-sharechats-moj-to-create-indias-largest-short-video-platform/articleshow/89480061.cms.

37. Dia Rekhi, "India Is 'Lighthouse Country' for Instagram Reels: Meta Executive", The Economic Times, 28 October 2022, https://economictimes.indiatimes.com/tech/technology/india-is-lighthouse-country-for-instagram-reels-meta-executive/articleshow/95129134.cms; Rohit Shewale, "36 Instagram Reels Statistics in 2023", DemandSage, 29 November 2023, https://www.demandsage.com/instagram-reel-statistics/; Satyam Joshi, "The Rise of Instagram Reels in India: How AI-powered Short-form Videos are Changing the Game", Linkedin, 29 April 2023, https://www.linkedin.com/pulse/rise-instagram-reels-india-how-ai-powered-short-form-videos-joshi/.

38. Take Back the Tech!, Luchadoras, and SocialTic, "13 Manifestations of Gender-based Violence", Gender IT, 12 November 2018, https://www.genderit.org/resources/13-manifestations-gender-based-violence-using-technology.

	Note: In determining what constitutes discriminatory speech, the social context of that country should be relevant.
Hate speech	Speech calling for women to be murdered, raped, etc. **Including:** 1. trolling – posting of messages, uploading of images or videos, and the creation of hashtags for the purpose of annoying, provoking, or inciting violence against women and girls 2. discriminatory speech based on caste, religion, race, disability, sex, sexuality, etc. 3. speech directed at outspoken or famous women 4. hate speech after an event relating to women's rights, targeting speakers/participants 5. cyber bullying 6. encouraging other people to harass you either online or offline 7. encouraging suicide
Harassment	Repeated and unsolicited acts against a person, perceived as intrusive, disturbing, or threatening. These acts may or may not be sexualised. **Including:** 1. cyber bullying 2. sexual harassment 3. live streaming of an offline act of harassment 4. encouraging suicide
Threats	Speech/content with violent, sexually aggressive, or threatening tones that expresses an intention to harm
Non-consensual sharing of private information	Unauthorised sharing or publication of information, data, or private details about a person. **Including:** 1. Doxing 2. Live streaming of offline acts without consent Note: In determining what constitutes private information, the social context of that country and that person's situation should be relevant.
Extortion	Forcing a person to act according to another person's will through threats and intimidation regarding something of value (like personal info, intimate images, etc.). Sub-category: Sextortion – the perpetrator threatens to release intimate pictures of the victim to extort additional explicit photos, videos, sexual acts or sex from the victim.
Disparagement	Defamation, smearing, and/or undermining of the credibility, professional career, work, or public image of a person, group, or initiative through the spreading of false, manipulated, or off-topic information. **Including:** the creation of imposter profiles with the intent to disparage or defame or spread false information.
Technology-related sexual abuse and exploitation	The act of exercising power over someone based on the sexual exploitation of their pictures and/or body against their will where technology is a fundamental means. **Including:** 1. Morphed pictures/videos used to abuse and exploit (by uploading online, using as blackmail, etc.) 2. Live streaming of an offline act of violence (like a sexual assault or rape or even consensual sex without consent to live stream)
Attacks on communication channels	Deliberate tactics and actions aimed at putting a person's or group's communication or information channels out of circulation. **Including:** 1. Censorship of a person's content online 2. Preventing a person's use of social media or the internet

Omissions by regulatory actors	Contempt or lack of interest, acknowledgement or action by actors (internet intermediaries, institutions, communities) who have the possibility of regulating, resolving, and/or penalising technology-related assaults.
Domestic violence online	Domestic violence perpetrated via online mediums, either wholly or in part.

Using this typology, we analysed the policies of the four apps in October–November 2022 by checking the policies for their inclusion of the above typology. In checking these policies, we looked for – mention of the form of harm, definition of the harm, comprehensive list of acts/behaviour/content that would fall within the purview of that definition, recognition of women or gender minorities and/or acts that disproportionately affect these groups in the definition/cover of that harm, and exceptions if any. We also compared the policies with the definition put together in our typology, as well as compared categories that found good or moderate cover across platforms; the findings of good cover have been highlighted in the discussion section.

Good coverage of a particular harm is determined by the types of behaviours and actions mentioned in the guidelines. Specificity in terms of the different kinds of behaviours allowed on the platform was a big contributing factor in the categorisation. If there were examples of unacceptable behaviours provided, user actions defined, or if consequences for violations were clearly defined for a particular type of harm, we categorised the coverage as good. Another factor considered (though not necessary) for this categorisation was the acknowledgement of the gendered nature of the harm in the policy, either through the recognition of gender as a category or through the recognition of gendered harms like sexual harassment.

Moderate coverage of a harm is a more nebulous category. If there was explicit mention of a type of harm but little explanation of the kinds of behaviours that comprise it, such that comprehensive protection from this type of harm was not provided by the platform even when different parts of the policies were read together, we categorised it as moderate cover.

The bad coverage category mostly applied to Josh. Bad coverage includes vague allusions to user behaviours with no clarity or definitions. When a type of harm is mentioned, it doesn't protect the user against its occurrence. It mostly exists in name only with no explanations or definitions, leaving the norms and the implementation of these guidelines open to a high degree of interpretation. Some sections go out of their way to assign liability to the user even in case of accidents or account takeovers.

Findings

Findings

In the guidelines and terms of service of the platforms we studied, we found almost no mention of the gendered effects of potential behaviours. This meant that our analysis of various types of behaviours that make up oGBV had to rely on how comprehensively a behaviour was defined, explained, or exemplified, and if there were consequences to the behaviour outlined by our typology.

Figure 3:

Summary of Findings

Good Moderate Bad Not mentioned

	oGBV forms	Moj	Roposo	Josh	Instagram
1	Unauthorised access and controlling access				
2	Control and manipulation of information				
3	Impersonation and identity theft				
4	Discriminatory speech				
5	Surveillance and stalking				
6	Hate speech				
7	Harassment				
8	Threats				
9	Non-consensual sharing of private information				
10	Extortion				
11	Disparagement				

	oGBV forms	Moj	Roposo	Josh	Instagram
12	Technology-related sexual abuse and exploitation				
13	Attacks on communication channels				
14	Omissions by regulatory actors				
15	Domestic Violence Online				

Figure 4:

Platforms and number of issues with good, moderate, bad, and no cover each.

	Moj	Roposo	Josh	Instagram
Good cover	3	6	1	5
Moderate cover	5	2	0	3
Bad cover	3	1	9	3
Not Mentioned	4	6	5	4

Discussions

Discussions

General Discussion

None of the analysed platforms have a separate policy or section in the CG on oGBV. There is no dedicated policy documentation to counter or even define the kinds of user behaviours that could result in gendered harms. A thorough reading of the policy texts is required to determine which forms of oGBV companies expect and address through more general sections in the CG.

Certain aspects stood out across the policies of all four platforms.

First, certain forms, including harassment, non-consensual sharing of personal information, extortion, impersonation and identity theft, discriminatory and hate speech, and threats, received strong coverage in some or most of the platforms. Seom of these policies features best practices when considering the most effective ways to tackle these harms. However, a broad trend we noticed was that some forms of oGBV were not explicitly covered in the policies but were subsumed under sections that focused on other rules pertaining to acceptable and unacceptable content. For instance, discriminatory speech was largely subsumed under hate speech, and threats were not found to be a separate category, but rather, were covered under specific forms of threats – such as threatening to share non-consensual personal content or intimate images.

Second, certain forms of oGBV were not mentioned in guidelines across platforms, such as online domestic violence, surveillance and stalking, attacks on communication channels, and omissions by regulatory actors – all of which are likely to have offline impacts as well, pointing to the inadequacy of policies in dealing with such harms.

Third, some policies or provisions were formulated with more obvious solutions or redressals, whereas others were phrased more generally, potentially leaving users uncertain about the specific actions that may be taken. For example, generally across policies, provisions on intellectual property violations or legal violations specify that the content will be taken down if found to be violative. However, across platforms, policies on the non-consensual sharing of private information, harassment, discrimination, and hate speech are simply listed as content that is restricted, prohibited, content with zero tolerance, etc., with no clarity on the specific actions that will be taken in cases of user misconduct.

Fourth, there was a lack of recognition of the networked nature of harassment in several cases, which would require different structures of reporting.

Fifth, an exception found across most platforms was the exception on **public awareness/public persona grounds.** Generally, the policies mention that content that may violate some aspects of the policy – like harassment (Moj), or hate speech/threats (Instagram) – may be permitted to raise public awareness, or if they are regarding a public persona. However, what is missing here are details on how this public awareness/personal exception may be weighed against the

violation of the guideline, or whether any such balancing will be conducted at all. This leads to a lack of clarity for users, especially for famous women or outspoken women, who are targeted more online.[39] Further, there is no clarity on what makes something 'newsworthy' or turns a person into a public figure. The platforms do not define a metric of popularity or explain how it might be measured given the technical features of the platform.

Sixth, another major gap across policies is the weak acknowledgement of gender minorities. Roposo's CG is the only one that recognises 'misgendering individuals' as a harm;[40] however, their policy does not otherwise cover gender minorities. Moj and Instagram both include 'gender identity'" as a basis for hate/discriminatory speech, but they fail to include gender minorities in other aspects of their policies.

Lastly, reporting mechanisms were also found to be lacking in some cases, or misaligned with the CG, making them redundant and leaving users unclear on what actions would be taken to resolve their issue.

Discussion on forms that find good or moderate coverage across platforms

Discriminatory and hate speech >

Discriminatory and hate speech found moderate coverage across three platforms – Instagram, Roposo, and Moj. In our typology, we clarify these as separate categories, with hate speech being understood as speech that calls for violent acts such as murder, rape, etc. to be committed against women, while discriminatory speech is considered to be speech that reflects cultural models that assign women or gender minorities a secondary, sexualised, or strictly reproductive role. However, across platforms, we see that hate speech and discriminatory speech are largely conflated and covered in the same section of the CGs. Having specific coverage is important as hate speech alone does not adequately cover covert discrimination through the use of slurs, stigmatisation, or the sharing of discriminatory images.[41] Comparatively, Moj has the best coverage of this form of oGBV as it explicitly disallows discriminatory content.[42] While Roposo does not explicitly ban discriminatory content, the illustrative list of content that is banned covers aspects pertaining to discriminatory content, such as name-calling, insults, and content that is racially/ethnically objectionable or *"victimizes, harasses, degrades, or intimidates an individual or group of individuals based on religion, gender, sexual orientation, race, ethnicity, age, disability, or other legally protected basis."*[43] Instagram, while failing to account for discriminatory speech, does ban content that supports hate groups, which is a unique aspect of its policy.

39. "Amnesty Reveals Alarming Impact of Online Abuse Against Women"; "Troll Patrol India: Exposing Online Abuse Faced by Women Politicians in India".

40. "Roposo Platform Content Policy", Roposo, A.8, https://www.roposo.com/content-guidelines, accessed 13 December 2023 (Roposo CG).

41. Elena Pavan, "Internet Intermediaries and Online Gender-based Violence", in Gender, Technology and Violence, Marie Segrave and Laura Vitis (Routledge Studies in Crime and Society, 2017).

42. "Your Commitments (d)", Terms of Use, Moj, accessed 13 January 2022, https://help.mojapp.in/policies/terms/ (Moj ToU); "Content Guidelines (f)", Content and Community Guidelines, Moj, accessed 13 January 2022, https://help.mojapp.in/policies/content-policy/ (Moj CG).

43. "Roposo User Terms and Conditions 6.4 (b) and (c)", Roposo, accessed 13 January 2022, https://www.roposo.com/tnc (Roposo TnC).

Further, both Moj and Roposo prohibit hateful and discriminatory content based on several protected characteristics, including gender. This is vital as most of the categories – while covered in the policies – do not explicitly recognise gendered forms of online violence. Instagram also prohibits attacking persons based on gender, among other protected characteristics.

Both Moj and Instagram carve out an exception for content that is intended to raise awareness. Moj has the additional requirement that the content should be clearly marked as such to ensure that hateful or discriminatory content shared to raise awareness is not misinterpreted.

Where the three platforms fall short of good coverage is by failing to account for social context and regional specificities in their policies. Hate and discriminatory speech are especially prone to being region- and language-specific, and if moderators are not local to the region where the speech is being made, they could miss out on such specificities, especially when it is not accounted for in the policies.[44]

Harassment

Harassment is prohibited across the policies of Moj, Roposo, and Instagram, although it is not defined under any of the policies. However, the behaviours that comprise harassment are covered fairly comprehensively in Moj's and Roposo's policies, including contacting people who have blocked you, making other users uncomfortable, sharing content relating to abuse, self-injury, and suicide, and various acts that would amount to sexual harassment. Moj's section on harassment specifically recognises the *"emotional and psychological distress"* users may face as a consequence – which is valuable as it recognises how the user felt and the impact of the act. Roposo's policy specifically and comprehensively recognises the gendered nature of harassment, alongside other protected characteristics, listing out specific forms and acts of harassment and including gendered cover for each of these. Both Instagram and Moj specifically discuss prohibiting the encouragement of self-injury, and hence cover this form of harassment. Roposo is the most comprehensive in terms of its cover of sexual harassment; its policies ban content that contains descriptions of sexual acts, sexualised terms, and explicit descriptions of body parts.

Non-consensual personal information sharing

Moj CG and policies offer fairly comprehensive coverage against the non-consensual sharing of personal information. It is addressed as a separate category, providing an overarching definition that includes doxing and unauthorised use. It is also covered under other categories such as nudity and harassment. However, what Moj overlooks is the recognition of the gendered aspect of this harm, especially concerning the sharing of sexual imagery. Roposo's coverage of this kind of violation is stronger in this sense; although Roposo restricts this harm to a privacy harm,

42. "Your Commitments (d)", *Terms of Use*, Moj, accessed 13 January 2022, https://help.mojapp.in/policies/terms/ (Moj ToU); "Content Guidelines (f)", *Content and Community Guidelines*, Moj, accessed 13 January 2022, https://help.mojapp.in/policies/content-policy/ (Moj CG).

43. "Roposo User Terms and Conditions 6.4 (b) and (c)", Roposo, accessed 13 January 2022, https://www.roposo.com/tnc (Roposo TnC).

44. Rima Athar, "From Impunity to Justice: Improving Corporate Policies to End Technology-related Violence Against Women", Association of Progressive Communications, 9 March 2015, https://www.apc.org/en/pubs/impunity-justice-improving-corporate-policies-end-0.

gendered harms are partially covered by prohibiting the sharing of footage of sexual assault and the sharing of pornographic material for sale. Instagram's policy, while failing to cover doxing, does take strong action against the posting of intimate images of others and covers the overlap of this harm with blackmail and harassment.

The three platforms also differ in their definition of private information. Moj has an inclusive list, which, among other types of personal information, specifically includes Aadhaar, which is a rare example of a context-specific policy term.

Extortion

Instagram and Moj's cover of extortion are the most comprehensive; they both address extortion based on the sharing of personal information and intimate or sexualised images. Moj additionally recognises and prohibits extortion or blackmail by posting false information or harassing based on other protected categories (not gender).[45]

Impersonation and identity theft

Moj and Roposo clearly and explicitly prohibit impersonation and identity theft. Moj also has a detailed, good, and nuanced exception for profiles that are fake but not malignant, including *"community profiles, informative profiles and fan profiles of public figures"*. Satire and parody accounts of 'public figures' are allowed if they are clearly described as such and do not mislead users. Roposo's policy prohibits impersonation for specific purposes like deceiving, misleading, and *"communicating information which is grossly offensive or menacing in nature"*, which adds additional nuance. It also prohibits various forms of deception like misusing another's phone number or email ID, using invalid numbers or IDs, and deceptive imagery. Instagram's prohibition on deception similarly includes a reference to purpose; however, this is more general, disallowing impersonation for the purpose of violating Instagram's guidelines. While this may have broader coverage, the specificity in Roposo's guidelines is appreciated.

Threats

Roposo and Instagram both have strong coverage of threats; they detail a variety of acts that would fall under the definition of threats and acknowledge gendered and sexualised threats. Roposo's guidelines provide strong coverage for sexualised threats by prohibiting content that *"is threatening... or contains explicit or graphic descriptions or accounts of sexual acts (including but not limited to sexual language of a violent or threatening nature directed at another individual or group of individuals)"*. Instagram's CG specifically covers posting intimate images, which accounts for the gendered nature of threats and the specific use of threats as a form of sexual harassment.

Social and cultural context

A concerning trend that we noticed was the lack of contextualisation and

45. "Content Guidelines (b) and (c)", *Community Guidelines*, Moj.

localisation of CG. While three of the apps we analysed were Indian-based apps, one of them, Instagram, has a wide global usership. In their content moderation policies, none of the platforms acknowledge how social and cultural contexts play a role in whether something amounts to oGBV, the kind of oGBV that manifests, and the consequences of the same in that region. The reporting mechanisms don't always have space for the person reporting to convey context, nor do they have requirements of checking for the local socio-cultural context when evaluating if content or behaviour amounts to oGBV.

Acknowledging how sociocultural contexts affect oGBV is important because behaviour/content that may not amount to oGBV in some countries or regions within it may do so in others. For example, slurs may be extremely offensive in certain languages, but they may not mean anything offensive when translated or read in another language by a moderator who is unaware of the local context. Another aspect of oGBV where context becomes particularly relevant is in determining what constitutes private information, and therefore, what would amount to non-consensual image/information sharing. For example, posting an image of a young unmarried couple in a public place without their permission may not amount to oGBV in a Western context, but it would amount to a privacy violation in a context where dating is considered unacceptable and could even lead to societal violence and ostracisation. It is therefore vital that policies, reporting flows, and moderators consider diverse socio-cultural contexts and languages and how gender plays out in the same.

Subjective interpretation of guidelines

The lack of specificity in most of the guidelines, especially with regard to the gendered nature of the abuse faced by users, leads to uncertain expectations regarding outcomes. There is no transparency or predictability in how decisions might be taken. It is unclear what action a platform might take for users who do not follow the guidelines and for how long the consequences would last. If a post is removed, does it stay offline forever? Are there ways for users to contest the blocking?

Further, users reporting content cannot adequately predict the implementation of policies since they are vague enough to protect platforms from liability, thus leading to moderators having to make subjective decisions without any precedents or examples for consistent moderation decisions.

Discussion on IP protections versus sexual harassment protections online

Another trend we noticed was that generally, protections against IP violations were the strongest and most detailed, often with a more clear-cut remedy, as compared to protections against other harms, such as gender-based violence, and specifically harms, such as revenge porn. We use revenge porn as a specific comparator because of the similarities

between copyright violations and revenge porn violations in that both cases entail information belonging to a user being shared without their consent.[46] Further, it has been argued that the notice and takedown procedure, which is generally extremely effective in cases of copyright infringement, can also be extended to revenge porn–related issues.[47]

Instagram allows for specific copyright/trademark reports to be filed in case of IP violations. Further, Instagram assigns the user a code when they report an IP violation, which can be used to follow up on the status of the report. Roposo contains a separate section on IP in its terms of use, and it has a 'take down' mechanism for copyright infringement, which it does not offer for other violations. Moj's content policy specifies that content that *"violates the intellectual property rights of third parties will be taken down and strict action will be taken against users who are repeat defaulters"*. Nowhere else in Moj's policy is there such a strict rule that certain kinds of content will be taken down. Similarly, Josh also specifies that it retains the right to remove or disable access to content that violates *"intellectual property or other rights of VerSe and/or other third parties"* – again, the rule seems to be stricter and more specifically applied in the case of IP violations.

In contrast, the guidelines for non-consensual sharing of private information (including revenge porn) are relatively less strict and specific. The most comprehensive policies we found were those of Moj. Moj specifies that they will remove posts featuring someone's personal or intimate photos or videos shared without their permission or consent or content that is invasive of someone's privacy. Josh simply mentions that a user "may not" share content that is *"invasive of another's privacy, including bodily privacy"*. Similarly, Roposo, in both its terms of use and content guidelines, generally prohibits content that violates privacy rights, or is footage of sexual assault, but it does not clarify what action will be taken or can be taken by the user to take it down or address this issue. In none of these platforms is revenge porn recognised as a separate category of violation, despite how significant a problem it is. Instagram's CG simply state that users should not share content that they do not own or do not have the right to post. It also does not permit users to threaten to post intimate images of others, though the CG do not specify what happens if a user does post intimate images of others. While the platform specifies that it will remove content that targets private individuals to shame or degrade them, this covers revenge porn incidentally and not specifically. That being said, their reporting mechanism specifies under the category of 'nudity or sexual activity' that they remove intimate images of others that were shared without permission.

This seems to point to the fact that even in cases where remedies are similar, other harms seem to be given more importance than gendered harms.

Networked harassment

One of the glaring omissions in CG is the

46. Phillip Takhar, "A Proposal for a Notice-and-Takedown Process for Revenge Porn", *Harvard Journal of Law and Technology Digest*, 5 June 2018, https://jolt.law.harvard.edu/digest/a-proposal-for-a-notice-and-takedown-process-for-revenge-porn; Amanda Levendowski, "Using Copyright to Combat Revenge Porn", *NYU Journal of Intellectual Property & Entertainment Law* 3, no. 2 (2014): 422–446.

47. Ibid.

One of the glaring omissions in CG is the coordinated nature of the oGBV faced by women and gender minorities. The reporting workflow too only allows reporting of single posts and sometimes offers a list of predefined reasons for reporting. If a user wanted to report multiple videos and/or comments, they would need to do so individually. This is insufficient for the kind of oGBV that is becoming increasingly common given the nature of threats that users face online.

Users not only face harassment at the hands of individual actors via a single comment or video, but harassment often takes the form of a relentless barrage of negative posts or comments directed at the victim. On the receiver's end, this often looks like an unending stream of negative content, and individually reporting each post/comment and waiting for moderators to take each of them down is impractical.

This negative content is often not created by individual actors either; a group of loosely coordinated users online will often harass a victim. Marwick and Caplan have shown how these abusive behaviours can foster a community of abusers who take pride in debilitating victims and forcing them offline.[48] Further, even when content is not directed towards the victim, it can still end up creating an environment ripe for harassment. This can be seen in the context of YouTube, where Lewis et al. have shown how response videos created by individual users can turn into a dogpiling event even if they're not directly targeted at the victim.[49] Instead, they serve as motivation for the harassment of the target among a networked audience, which can then take actions into their own hands, allowing the large creator plausible deniability.
This seems to point to the fact that even in cases where remedies are similar, other harms seem to be given more importance than gendered harms.

Limitations

This report has certain limitations in terms of the sample selection of applications as well as the inclusion of gender minorities.

The process of selecting applications was done in December 2020, while the analysis was completed in December 2023. Therefore, the applications chosen are based on 2020 data. Since at the time, Chingari did not have as many downloads as the applications chosen, it was not included in the analysis; however, it does match up to these applications now.

The literature that we drew from to create the typologies of oGBV was largely female-focused. While some definitions of oGBV do include gender minorities, and we too have relied upon such definitions, the literature from which we drew the typologies did not always do so. The literature also did not specifically look at forms of oGBV faced by gender minorities who were not cis women. This could potentially result in some shortcomings, such as the non-inclusion of a specific category to cover the outing of a gender-minority person online, which may otherwise be covered in harassment and non-consensual sharing of private

48. Alice E Marwick and Robyn Caplan, "Drinking Male Tears: Language, the Manosphere, and Networked Harassment", Feminist Media Studies 18, no. 4 (2018): 543–559, https://doi.org/10.1080/14680777.2018.1450568.

49. Rebecca Lewis, Alice E Marwich, and William Clyde Partin, "'We Dissect Stupidity and Respond to It': Response Videos and Networked Harassment on YouTube", American Behavioural Scientist 65, no. 5 (2021): 735–756.

information. However, this paper features a specific discussion on the inclusion of gender minorities in CG, and it analyses these guidelines for the same.

50. Bijin Hose, "Alia Bhatt Is the Latest to Fall Prey to Deepfakes: 12 Ways to Stay Safe Online", *Indian Express*, 30 November 2023, https://indianexpress.com/article/technology/artificial-intelligence/alia-bhatt-deepfake-video-ways-to-stay-safe-online-9045902/.

Conclusion

information. However, this paper features a specific discussion on the inclusion of gender minorities in CG, and it analyses these guidelines for the same.

Conclusions

From gendered cyberbullying resulting in suicides to the impersonation of famous women using deepfakes,[50] oGBV is as rampant as ever and continues to become even more pervasive. In this paper, we sought to understand the stance taken by short-video platforms in India towards oGBV, the harms posed by it, and the recourse and protection that they offer users.

Our research focused on four such platforms – Instagram, Roposo, Moj, and Josh. We analysed the CG and terms of use of these platforms as well as the reporting flows. We tallied these against a typology of oGBV – checking for guidelines that mentioned it, the comprehensiveness of definitions, and the recognition of gendered harms. None of the analysed platforms recognised oGBV separately; however, some forms of oGBV were recognised in these platforms, with others being covered in more generalised contexts and a few forms finding no mention at all.

We noticed certain broad trends across the platforms. While Josh was an outlier with an extremely vague policy that had poor cover of oGBV, the other three platforms – Moj, Roposo, and Instagram – had some strengths in their policies in the coverage of certain forms. Some forms of oGBV are addressed to a good or moderate extent across these three platforms, including harassment, sharing of non-consensual personal information, extortion, impersonation and identity theft, and threats. Across these platforms, these forms were explicitly mentioned, well defined, and often included lists of behaviours that would fall within them; in some cases, they specifically referred to gendered nuances. Certain forms of oGBV were missing across platforms, such as online domestic violence, surveillance and stalking, attacks on communication channels, and omissions by regulatory actors – all of which are likely to have offline impacts as well, pointing to the inadequacy of policies to deal with such harms. While oGBV may be recognised to some extent, gender minorities are not specifically mentioned within the provisions.

The policies were generally found to be vague and lacking specificity, especially in terms of recognising the gendered nature of abuse, which leaves users unclear regarding the redressal they may hope to receive and if they will get any. While some provisions of CGs provide solutions or redressals, such as IP violations, others are vague and more general, even when similar redressals would apply – as in the case of non-consensual sharing of personal information. This points to the lack of solutions for these harms, which include oGBV harms. Further, the guidelines also fail to consider the coordinated nature of oGBV faced by women and gender minorities. Finally, reporting flows were also found to be lacking; in some cases, they did not align with the CG and lacked a comprehensive list of categories to report under, leaving users unclear on how reporting decisions are made, and

50. Bijin Hose, "Alia Bhatt Is the Latest to Fall Prey to Deepfakes: 12 Ways to Stay Safe Online", *Indian Express*, 30 November 2023, https://indianexpress.com/article/technology/artificial-intelligence/alia-bhatt-deepfake-video-ways-to-stay-safe-online-9045902/.

failing to give an option to track one's complaint in some cases.

All of this is doubly concerning since these platforms are second-order social media platforms. The harms of gendered violence are well documented and known by this point. So, the lack of clarity on implementation and policy is no longer an oversight but an active choice not to take care of their users.

Appendix

Appendix

Moj

Good Cover

Moj prohibits **unauthorised and controlling access of accounts** (🔒) through its terms of service by explicitly calling out that users shall not "attempt to use another user's account, username, or password without their permission." or "solicit login credentials from another user."[51] Electronic sabotage through the platform in the form of viruses or "any other computer code" is explicitly prohibited as well.

Impersonation and identity theft (🎭) is clearly prohibited in the content guidelines with a blanket policy against "*Impersonating another person (such as your family, friends, celebrities, brands or any other individuals/organisations).*" There is also a detailed and nuanced exception for profiles that are fake but not malignant in the form of "*community profiles, informative profiles and fan profiles of public figures*". Satire and parody accounts of "public figures" are also allowed if they are clearly described as such and do not mislead users. However, there is no definition of what constitutes a "public figure", nor are there any rules on multiple profiles of these public figures with an intent to harass.

Harassment (🚷) is prohibited on Moj, and there is a recognition of the "emotional or psychological distress" caused by harassing and bullying behaviours. There are also examples of behaviours that violate the anti-harassment clause and include a wide range of behaviours like abusive language, blackmailing and extortion. Even though there is no section that comprehensively covers online sexual harassment, many behaviours that comprise it are covered in the sections on Nudity and Pornography, and Harassment and Bullying that when seen together form an extensive set of guidelines than what we have observed from other India founded players in the space. There are exceptions to the harassment guidelines for "individuals who are featured in the news or tend to have a large public audience", but it is unclear what these are or how they shall be implemented.

Moderate Cover

Content Guidelines f. Hate Speech and Propaganda lays down an overview of content that attempts to cover both **hate speech** (🗯️) and **discriminatory speech** (✉️) forms of OGBV.[52]

Hate speech (🗯️) is covered through prohibition of content that *"promotes violent behaviour against"* or *"produces hatred or has the intention of creating or spreading hatred or hate propaganda"* across a number of protected characteristics including gender. Promotion of hatred and propaganda are not concepts that are easily defined,

51. Moj ToU, Safety, accessed February 2022

52. Moj CG, accessed December 2021

however, and it is unclear how these will be judged by the company stance.

We did not find any mention of further specifics like trolling. There is no separate discussion about celebrities, public figures or public events pertaining to women's rights leading to the assumption that hateful content shall be prohibited in all cases. However, given the prior discussion about "public figures" in section c. of the content guidelines, there is room to clarify what the rules for content about them might be. There is a notable exception outlined to the rules against hate speech: content that raises awareness of these issues may be exempt from takedown if it is clearly marked as such. Finally, there is no acknowledgement of regional contexts and how that plays into hate speech.

Discriminatory speech is disallowed by the same section that says

> *"We do not entertain content that spreads discrimination, intends to justify violence based on the above-mentioned attributes and refers to an individual or a group of individuals as inferior in any sense or with negative connotations."*

Again, there are no guidelines describing content about/ directed towards popular women, or socio-cultural contexts.

There is a separate section for the **non-consensual sharing of private information** but it does not focus on the gendered harms of unexpected disclosure. The behaviours mentioned classify doxing and posting images and videos without consent of the subjects as unacceptable.

Threats are generally prohibited by the Terms of Service, while the content guidelines outline problematic threats i.e. threats to reveal personal information about a user or content that threatens the country.

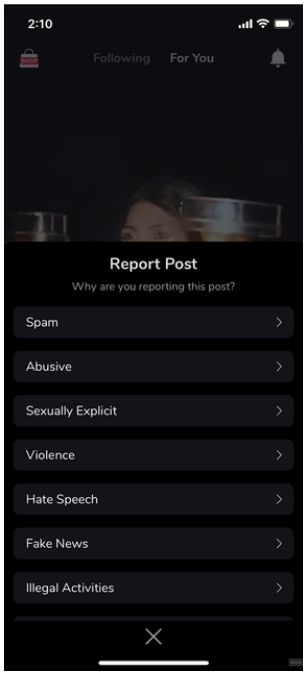
Control and Manipulation of information finds mention in a roundabout way through a blanket policy prohibiting automated means of accessing the platform in their TOS: *"You will not use any robot, spider, crawler, scraper, or other automated means or interface to access the Services or extract other user's information."* There is no recognition of intent behind the information collection, and no caveats outlined for legitimate uses of information like academic research. Moj's TOS fail to recognise that gathering and use of information is contextual.

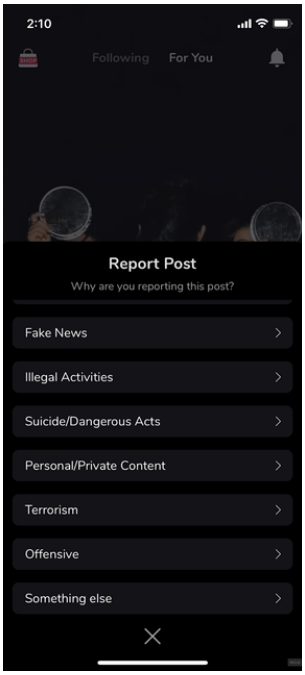
Bad Cover

Extortion, disparagement, and technology related sexual abuse and exploitation do not have separate focus in the policies. However, prior guidelines on harassment, fake profiles, misinformation, and nudity offer broad strokes protections.

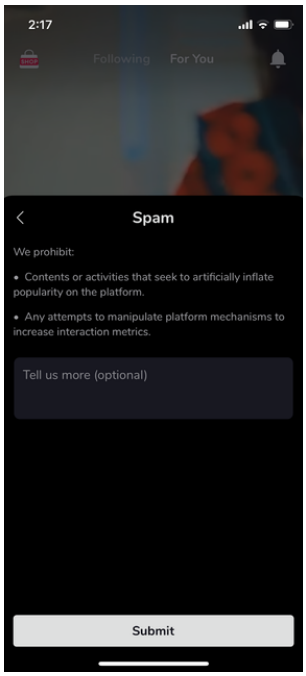
Reporting Workflows

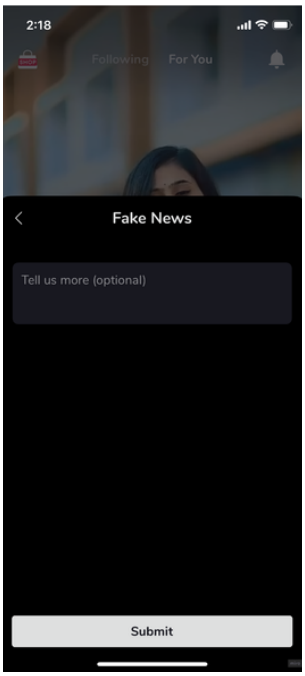
Moj allows users to select broad categories of reasons when they report a video. These categories are spam, abusive, sexually explicit content, violence, hate speech, fake news, illegal activities, suicide/dangerous acts, Personal/Private Content, Terrorism, Offensive and "something else". Out of all the Indian platforms surveyed, Moj was the most comprehensive of the Indian platforms giving users a high level of granularity to their complaints.



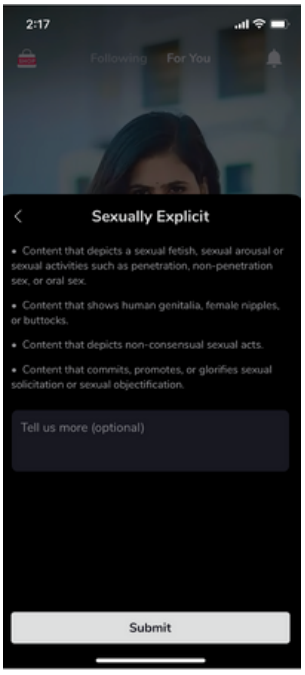


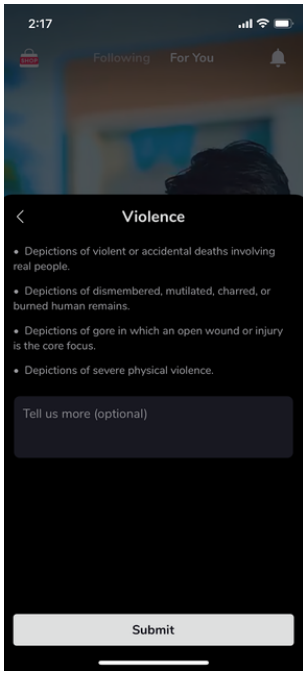
Each of these options further opens up another screen which may provide further clarity on what each term means and allows the user to enter more details in case they wish to elaborate on the reasons for reporting. As of evaluation of the screens, the options for abusive content, fake news, offensive, personal/private content, and terrorism do not have any language that clarified the kind of content that was to be reported under that category.





"Violence" emphasises the graphic and/or gory nature of the violence being perpetuated in a video. "Sexually explicit content" reads like a blanket anti-sexual content policy that also includes the first explicit mention of gender in the reporting workflow when it encourages a ban on "female nipples".





There is no mention of what happens after a complaint is submitted or how to follow up on it.

Roposo

Good Cover

Roposo's terms and conditions address **unauthorised or controlling access** quite comprehensively. The use of a Roposo account to *"breach security of another account or attempt to gain unauthorised access to another network or server"* is prohibited.[53]
Spamming is prohibited by the terms and conditions,[54] and the content guidelines.[55] Further, electronic sabotage in the form of spam or malignant viruses is also covered in the TnC which prohibits content which

> *"contains software viruses or any other computer code, files, or programs that are designed or intended to disrupt, damage, or limit the functioning of any software, hardware, or telecommunications equipment or to damage or obtain unauthorized access to any data or other information of any third party."* [56]

Impersonation or identity theft is clearly prohibited under the TnC, and specifically includes impersonation for the purposes of deception, misleading, or *"communicating information which is grossly offensive or menacing in nature"*.[57] The TnC also specifies that users are considered fraudulent if they misuse another person's phone number or email ID, or use invalid phone numbers or email IDs.[58] Additionally, the content guidelines prohibit deception and the posting of content which is *"intended to look like someone else/ some other channel is posting the content or otherwise impersonating any person or entity"*.[59]

Harassment is also specifically prohibited under both the TnC and CG. The TnC does not allow content that is harassing, or content that harasses an individual or group on the basis of *"religion, gender, sexual orientation, race, ethnicity, age, disability, or other legally protected basis"*. Further, the CG also does not permit *"not permit any violent, criminal, dangerous and obscene content directed towards any individual, gender or community at large"*[60]- within this specific forms of harassment like name calling, malicious insults, and glorifying violence including mob-lynching are also listed. Finally, the CG also includes a specific set of acts which would classify as harassment and are not permitted, including - *"content containing statements that degrades a private or group of individuals by using targeted curse/slur words or sexualised terms"*, and *"describing body parts in an explicit, obscene or otherwise hurtful manner."*[61] This listing out of forms of harassment makes the policy quite comprehensive.

Threats are prohibited under the TnC, which specifically prohibits content that *"is threatening... or contains explicit or graphic descriptions or accounts of*

53. TnC 3.4
54. Roposo TnC, 6.4(e)
55. Roposo TnC, 3
56. Roposo TnC, 6.4 (f)
57. Roposo TnC, 6.4(g)
58. Roposo TnC, 13.2
59. Roposo CG, A.3
60. Roposo CG, A.1.
61. Roposo CG, A.4

sexual acts (including but not limited to sexual language of a violent or threatening nature directed at another individual or group of individuals)". This would cover threats that are specifically sexualised as well.

Non-consensual sharing of private images is prohibited if we see it as a privacy harm, but not recognised separately as a gendered harm for most part. The TnC does not allow use of content that violates privacy rights of any user.[62] The CG also prohibits sharing content that contains[63] personal or private data of any individual. The TnC and CG recognise gendered harms somewhat partially - the TnC prohibits sharing of footage containing sexual assaults under its prohibition of *"violent, criminal, dangerous and obscene content directed towards any individual, gender or community at large"*,[64] and the CG prohibits sharing of pornographic material for the purposes of buying.

Hate Speech is covered in the Content Guidelines which does "*not permit any violent, criminal, dangerous and obscene content directed towards any individual, gender or community at large*." Within this, there are a range of examples that illustrate the various kinds of content that would be prohibited including content with prolonged name calling or malicious insults; intent to shame, deceive or insult; showing viewers how to perform activities meant to kill or harm; promoting or glorifying violence; overtly religious or political content, and illegal activities. The TnC also prohibits content that is hateful, or racially/ethnically objectionable and content that *"victimizes, harasses, degrades, or intimidates an individual or group of individuals based on religion, gender, sexual orientation, race, ethnicity, age, disability, or other legally protected basis."*[65] While this is a fairly comprehensive provision, it does not specifically address issues of hate speech levelled against women in the public eye, or hate speech directed at women's rights campaigns/events. Further, hate speech is not defined in relation to social and cultural contexts, or expressly interpreted in such a manner.

Moderate Cover

The Content Policy does not allow you to "re-order, re-purpose, modify, edit, obscure or truncate in anyway the Content, Ad-Content or ROPOSO Platform".[66] This covers control and **manipulation of information** to some extent, though it does not recognise specifically gendered harms of such acts, for example, deep fakes. Further, the terms and conditions prohibit the uploading, sharing, posting or distributing of content that *"belongs to another person and to which the user does not have any right to and/or infringes on any right of publicity, moral right, or other proprietary right of any party."*

Disparagement is prohibited under the TnC which do not allow use of content in a disparaging manner or content that is disparaging.[67] However, the TnC does not elaborate upon what constitutes disparaging content and what behaviours would come within this.

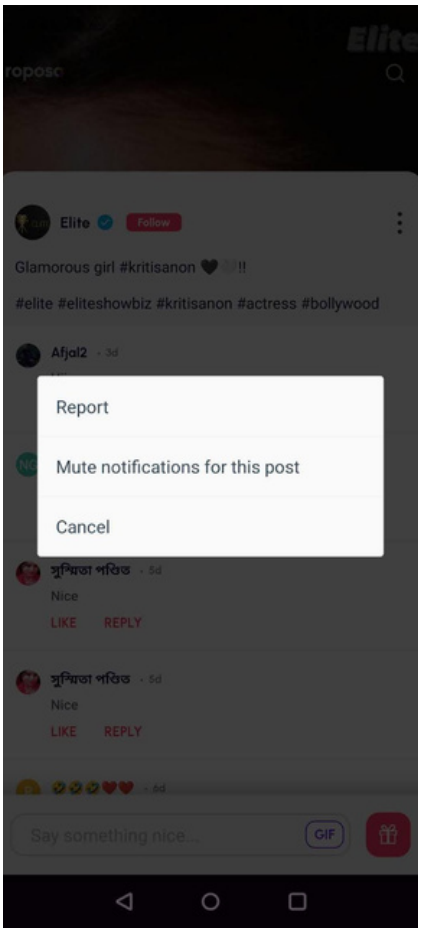
62. Roposo TnC, 5.2(a) and 6.4(a)
63. Roposo CG, A.5
64. Roposo CG, A.1.
65. Roposo TnC, 6.4(b) and (c)
66. Roposo TnC, 5.2(d)
67. Roposo TnC, 5.2(a) and 6.4(b)

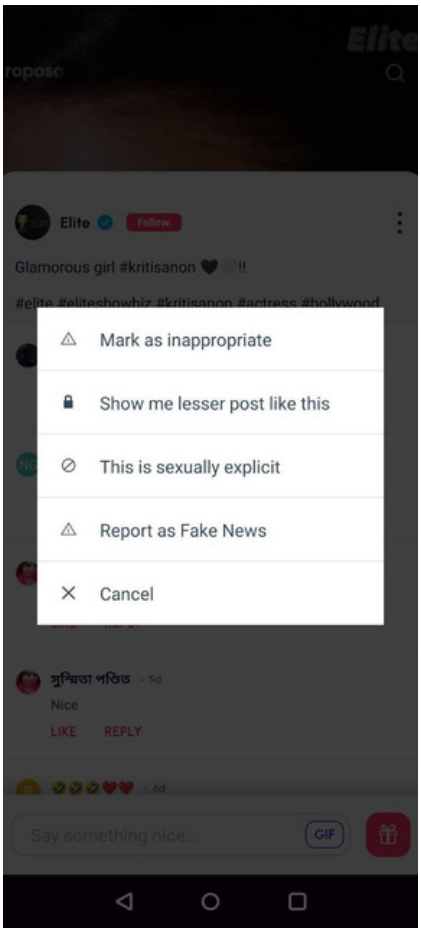
Bad Cover

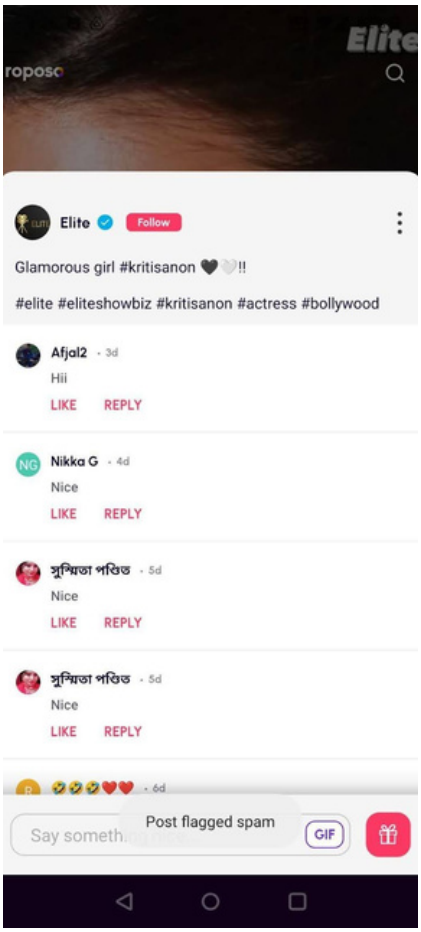
Discriminatory speech (🎭) could to some extent be covered in the same provisions that cover hate speech within content that is obscene towards a gender/community, involves name calling, malicious insults, and intends to shame/insult. However, these are all aspects that would fall within hate speech as well and additionally may constitute discriminatory speech. There is no separate categorisation of discriminatory speech and content that may not amount to hate speech, but would be discriminatory is unlikely to find cover within Roposo's policy.

Reporting Workflows

Reporting only allows for four options – inappropriate, show lesser posts like this, sexually explicit, and report as fake news. If you select any of these options, only a pop-up with *"post flagged as spam"* comes up – no other information is provided, and the post is not even flagged as per the option you choose. There is no way to follow up on your complaint.







Instagram

Good Cover

Disparagement is covered in the CG which prohibits content that targets private individuals to degrade or shame. **Impersonation and identity theft** is specifically prohibited in the ToU and CG. The CG also specifies that you can't impersonate for the purposes of violating Instagram guidelines – while this does not specifically mention intent to defame, disparage, or spread false information, this can be implied given that the CG also prohibits misinformation and content that targets private individuals to degrade or shame.

Threats are fairly comprehensively covered in the CGs - content containing credible threats is removed, and serious threats of harm to public and personal safety (including physical harm, theft, vandalism, financial harm) are not allowed. Further, threats to post intimate images are specifically covered in the CG. This is comprehensive in that it covers sexualised threats, however the CG does not specifically talk about violent, sexually aggressive content more generally – for example, threats which include threats to sexual harassment or rape are not specifically recognised separate from threats of physical harm, and thus the genered component of these acts are not recognised.

Extortion is covered by the CG providing for removal of content with personal information meant to blackmail, and including financial harm within the definition of harms caused by threats. This is fairly comprehensive, as it includes extortion based on sharing of personal information and intimate or sexualised images.

Moderate Cover

Some aspects of **control or manipulation of information** are covered through the prohibition of access or collection of information in unauthorised ways. Further, posting of another's private or confidential information without permission is also not allowed. This provision also includes that one cannot *"do anything that violates someone else's rights"* - while it could be interpreted to include rights violations in a gendered context, it is overbroad and too vague for us to conclusively say that this would be done. Manipulation of information is covered to some extent through the provision that provides that users *"can't modify, translate, create derivative works of or reverse engineer Instagram's products or their components"*, assuming that user posts are considered components of Instagram. The provisions focus excessively on the technical manipulation of the platform and does not seem to be focused on the harms that can arise out of loss of control of information.

Harassment is prohibited through a guideline that removes content *"that contains credible threats or hate speech, targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages"*. There is no definition of "credible", leaving users unclear about what would be the threshold for credibility. Encouraging suicide is also specifically prohibited in the CG. However, in all three of these there is no specific mention of gender or a recognition of sexualised forms that these violations can take.

Non-consensual sharing of private information is prohibited in the Terms of Use which does not allow posting of

"private or confidential information without permission or do anything that violates someone else's rights, including intellectual property rights ". The CG also does not allow you to *"post anything you've copied or collected from the Internet that you don't have the right to post"*. However, these provisions and their explanations, including examples of rights that may be violated, are focussed primarily on intellectual property violations, and do not talk about doxing, non-consensual sharing of intimate images, etc. Non-consensual sharing of intimate images specifically is covered by Instagram's no nudity policy, and they have zero tolerance when threatening to post intimate images of others. This form of oGBV is further covered to some extent by the guideline which removes content containing *"personal information meant to blackmail or harass"*, but this only covers non-consensual sharing when there are specific intents present, and how this intent is evaluated is unclear. A major issue that arises here is that there is no clear definition in the CG of what private, personal or confidential information is, which leaves it open as to what would be included within these categories, and fails to account for gendered and contextual nuances of what would be considered as private, personal or confidential.

The CG, which mentions one of its purposes as to foster a diverse community, specifically prohibits hate speech. Content that contains hate speech is removed, and the CG does not allow content that *"encourage(s) violence or attack(s) anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities or diseases."* The CG also does not support hate groups. The FAQs also mention that they *"do not allow attacks or abuse based on race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religion, disability or disease."* Therefore, gender and gender identity are specifically recognised in the context of hate speech. It is important to note that the CG allows for *"stronger conversation around people who are featured in the news or have a large public audience"* - this may exclude hate speech against famous or outspoken women, and it is unclear to what extent *"stronger conversation"* is allowed or what it involves. Further, another exception is that hate speech may be shared for the purposes of awareness, and the CG is clear that this intent needs to be clearly stated. Finally, there is no mention of how social context is considered and determines what constitutes hate speech.

Bad Cover

The CG provisions on hate speech which do not allow content that *"encourage(s) violence or attack(s) anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities or diseases."* can be interpreted to include some forms **discriminatory speech** (✉) within it, however it is not explicitly or comprehensively covered. There are no other provisions which prohibit discriminatory speech or behaviour that would fall within it, that would not amount to encouraging violence or attacking someone.

Unauthorised or controlling access (🔒) is not explicitly included in the CG/ToU. Both the CG and the ToU do not allow the creation of accounts or access of others' information in unauthorised ways, and do not allow you to use others' accounts. However, there is no specific mention of attempting to or actually accessing others' accounts in unauthorised ways.

There is also no mention of electronic sabotage in the form of spam or malignant viruses.

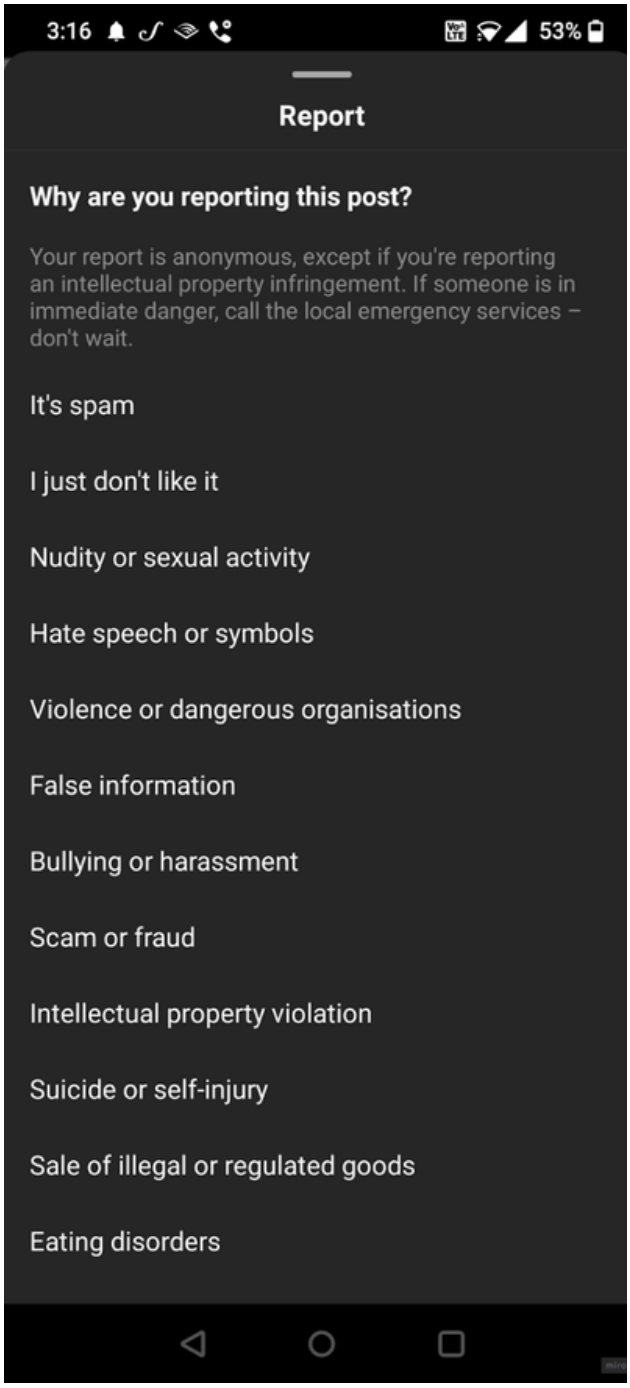
Technology-related sexual abuse and exploitation is not covered explicitly in the policy, however some forms of it are covered through various provisions. The CG prohibits sharing of graphic images for sadistic pleasure or to glorify violence, does not allow threats to share intimate images or using personal information to blackmail. However, these provisions only cover specific acts that could form a part of technology-related sexual abuse or exploitation - what this leaves out or does not recognise is the act of using technology to sexually abuse or exploit regardless of what form it may take, and these provisions lack a specific focus on sexual abuse or exploitation.

Reporting Workflows

Instagram allows the user to report under a variety of heads, as seen in the picture to the right. This includes - spam, *"I just don't like it"*, nudity or sexual activity, hate speech or symbols, violence or dangerous organisations, false information, bullying or harassment, scam or fraud, intellectual property violation, suicide or self-injury, sale of illegal or regulated goods, and eating disorders. Some of these have further sub-categories within them. Generally, once a post is reported (regardless of its category), the post is hidden from view and the user is given the option to either restrict or block the person who posted it. Advertisements have a different set of reporting categories.

The reporting workflow treats different categories differently. For some categories, such as spam and *"I just don't like it"* they just let you know that they use the reports to understand problems people have with content and show them less of that content and allow you to block/report a user.

For some categories, such as hate speech or symbols, the users submit a report and then explain what kind of posts they remove under that heading. For example, for hate speech, Instagram removes photos or videos of hate speech



or symbols, posts with captions that encourage violence or attack anyone based on who they are, and specific threats of physical harm, theft and vandalism. This is similar to what is specified in the CG as well.

Some categories require further inputs from the user and/or provide further resources/information once selected, apart from just information on what is included within a particular category or options to restrict/block the user who has posted it. For nudity or sexual activity, they inform the user that their report has been received, and is awaiting review - they review it using technology or a review team. They also inform the user that Instagram will *"send you a notification to view the outcome in your support requests as soon as possible"*. For bullying/harassment the reporting user is asked a question on who is being bullied or harassed, whether it is the user, someone they know, or someone else. For IP violations, there is a code assigned to users through which they can submit a report to the Help Centre. The Help Centre with detailed information about IP, its different forms, and links to information about violations for each of these forms. For self-harm/self-injury and for eating disorders they have a set of resources on staying safe, and how to help friends struggling with these issues.

Josh

Josh has by far the most sparse policy documents,The policies are extremely broad strokes. Josh does not recognise OGBV in any way that would be meaningful to its users.

There is a vague blanket guideline against posting content with "bad intentions" which is laughably broad and unhelpful as a measure of content moderation. Their attitude towards users and platform policies is exemplified by the following quote in the Content Policy: "If You disagree with any of these terms, please stop using the platform" and does not cover any parties, not signed up on the platform that may gain access to the content and data.

Good Cover

There is a categorical prohibition on impersonation of another person through multiple sections in the content guidelines,[68] terms of service and the user agreement. Identity theft, however, does not find explicit mention. There is no nuance in the policies for parody accounts or fan accounts of famous people. Since Josh also prohibits any *"news/ current affairs content"* there are no exceptions to this for educational or awareness content.

Bad Cover

Josh's user agreement and content guidelines disallow unauthorised and controlling access of another account in extremely vague and nonspecific terms.

The user agreement section 3.1.1 allows the use of the platform *"only in a single mobile or computer device at a time."* The content guidelines disallow accessing or posting content on the platform via *"unauthorized means including but not limited to, by using an automated device, script, bot, spider, crawler or scraper"* but do not mention mislaid passwords or hacked accounts. Neither do the guidelines or terms of service mention any way in which users can retrieve their accounts if it has been hacked or stolen or otherwise rendered inaccessible through third-party attacks.

> The terms and conditions actually place all liability on the user for keeping their account access confidential. Section 6.2 states that *"You are entirely responsible for maintaining the confidentiality of Your password and account. You agree that you are solely responsible (to us and to others) for the activity that occurs under your account."*

Further, section 6.3 emphasises that **"you"** should not share account information or use someone else's account. It is unclear how this would be enforced or who this **"you"** is supposed to be since someone accessing the account through a third party will not have their own login on the website.

This section goes on to put the blame/ liability of any content posted through a user's account on them even if the loss of access to the account is reported to the platform. The terms emphasise -

68. "Community Guidelines", *Josh*, https://share.myjosh.in/content-policy, accessed January 2022 (Josh CG).

> *"We will not be liable for any loss that You may incur as a result of unauthorized use of Your password or account. However, you could be held liable for losses incurred by the Company or another party due to someone else using your account or password."*

This only succeeds in increasing the emotional and logistic burden on the victim.

Spam is prohibited by the guidelines as well, but in context, the company is more concerned about commercial spam rather than spam meant to harass or gain user data. User harms have not been considered at all in the drafting of these policy documents.

Control and manipulation of information is prohibited too, but the context of oGBV is not clarified. Any form of manipulation that affects the functioning of the platform is prohibited, and there is an attempt to define what are some ways in which this can happen. These include the transmissions of "*viruses, worms, defects, Trojan horse, cancelbots, spyware, other items of a contaminating or destructive nature, denial of service attacks, adware, packet or IP spoofing, forged routing or electronic mail address information or similar methods or technology harmful code, flood pings, malware, bot, time bomb, worm, or other harmful or malicious component, which does or might overburden, impair or disrupt the Platform , or which does or might restrict or inhibit any third-party user's use and enjoyment of the Platform.*" [69] Unfortunately, none of these terms, which often have technical specifications and contexts are defined or clarified in scope.

VerSe, the parent company of Josh, denies all liability in cases where events are beyond the *"reasonable control of VerSe"*, and these events include *"hacking, data theft, unauthorised access to User account, impersonation, fraud, misrepresentation and so on."* Essentially, both of these forms find mention but put the onus on the victim to protect themselves and assume all liability.

There are no clear sections against **Surveillance and stalking** of and by users of the platform. The only time stalking finds a mention is in section 10.6 of the Terms and Conditions, which prohibits stalking other users or employees of the parent company VerSe. The behaviours that constitute stalking and shall thus result in consequences when identified are not defined anywhere. "*You shall not stalk, exploit, threaten, abuse or otherwise harass another User, or any VerSe employees and/or affiliates.*"

The Josh policies hint at specific keywords but do not usually have any context, or even definitions or explanations for when those keywords might become relevant for making moderation decisions or de-platforming users.

There is an allusion to the prohibition of **discriminatory speech** as well while providing no details on what constitutes discriminatory behaviour or intentions. Section 6.5 of the T&C document includes that *"You may not intimidate or harass another, or promote sexually explicit material, violence or discrimination based on race, sex, religion, nationality, disability, sexual orientation or age;"* and further in the list refers to *"any material that is racist or*

69. "JOSH Terms of Service", *Josh*, https://share.myjosh.in/terms-conditions, accessed January 2022 (Josh ToS).

discriminatory, including discrimination on the basis of someone's race, religion, age, gender, disability or sexuality;"

Josh policies do not have separate guidelines on **hate speech** (💬) **or harassment** (👥) either, but a few provisions when combined can be seen to allude to prohibiting hate speech in the broadest sense.

From the content policy asks the user to ensure that the content they post is not hateful or harassing or *"any manner abusive, obscene, defamatory, harassing, vulgar, pornographic, indecent, libellous, racist, hateful, threatening, or otherwise illegal"*. Section 6.5 of the Terms of Service also calls out *"obscene, offensive, pornographic, hateful or inflammatory"* content in the broadest of terms, and content that is "intended to harass". Intention to harass shows up multiple times in the terms of Service document but has the same problems of lack of nuance, undefined behaviours and thus inadequate predictability in moderation decisions that fail to set safe norms for the platform.

Threats (☣) are generally disallowed with a specific call out to *"threats of physical violence."*[70]

Non-consensual sharing of private information (🔒) by other users is prohibited through a patchwork of provisions. While the policies don't use the word "doxing" there is an awareness of the kinds of information that can be shared including *"addresses, phone numbers, email addresses, number and feature in the personal identity document (e.g., National Insurance numbers, passport numbers) or credit card numbers"* that can be disclosed without permissions.

However, such provisions are not extended to the company itself. The company retains the right to use any information even after the user has removed it from the platform.[71] Further, the platform does acknowledge that harm could come from other people disclosing private information, but only in the context of shunning any responsibility for it.[72]

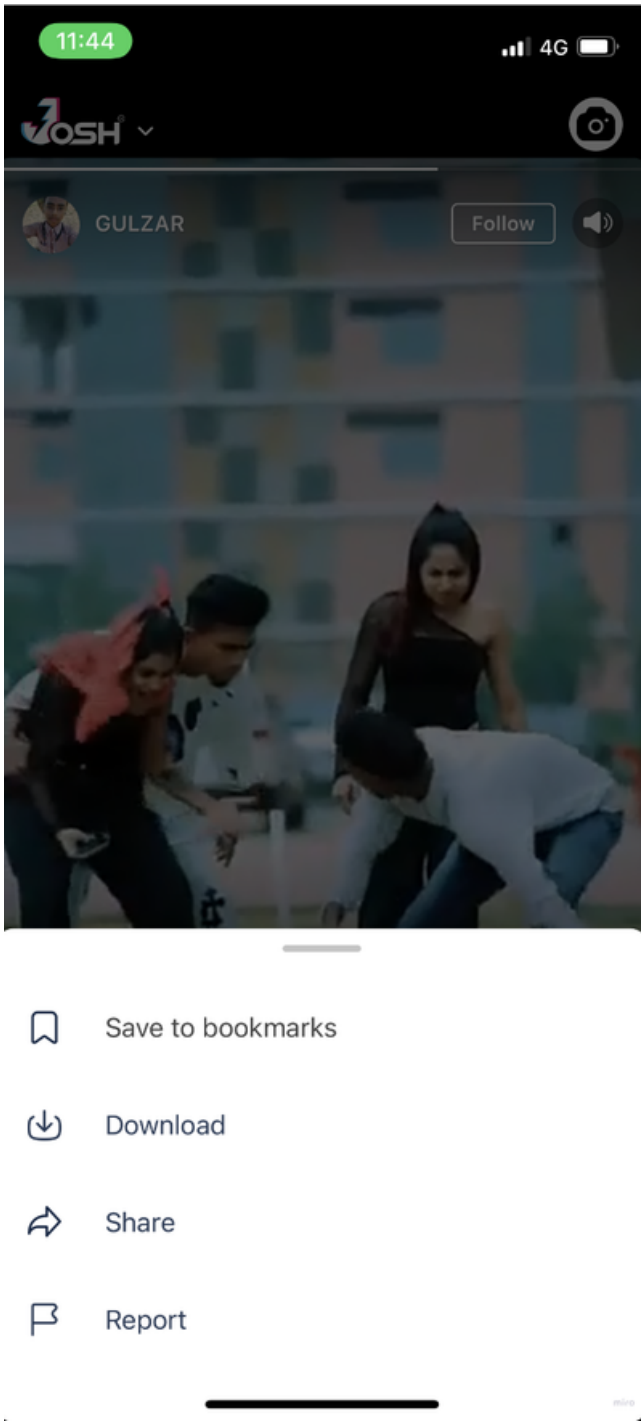
Disparagement (👎) can be seen to be prohibited only through very broad prohibitions towards defamation. Since the creation of impersonated profiles is banned as well, as a side effect, they cannot be used for disparagement.

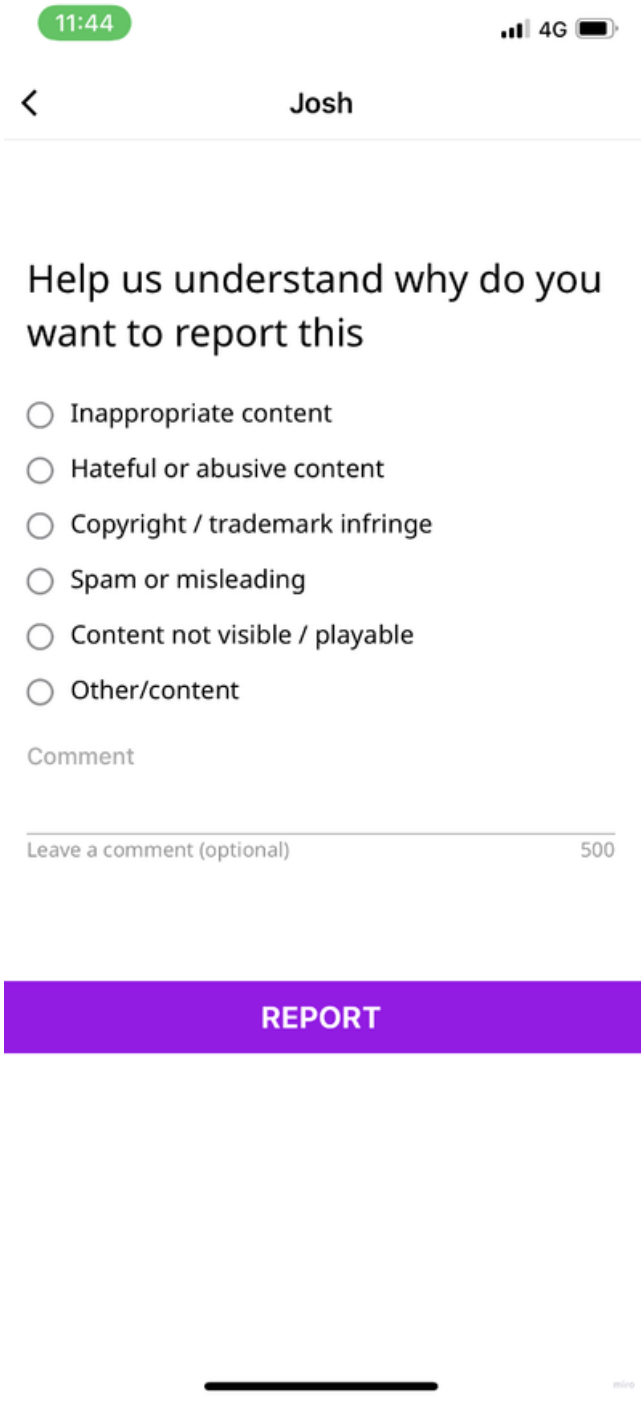
There is also a blanket policy against any sexual content or adult content, so no nuance is recognised for any content that may be for artistic merits, self-expression etc.

Reporting Workflows

The reporting workflow is a single screen that comprises of six options - inappropriate content, hateful or abusive content, copyright/trademark infringement, spam or misleading, content not visible or playable and other content. There is an optional textbox to include further information about the complaint before submitting. There is no way to follow up on complaints.

70. Josh ToS, 6.5.
71. Josh CG, 15
72. Josh ToS, 13.2


11:44
4G
GULZAR
Follow
Save to bookmarks
Download
Share
Report


11:44
4G
Josh
Help us understand why do you want to report this
Inappropriate content
Hateful or abusive content
Copyright / trademark infringe
Spam or misleading
Content not visible / playable
Other/content
Comment
Leave a comment (optional)
500
REPORT