Rethinking Legal-Institutional Approaches to Sexist Hate Speech in India

Regulating Sexist Online Harassment: A Model of Online Harassment as a Form of Censorship

Amber Sinha

IT for Change
February 2021

Regulating Sexist Online Harassment: A Model of Online Harassment as a Form of Censorship

Amber Sinha

Amber Sinha is the Executive Director of the Centre for Internet and Society, India.

This paper is part of a series under IT for Change's project, Recognize, Resist, Remedy: Combating Sexist Hate Speech Online. The series, titled Rethinking Legal-Institutional Approaches to Sexist Hate Speech in India, aims to create a space for civil society actors to proactively engage in the remaking of online governance, bringing together inputs from legal scholars, practitioners, and activists. The papers reflect upon the issue of online sexism and misogyny, proposing recommendations for appropriate legal-institutional responses. The series is funded by EdelGive Foundation, India and International Development Research Centre, Canada.

February, 2021

Conceptualisation

Anita Gurumurthy, Nandini Chami, Bhavna Jha

Editors

Anita Gurumurthy, Bhavna Jha

Editorial Support

Amay Korjan, Ankita Aggarwal, Sneha Bhagwat, Tanvi Kanchan

Design and Layout

Sneha Bhagwat

Amber Sinha

Regulating Sexist Online Harassment: A Model of Online Harassment as a Form of Censorship

Introduction

The proliferation of internet use was expected to facilitate greater online participation of women and other marginalised groups.[1] However, over the past few years, as more and more people have come online, it is evident that social power in online spaces mirrors offline hierarchies. While identity and security thefts may be universal experiences, women and the LGBTQ+ community continue to face barriers to safety that men often do not, aside from structural barriers to access. Sexist harassment pervades the online experience of women, be it on dating sites, online forums, or social media.[2]

In her book, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*,[3] Zeynep Tufekci argues that the nature and impact of censorship on social media are very different. Earlier, censorship was enacted by restricting speech. But now, it also works in the form of organised harassment campaigns, which use the qualities of viral outrage to impose a disproportionate cost on the very act of speaking out. Therefore, censorship plays out not merely in the form of the removal of speech but through disinformation and hate speech campaigns.

In most cases, this censorship of content does not necessarily meet the threshold of hate speech, and free speech advocates have traditionally argued for counter speech as the most effective response to such speech acts. However, the structural and organised nature of harassment and extreme speech

---

[1] Hilbert, Martin, Digital Gender Divide or Technologically Empowered Women in Developing Countries? A Typical Case of Lies, Damned Lies, and Statistics (November 1, 2011). Women's Studies International Forum, Vol. 34, No. 6, 2011, Available at SSRN: https://ssrn.com/abstract=2039116

[2] Lewis, Ruth & Rowe, Michael & Wiper, Clare. (2016). Online Abuse of Feminists as An Emerging form of Violence Against Women and Girls. British Journal of Criminology. 57. 10.1093/bjc/azw073.

[3] Tufekci, Zeynep. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, 2018.

often renders counter speech ineffective. This paper will explore the nature of online sexist hate and extreme speech as a mode of censorship. Online sexualised harassment takes various forms including doxxing, cyberbullying, stalking, identity theft, incitement to violence, etc. While there are some regulatory mechanisms – either in law, or in the form of community guidelines that address them, this paper argues for the need to evolve a composite framework that looks at the impact of such censorious acts on online speech and regulatory strategies to address them.

Online Harassment and the Limits of Counter Speech

Censorship has traditionally involved the participation of gatekeeping actors who control the dissemination of information. These actors, whether in response to judicial or executive orders, or in pursuance of their own editorial policies, would prohibit the dissemination of what was perceived as offending content.

Tufekci argues that censorship in the digital age is not dichotomous. Unlike the pre-digital era when gatekeepers such as governments or mass media decided whether to censor content, censorship now operates as denial of either access or attention through multiple means, "including inundating audiences with information, producing distractions to dilute their attention and focus, delegitimizing media that provide accurate information (whether credible mass media or online media), deliberately sowing confusion, fear, and doubt by aggressively questioning credibility (with or without evidence, since what matters is creating doubt, not proving a point), creating or claiming hoaxes, or generating harassment campaigns designed to make it harder for credible conduits of information to operate, especially on social media which tends to be harder for a government to control like mass media".[4] The aspect of digital censorship that we will focus on here is the generation of harassment campaigns designed to either silence or attempt to delegitimise sources of information or groups, particularly women and the LGBTQ+ community.

As specific groups or individuals, women and the LGBTQ+ community have a greater magnitude of persistent abusive and denigrating content directed towards them online with the clear intent of either intimidating them into silence or delegitimising their opinions.[5] These techniques ordinarily entail continual comments, replies, and direct messages. In some cases, they may also involve more extreme measures such as doxxing, phone calls, or setting up fake profiles.

[4] *Ibid.*

[5] Gudipaty, Nagamallika. (2017). "Gendered public spaces: Online trolling of women journalists in India". Comunicazione Politica. 18. 299-310. 10.3270/87226.

The most palpable response to what may be broadly termed problematic speech has always been counter speech. In 1927, Justice Louis Brandeis extolled the virtues of counter speech, calling "more speech and not enforced silence" as the remedy to "falsehoods and fallacies".[6] Counter speech also appears in varying forms. The most commonly referenced form of counter speech views speech as a tool of persuasion. In its online form, counter speech is most often witnessed as one user addressing another in an attempt to change their opinions, beliefs, or behaviour. The efficacy of this form of direct counter speech remains debatable and often depends on factors such as the size of the group being addressed through the act of counter speech,[7] the form of the content,[8] and the tone of the message.[9] In some cases, direct counter speech has had what is described as the "contagion effect"[10] where exposure to counter speech influences others who are not the direct addressees of the speech. There is also some research on interventions made by bystanders in an online context where others may have jumped in to defend a victim of cyberbullying.[11]

> Those who are most marginalised and disadvantaged often suffer the most at the hands of dangerous speech with the least agency to respond through counter speech, as is the case with women and the LGBTQ+ community who have been subject to sexist hate speech and misogyny online.

The proponents of counter speech, from the US Supreme Court through Justice Brandeis to the most recent defence of counter speech by Justice Kennedy in his plurality opinion in *United States v. Alvarez*,[12] give value to the power of speech. It is this belief that gives preference to counter speech over censorship, except in limited cases such as incitement to violence. Yet, counter speech is limited in its power and reach. Those who are most marginalised and disadvantaged often suffer the most at the hands of dangerous speech with the least agency to respond through counter speech, as is the case with women and the LGBTQ+ community who have been subject to sexist hate speech and

[6] *Whitney v. California*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring).

[7] Schieb and Preuss (2016). "Governing hate speech by means of counterspeech on Facebook". Conference: 66th ICA Annual Conference. https://www.researchgate.net/publication/303497937_Governing_hate_speech_by_means_of_counterspeech_on_Facebook

[8] Bartlett, Jamie and Alex Krasodomski-Jones (2015). "Counter-speech: Examining content that challenges extremism online." Demos.

[9] Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. (2018) "Countering hate speech on Facebook: The case of the Roma minority in Slovakia." Social Science Computer Review.

[10] Buerger, Cathy and Wright, Lucas. (2019). "Counterspeech: A literature review". https://dangerousspeech.org/wp-content/uploads/2019/11/Counterspeech-lit-review_complete-11.20.19-2.pdf

[11] Allison, Kimberly R. and Kay Bussey. (2016). "Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying." Children and Youth Services Review, 65, 183–194.

[12] Hudson, David Jr. (2012). "United States v. Alvarez (2012)". The First Amendment Encyclopedia. http://mtsu.edu/first-amendment/article/1479/united-states-v-alvarez#_blank

misogyny online. This poses a continual danger to their social, psychological, and economic health. A survey by Data & Society observed that "four in ten young women say they have self-censored to avoid harassment online".[13]

Arguments in favour of counter speech seem to advocate for something resembling an equilibrium which suggests that increased positive speech can somehow cancel out the harms of dangerous speech. Counter speech can contribute to the voices ensuring that dangerous speech does not become the dominant discourse, and in some instances also use speech as an important tool of persuasion. However, free speech fundamentalists ignore the very real harms caused to women and the LGBTQ+ community by dangerous sexist speech. Further, this argument assumes equity in opportunity and capacity for speech between the abuser and the abused, the harasser and the harassed. Online sexist harassment not only has a chilling effect on those who have been harassed but also on members of the community to which the harassed person belongs. The fears of harassment and abuse are most acutely felt by groups most oppressed and disadvantaged, and impact women and the LGBTQ+ community significantly. Thus, counter speech can only be effectively employed by those who have the freedom and privilege to exercise it.

A Blueprint for Regulatory Response to Sexist Dangerous Speech

In general, categories of speech seen as outside of the scope of protected speech, such as threats, hate speech, and incitement of violence are clearly regulated. However, most of the problematic behaviour online does not meet the threshold for these speech categories that are outside the purview of protected speech.

There have been several attempts to regulate cyberbullying and other forms of online harassment through criminal penalties. In the US, various states have either considered or passed legislation criminalising cyberbullying.[14] Some of these attempts, however, are based on an overly broad characterisation. For instance, cyberbullying, much like dangerous, extreme speech, includes a wide range of behaviours such as "extortion, threats, stalking, harassment, eavesdropping, spoofing

[13] Lenhart, A, Ybarra, M., Zickuhr, K, and Price-Feene, M. (2016). "Online harassment, digital abuse, and cyberstalking in America". Data & Society. https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf

[14] Lidsky, Lyrissa Barnett and Garcia, Andrea., "How Not to Criminalize Cyberbullying". (July 2, 2012). Missouri Law Review, Forthcoming, Available at SSRN: https://ssrn.com/abstract=2097684.

(impersonation), libel, invasion of privacy, fighting words, rumor-mongering, name-calling, and social exclusion".[15]

To tackle this impasse, it is may be useful to rethink the contours of protected speech online by viewing the offending act as censorious towards the capacity for self-expression of targeted demographics.

There are two core questions that one must resolve to make any progress in this line of thinking. The first is, in the broad categories of speech acts described as cyberbullying or dangerous speech, which ones constitute censorious speech acts? The second is, when one reaches a conclusion that a certain speech act is indeed censorious or chilling, what kind of regulatory response is legitimate?

The first key factor that we must keep in mind is that context matters. The scope of protected free speech depends profoundly on the context. While evaluating thresholds of free speech protection, some contextual factors must be considered. For instance, speech in schools and workplaces is subject to very different standards from speech in public spaces. More granularly, the context of the speech-makers in schools must be accounted for – are they minors or majors, are they younger or older minors, and is the speech a part of their curriculum or political speech? Further, the subject of the speech – whether it is on matters of public concern – may also determine the extent of the protection.

The second key factor that may help answer these questions is the degree of disruption the online harassment may cause. Constant abuse, for instance, is designed to create substantial disruption and may move beyond the scope of protected speech even if it does not clearly fall under incitement or hate speech.[16] This has judicial precedent in common law countries where repeated abusive behaviour in public spaces constitutes unprotected speech. Therefore, cyberbullying laws which focus on targeted and repeated abusive behaviour aiming to silence, censor, or chill the recipient may not fall foul of free speech laws.

While the standard for hate speech itself is higher, while dealing with sexist abusive speech, it may be more appropriate to use the definition of dangerous speech. In order to constitute hate speech, the speech act needs to necessarily be aimed at those who share a particular protected characteristic of a religion, race, gender etc. Hate speech is harder to regulate due to its focus on intent which is more difficult to ascertain. To address this, we can rely on Susan Benesch's dangerous speech framework. [17]

---

[15] *Ibid.*

[16] *FCC v. Pacifica Found.*, 438 U.S. 726, 739 (1978); In re S.J.NK., 647 N.W.2d 707, 712 (S.D. 2002).

[17] Benesch, S. (2013). "Dangerous Speech: A Proposal to Prevent Group Violence". *Dangerous Speech Project.* Available at: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf.

This framework focuses on a more specific category, defined not by a subjective emotion such as hatred, but by its capacity to inspire a harm. Benesch defined dangerous speech as “any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group”.

I have highlighted some key elements from this framework, which have a censorious and chilling effect on the speech rights of women and the LGBTQ+ community:

Promotion of fear: Speech acts promoting fear and malevolently portraying groups as threats.

Resorting to untruths: Speech acts which often employ disinformation to spread fear and hatred, and incite violence.

Causing direct harms: Direct harms offend, denigrate, humiliate, or frighten the people that the speech acts purport to describe.

Causing indirect harms: Indirect harms are caused by speech acts that motivate others to think and act against members of the group in question.

The primary discomfort with further regulation of online sexist speech arises from the over-criminalisation of speech. Despite speech being protected as a fundamental right in most common and civil law countries, there are a plethora of vague laws ranging from sedition to blasphemy which have severe chilling effects on speech. Although many of these laws are plagued by vagueness and overbreadth, they often continue to remain constitutionally valid on the basis of protection of ‘public order’, an equally ambiguous restriction on rights.

> The use of a higher threshold, like dangerous speech, which focuses on the impact of criminal speech allows for carving out a clearly articulated legal exception for free speech.

The use of a higher threshold, like dangerous speech, which focuses on the impact of criminal speech allows for carving out a clearly articulated legal exception for free speech.

For other forms of speech-based sexist harassment, which may not necessarily qualify the threshold for dangerous speech, it is more prudent to explore non-criminal law remedies. As stated earlier, repeat attacks, or clear patterns of behaviour intended to abuse and silence the targeted people can be regulated through a calibrated set of responses.

Regulatory agencies have considerable discretion over the task of enforcement. Broadly, they can choose between two very different enforcement strategies: 'deterrence' or 'advise and persuade',

sometimes referred to as a compliance strategy. The deterrence approach is an adversarial style of enforcement based on sanctions for rule-breaking behaviour, and is built on a model of economic theory that those regulated are rational actors who would respond to incentives and disincentives. On the other hand, the compliance strategy emphasises cooperation over confrontation and conciliation over coercion. It seeks to prevent harm rather than punish evil. Its conception of enforcement centres upon the attainment of the broad aims of legislation, rather than sanctioning their breach.[18]

It is important to have intermediate strategies like enforced self-regulation, where clear thresholds set by a responsive regulator may inform the community guidelines for content regulation on an online platform.

Given the complex nature of sexist speech online, a judicious mix of compliance and deterrence may be the optimal regulatory strategy. This requires a gradual escalation up the pyramid of regulatory tools and the existence of a credible tip that, if activated, will be sufficiently powerful to deter even the worst offenders.[19] In this case, it is important to have intermediate strategies like enforced self-regulation, where clear thresholds set by a responsive regulator may inform the community guidelines for content regulation on an online platform. The censorious nature of repeated attacks, particularly when focused on vulnerable demographics either as individuals or as a group, ought to be recognised for its impact on these individuals or groups. In such cases, we could look at calibrated responses by platforms which may include the use of technological design to ensure better access to redressal mechanisms and use pedagogical friction in response to suspicious behaviour to slow down concerted attacks. It is also imperative for the platform to provide complete transparency about its decision-making process, and allow for a robust complaint and redressal mechanism for users to challenge such decisions. This combination of recognising the censorious impact of online harassment, setting a higher threshold for criminalisation, and developing a responsive strategy for other forms of harassment could be the basis of a composite strategy for combating online sexist hate speech.

---

[18] Baldwin, Robert, Martin Cave and Martin Lodge, "The Oxford Handbook of Regulation", Oxford, Oxford University Press, 2010.

[19] Ayers, Ian. & J. Braithwaite, "Responsive Regulation: Transcending the Deregulation Debate", Oxford: Oxford University Press, 1992.