Submission of Input: Report on freedom of expression and the gender dimensions of disinformation

A. Conceptual Issues

What do you consider to be 'gendered disinformation'?

The Institute for Strategic Dialogue (ISD) uses the definition of gendered disinformation by Nina Jankowicz et al., as defined in a 2021 Wilson Center report: "a subset of online gendered abuse that uses false or misleading gender and sex-based narratives against women, often with some degree of coordination, aimed at deterring women from participating in the public sphere." This definition has guided ISD's publications on this topic to date but is supplemented by a research approach that acknowledges that gendered disinformation impacts people of all gender identities, not just women.

Over the last few years, definitions of online disinformation have typically concentrated on the 'truthfulness' of the content produced and reproduced by users of tech platforms, along with the intent with which that content is shared. This dichotomy between mis- and disinformation centers the objective of the sharer, rather than the impact on those targeted or the impact on the broader information environment.

ISD's portfolio of work on information manipulation, which includes disinformation targeting protected characteristics (for example gender), is focused on understanding and mitigating the strategies, tactics and technologies employed by actors seeking to manipulate. In this way, our objective is not only to detect false and misleading content but also to maintain an advanced understanding of the methods and behaviors used to compromise information integrity.

How is 'gendered disinformation' similar to or different from online gender-based violence?

Gendered disinformation alone is not necessarily violent, but it can act as a precursor to online gender-based violence. ISD research has found that misleading or false gender and sex-based narratives against women often prompt or encourage misogynistic vitriol, threats and violence against the woman or group of women targeted in the narratives. Across the field, there has been a concerted effort to address gendered disinformation that is sexual in nature, on account of its often graphic and egregious nature. However, some of the long-term impacts of this content are equally insidious; as such, our research approach also encompasses messaging that discourages women from political participation and reinforces gender inequalities.

Gendered disinformation and online gender-based violence are similar in the sense that they both serve to mainstream misogyny and normalize hate against women online. Gendered disinformation does so more subtly than online gender-based violence, which is why it is more commonly found on mainstream social media platforms and why gendered disinformation campaigns are able to turn everyday social media users against their targets. Additionally, a lack of public awareness about the issue of gendered disinformation means that users are less likely to spot it, call it out, or push back.

B. Responses of States, companies and organizations

What measures have States, digital companies or international organizations taken to combat 'gendered disinformation'?

In general, States, digital companies (social media platforms) and international organization consider sex, gender, and gender identity as protected attributes under their hate speech policies. They also have policies in place for violent or inciteful rhetoric, and broader policies covering mis- and disinformation which should theoretically encompass gendered disinformation. However, there seems to have been less progress in addressing this specific issue than others.

The European Union's new Digital Services Act, for example, provides novel due diligence obligations for Very Large Online Platforms. One of the new obligations may include carrying out a risk assessment on the potential for their services having "any actual or foreseeable negative effects in relation to gender-based violence," (Article 34). The recent establishment and implementation of the Global Partnership on Women's Rights, which now includes 12 different countries, also represents a positive step forward for cooperation on 'technology-facilitated gender-based violence' or TGBV.

The trust and safety teams for social media companies have reportedly not been able to settle on their own internal definitions for what constitutes violative gendered disinformation, as opposed to gendered hate speech, making it more difficult to enforce policies that address other forms of gendered and/or sexualized mis- and disinformation.

How effective have these measures been in addressing 'gendered disinformation'?

Platform policies aimed at curbing disinformation have generally been ineffective in addressing gendered disinformation. While there has been periodic attention paid to the issue, usually following seminal research publications or after incidents targeting women politicians in the US, this usually does not translate into policy change. Many policies do not account for nuances of language, which is particularly problematic when it comes to gendered disinformation since words or phrases that are weaponized against women often do not appear misogynistic at first glance but are clearly harmful in context.

Twitter's rules and policies prohibit "behavior that targets individuals or groups with abuse based on their perceived membership in a protected category." This includes women, people of color, LGBTQ+ people, and marginalized and historically underrepresented communities. According to its policies, the use of slurs or sexist tropes and hateful imagery is forbidden on the platform. However, forthcoming research by ISD found multiple examples of this exact content on Twitter, including tweets containing sexist tropes targeting Amber Heard and gendered disinformation narratives about Liz Cheney and Nancy Pelosi.

YouTube's hate speech policy claims that the platform aims to "remove content promoting violence or hatred" against individuals based on a list of attributes including gender identity or expression, sex, and gender. Content includes videos and comments, and the policy is applied more strictly to repeat offenders. Based on the same forthcoming report, YouTube still lacks efficient enforcement mechanisms when it comes to comment sections: analysts were able to find comments targeting women using derogatory and misogynistic terms such as "whore", "cunt", "bitch", and more.

Finally, the platform in ISD's recent study with the most misogynistic and abusive content and seemingly the most "hands-off" content moderation and platform policies was Telegram. The platform lacks any policies specific to misogynistic speech and images and address only calls to violence, spam and scam content, and "illegal pornographic content on publicly viewable Telegram channels." This holds true for many alternative and fringe platforms, some of which are used by male supremacists, 'men's rights activists' and incel communities to organize and spread their messaging.

Many of these platforms were found to have a problem with 'repeat offenders' of gendered disinformation and misogynistic abuse and are not taking this into account effectively when designing moderation practices. This is most often when an actor is removed from one platform, yet their content continues to circulate widely, shared by fans and at times media outlets. Platforms should implement clearer policies on how to handle content from banned users and proactively monitor the spread of their content, especially if that content violates hate speech policies.

Research by ISD has shown frequent violations of policies on multiple platforms, which are shared with the relevant companies in advance of publication where appropriate. This includes targeted campaigns against women, particularly high-profile women (politicians, celebrities, and activists). In these cases, women face a large volume of hateful content, sometimes over long periods of time, that platforms do not appear to mitigate despite violative behavior. This appears particularly true in discussion areas of platforms, such as comment sections. Recent developments within these companies, notably Twitter, has had a knock-on effect on the relationship between researchers and internal policy and trust and safety staff.

C. Finding Solutions

What issues or areas of gendered disinformation require further research in your opinion?

Researchers and practitioners should do more research on misogyny and gendered disinformation in non-English-speaking countries. Due to language and data access limitations, many high-profile studies have focused on English-speaking content. Events in non-English speaking countries that attract gendered disinformation, for example the sustained attacks on women's rights in Iran, require linguistic as well as cultural expertise to identify problematic content and narratives.

Researchers and practitioners need to keep abreast of malign actors' evolving tactics in producing, spreading, and amplifying misogynistic content, and their use of different platform features. As research from ISD and others has shown, misogynistic content takes various forms – from hateful speech and calls to violence to demeaning narratives and gendered disinformation – and continues to evolve. Analysis of extremist movements and disinformation purveyors has consistently shown that new platform features can be quickly and easily exploited by malicious actors. Researchers need to pursue opportunities to collaborate and share knowledge with policymakers and relevant stakeholders about the evolving language and tactics of online misogyny.

There needs to be further research and consideration for how women with intersecting identities are targeted with misogynistic online speech and gendered disinformation campaigns. Upcoming ISD research found that across 2022, women politicians of color, along with both trans and cis women athletes, were notable targets of abuse.

Additionally, more in-depth analysis is required to understand and characterize the interconnected nature of the online and offline dimensions of gendered disinformation and violence. For example, to ascertain whether more frequent posters of misogynistic content are more likely to perpetrate offline misogynistic behaviors. The dynamic of interplay between these on- and offline worlds would help to inform best practices for response and mitigation.

Lastly, the weaponization of deepfake pornography against women existed before the recent boom in generative artificial intelligence, but the rapid adoption of these systems in recent months could exacerbate the problem. As generative AI systems grow more sophisticated, the threat to women could become more acute. The existence of deepfake pornography, or the threat of it, may intimidate women into censoring themselves online or withdrawing from public life. The sharing of non-consensual intimate images, even if the images are fake or deceptively edited, has been proven to have adverse effects on women's mental health and wellbeing. An amendment to online safety legislation in England and Wales is expected make it illegal to share deepfake pornography. Further research into the extent of this issue and its effects is warranted.

Please provide references or links to relevant research or reports.

Hate in Plain Sight: Abuse Targeting Women Ahead of the 2022 Midterm Elections on TikTok and Instagram

- Authors: Cécile Simmons and Zoé Fourel
- Publication date: December 1, 2022
- Description: Researchers analyzed hashtag recommendations served to users on both platforms when searching for content related to several key women in US politics in the days before the election. This report finds that platforms recommend abusive hashtags when people search for the names of certain female political figures, and also promote abusive content that violate their own terms of service, showing that harmful and abusive content targeting women running for, and in-, office remains in plain sight of the platforms.

Public Figures, Public Rage: Candidate abuse on social media

- Authors: Cécile Guerin and Eisha Maharasingam-Shah

- Publication date: October 5, 2020
- Description: This report presents the findings of a research project measuring the scale of online abuse targeting a variety of Congressional candidates in the 2020 US election. It found that women and candidates from an ethnic minority background are more likely than men and those who do not have an ethnic minority background to receive abusive content on mainstream social media platforms (Facebook and Twitter).

She Drops: How QAnon Conspiracy Theories Legitimize Coordinated and Targeted Gender Based Violence

- Authors: Marc-André Argentino, Adnan Raja & Aoife Gallagher
- Publication date: October 10, 2022
- Description: In this report, based on analysis conducted in early 2021, and examining upwards of 9 million posts and mentions across Facebook, Instagram and Twitter, we examine the role of gender-based violence against celebrities who were of particular significance to the QAnon community's conspiracy theories in late 2019 and into the end of 2020: Chrissy Teigen, Tom Hanks, Ellen DeGeneres, Anderson Cooper, Jussie Smollett and Oprah Winfrey. The resulting analysis confirmed the suspicion that the most prominent type of harassment came in the form of brigading individual targets with accusations and slanderous mentions of pedophilia, often with graphic and disturbing language in their accusations.