Meedan

Meedan's submission to the consultations of the UN Special Rapporteur report on "Freedom of expression and the gender dimensions of disinformation"

The Honourable Irene Khan
UN Special Rapporteur on Freedom of Opinion and Expression, United Nations

Executive Summary & Recommendations

The Internet is our five billion person-strong town square. In a world increasingly connected both within and across geographic borders, the Internet provides a space for connection and community building, and serves a unique function in enabling communities to mobilize in times of civic importance, in crises, and in the pursuit of human rights. As such, digital conversations, and the governing principles and norms they exist within, inevitably impact our offline reality. An equitable Internet — one that is safe, inclusive and accessible — enables communities, especially those historically underserved, to access the quality information needed to mobilize, organize and foster long-term social change.

Gendered disinformation is a barrier to participation in this town square. We, at Meedan, in our commitment to making the Internet a safe and inclusive space, work directly with technology companies, social media platforms, and civil society organizations, to better understand whether existing interventions to improve Internet equity are working, where they can be improved, and how to ensure that community perspectives are integrated into policy decision-making.

Through this submission, we recommend the following considerations be taken to strengthen preparedness for, and responses to, gendered disinformation online:

1. **Ensure that no community is left behind in definitions, case studies, policies and interventions to address gendered disinformation.** To date, few country specific examinations on gendered disinformation exist.[1] [2] [3] [4] Systematic documentation of gendered disinformation is limited, especially in the Larger World context. The issue of gendered disinformation is a global one, but there is also a need for the issue to be understood and addressed from regional and hyperlocal perspectives.[5]

   We urge donor agencies and academic institutions to promote and support research and programmatic initiatives that build a contextual understanding of gendered disinformation that incorporates local nuances and experiences in order to deepen perspective and

[1] She Persisted, "#MonetizingMisogyny Country Reports", 2023, https://she-persisted.org/our-work/research-and-thought-leadership/
[2] National Democratic Institute, "Examining State-Based Disinformation Campaigns During Times of Crises", http://gendereddisinfoandshocks.demcloud.org.s3-website-us-east-1.amazonaws.com/?lang=eng
[3] Julie Posetti, "Maria Ressa: Fighting an Onslaught of Online Violence," International Center for Journalists, 2021, https://www.icfj.org/our-work/maria-ressa-big-data-analysis
[4] Digital Rights Foundation, Digital 50.50 | 2022 | Issue 2, 2022, https://digitalrightsfoundation.pk/digital-50-50-threats-and-harms-of-gendered-disinformation-targeting-women-in-public-life/
[5] Maria Giovanna Sessa, "What is Gendered Disinformation?", Heinrich Böll Foundation, 2022, https://il.boell.org/en/2022/01/26/what-gendered-disinformation

understanding of the issue, and to build pathways for hyperlocal terminology to integrate into detection and reporting systems.[6]

There is also a disproportionate focus on high impact cases facing public figures, such as women politicians and those who are in the public domain.[7] There is a need to broaden the focus to gendered disinformation attacks on trans and non-binary people and rights defenders in less visible communities, as well as those who are not necessarily in the public sphere. Relatedly, we also recommend media training for press to cover gendered disinformation responsibly, including a shared code of conduct for media publications/writers.

2. **Improve mechanisms used to flag gendered disinformation**. Social media platforms should provide accessible, trauma-informed reporting tools that allow the flagging of harmful content to center the experience and knowledge of the person being targeted or the allies reporting on their behalf, rather than relying on top-down and automated detection of harmful content. It is important that platform reporting mechanisms are accessible to vulnerable people with variable technical literacy and localized interpretations of online abuse, so they can provide the context needed for companies to make informed content moderation decisions and identify where gaps lie in their systems currently.

3. **Examine the successes where technical systems have supported gendered disinformation response efforts.** The development of technology that can monitor and detect online gendered disinformation and can group similar issues together is helpful for practitioners and researchers engaged in understanding the issue at scale. Our work with media, research and fact-checking partners in the Larger World involves the use of such tools. We recommend greater investment in the development and promotion of such tools that are open source, that enhance collaboration between different stakeholders and that make the work of practitioners efficient and scalable. Of course, when implemented poorly, such systems can further harm individuals and communities with marginalized or otherwise targeted gender identities. In this report, we emphasize the need to examine how technology (specifically machine learning and large language models developed in collaboration with hyperlocal communities) is used to reduce the prevalence and impacts of gendered disinformation and improve the detection and fact-checking workflows required to effectively respond to gendered disinformation. This is something that Meedan is currently examining in the Asia-Pacific region, with methods that can be expanded to understand other contexts.

   Such efforts can ensure that constructive dialogue, effective participation and progress in addressing power dynamics on and offline is not impeded by the gendered disinformation.

Defining Gendered Disinformation

---

[6] Maria Giovanna Sessa, “What is Gendered Disinformation?”, Heinrich Böll Foundation, 2022, https://il.boell.org/en/2022/01/26/what-gendered-disinformation

[7] Consortium for Elections and Political Process Strengthening, ‘Understanding the gender dimensions of disinformation, 2021, https://counteringdisinformation.org/topics/gender/3-current-approaches-countering-gendered-disinformation-and-addressing-gender

Gendered disinformation is a manifestation of online violence and often coexists with other forms of online violence. According to Association for Progressive Communications online violence are "acts of gender-based violence that are committed, abetted or aggravated, in part or fully, by the use of information and communication technologies (ICTs), such as mobile phones, the internet, social media platforms, and email." This often takes the form of infringement of privacy, surveillance and monitoring, damaging reputation and/or credibility, harassment, direct threats and/or violence.[8]

In the case of gendered disinformation, the mode of attack is to deliberately spread false or maligning content and to weaponize misogyny and online tools to malign, discredit and disempower targets, mislead readers and hinder an ecosystem for women and gender diverse persons to participate in the public sphere.[9]

Key Conceptual Issues

Verification and fact-checking: Gendered disinformation as distinct from abuse

Researchers have developed definitions for gendered disinformation and gender-based abuse that acknowledge the intersections between the two. Researchers point to the importance of defining gendered disinformation as distinct from gender-based abuse. The underlying infrastructure and business model that powers today's internet (i.e. incentivization of sharing and scale) has a unique impact on the spread of gendered disinformation.[10] [11] [12] [13] Further, verification and fact-checking play a distinctive role in identifying content as disinformation.

Turning the concept of gendered disinformation into something that can be measured and responded to should involve: better models for content flagging and detection that can be integrated into fact-checking workflows; stronger collaborations with community organizations that are best positioned to identify and localize key terms needed to search for harmful content; and better systems for users to report gendered disinformation, especially as it continues to contribute to offline harm.

---

[8] Association for Progressive Communications (APC), "Online gender-based violence: A submission from the Association for Progressive Communications to the United Nations Special Rapporteur on violence against women, its causes and consequences", 2017, https://www.apc.org/sites/default/files/APCSubmission_UNSR_VAW_GBV_0_0.pdf

[9] EU DisinfoLab, "Gender-Based Disinformation: Advancing Our Understanding and Response", 2021, https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response/

[10] Internet Governance Forum, "Best Practice Forum on Gender and Digital Rights: Exploring the concept of gendered disinformation", 2021, https://intgovforum.org/en/filedepot_download/248/21181

[11] EU DisinfoLab, "Gender-Based Disinformation: Advancing Our Understanding and Response", 2021, https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response/

[12] The following definitions of gendered disinformation help illuminate this distinction: "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. It combines three defining characteristics of online disinformation: falsity, malign intent, and coordination." – Nina Jankowicz, Jillian Hunchak, Alexandra Pavliuc in *Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online.* "Gendered disinformation is the spread of deceptive or inaccurate information and images against women political leaders, journalists, and female public figures." - Lucina Di Meco, Kristina Wilfore, *She Persisted*. "'Gendered disinformation' refers to information activities (creating, sharing, disseminating content) which attacks or undermines people on the basis of their gender; weaponises gendered narratives to promote political, social or economic objectives." – Ellen Judson et al., National Democratic Institute

[13] Nina Jankowicz et al.,"Malign Creativity: How gender, sex and lies are weaponized against women online," Wilson Centre, 2021, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online

Platforms have become spaces for prolific online abuse and disinformation that targets individuals or communities, and the burden of proof for reporting these issues falls mostly on users and targets. User research and surveys have pointed to rising demands for easier and transparent reporting mechanisms.[14] [15] Platform designs are a major hindrance to sound reporting mechanisms, because of poor capabilities of machine learning systems used to identify non-English languages.

A majority of social media users think platforms are not doing enough to counter online abuse.[16] At the same time, networks of and networks of fact-checking organizations, human rights defenders and archivists are already building capacity to support gendered disinformation response efforts on social media. [17] [18] [19]

Detection of gendered disinformation is inherently challenging because of the variety of content, the variance of targets from individual to identities, explicit and implicit content and different articulations of abuse, as well as misaligned incentives for social media platforms to address viral content.[20] [21] While experts agree that this indeed is a challenge, they also point that it is not impossible and should not be a reason to not have sound strategies to counter harmful content.[22]

In Meedan and PEN America's investigation examining reporting mechanisms on social media platforms, the authors refer to reporting on social media platforms as 'shouting into a void' and point to flawed designs as major barriers to effective reporting and accountability of platforms.[23] The major barriers identified were: vague definitions of violative content, no room for adding context, and poor mechanisms to report coordinated and repeated harassment. The authors recommend a series of product design fixes for platforms to adopt collectively to help make reporting more transparent, efficient, equitable, and effective, and reiterates the importance of accessible and trauma-informed reporting tools.[24] These include clear definitions of harmful content, transparent communication on reporting processes, giving users space to add context on abuse, transparency in tracking of content moderation requests, trauma-informed designs and adaptations to address coordinated and repeated harassment. Of these, the recommendation to

---

[14] Anti-Defamation League, "Online Hate and Harassment: The American Experience 2021," 2021, https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021

[15] Pew Center Research, "A majority say social media companies are doing an only fair or poor job addressing online harassment", 2021, https://www.pewresearch.org/internet/pi_2021-01-13_online-harrasment_0-08a/

[16] Ibid

[17] Ibid

[18] FactsFirstPH, https://factsfirst.ph/about

[19] Ekta News Coalition, https://ekta-facts.com/about

[20] Bertie Vidgen et al, Challenges and frontiers in abusive content detection, 2019, https://ora.ox.ac.uk/objects/uuid:3864e746-88c8-4f99-b912-52f4b4be289a/download_file?file_format=application%2Fpdf&safe_filename=W19-3509.pdf&type_of_work=Conference+item

[21] Lucina Di Meco, "Monetizing Misogyny: Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally," She Persisted, 2023, https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf

[22] Kat Lo and Viktorya Vilk, "Shouting into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It," Pen America, 2023, https://pen.org/report/shouting-into-the-void/

[23] Ibid

[24] Ibid

strengthen networked and mob harassment on multiple platforms is particularly relevant for addressing gendered disinformation campaigns targeting women and gender diverse people.[25]

Furthermore, there is limited information available to civil society and governments to understand the nature of online gendered disinformation and evaluate platforms' efficacy in countering it. In Meedan's recent collaboration with National Democratic Institute, we examined the role of gendered disinformation in politics, and determined the importance of coordination efforts in important civic and political events to mitigate the impacts. We recommend that platforms develop a coordination mechanism at the country level, with the involvement of community organizations, in order to better identify, receive and escalate incidents of online gendered disinformation that are likely to have an impact on political discourse or outcomes.[26] Policymakers should also introduce legislation on social media transparency to include specific reporting requirements for online violence against women, nonbinary and trans individuals, including gendered disinformation, in order to make this data available to external researchers, auditing organizations, and the public.[27]

The application of current technical innovations

The integration of verification or fact-checking workflows and processes is a key part of this response effort.[28] Current human-led fact-checking efforts tend to address the largest claims connected to civic, political and cultural moments that are greatest in spread and reach. However, these may not include the instances of gendered disinformation, especially in regions underserved by fact-checking efforts or programs. As disinformation worldwide continues to escalate, it is becoming harder and harder for fact-checking efforts to keep up. The fact-checking field is a time-consuming endeavor, often relying on domain expertise.[29] Technical infrastructure that supports the detection of gendered disinformation by tracking the spread of false narratives within and between languages can support fact-checking systems and stakeholders.[30] A key input into these detection systems is the development of models in languages currently under-resourced in machine learning.[31]

The algorithms developed to support content moderation at scale have yet to benefit from the topic area and lived expertise of communities best positioned to annotate disinformation for relevance, context, and risk.[32] In addition to explicitly gathering communities of practice to generate meaningful datasets that can be used to better flag instances of gendered

---

[25] Ibid

[26] National Democratic Institute, "Interventions to End Online Violence Against Women in Politics", 2022, https://www.ndi.org/publications/interventions-end-online-violence-against-women-politics

[27] Ibid

[28] Kunze, Alex, et al. "AUTHORITARIANISM AND GENDERED DISINFORMATION." (2021).https://she-persisted.org/Authoritarianism_and_Gendered_Disinformation_May_2021.pdf

[29] Preslav Nakov et al.. Automated fact-checking for assisting human fact-checkers. International Joint Conferences on Artificial Intelligence Organzation, 8 2021. doi: 10.24963/ijcai.2021/619. URL https://doi.org/10.24963/ijcai.2021/619. Survey Track.

[30] Kazemi, Ashkan, et al. "Claim matching beyond English to scale global fact-checking." *arXiv preprint arXiv:2106.00853* (2021). URL https://arxiv.org/pdf/2106.00853.pdf

[31] Dori-Hacohen, Shiri, and Scott A. Hale. "Information Ecosystem Threats in Minoritized Communities: Challenges, Open Problems and Research Directions." *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022. URL https://dl.acm.org/doi/abs/10.1145/3477495.3536327

[32] In the early months of the COVID-19 pandemic, Meedan's Digital Health Lab launched an initiative, Health Desk, to ensure that topic area experts support fact-checkers in disinformation response, and that these exchanges can get documented into algorithmic flagging systems.

disinformation, the final component of this intervention pathway must involve the development of data partnerships with vetted research communities. The goal must be to enable open-access, non-commercial research that can inform the development algorithms that individuals and communities can use to counter the rise of hate, harassment, and disinformation in their digital environments. In the long-term, this serves an ecosystem goal of continuing to build open infrastructure between subject matter experts and content annotators, platform trust and safety teams, fact-checkers and journalists in responding to disinformation events.

This requires close collaboration between computer scientists, community organizations, and practitioners trained in gender analyses, working together to develop data standards for documenting gendered disinformation, detecting and flagging instances on social media, and developing more comprehensive response systems. [33]

In conclusion, accountability and technical collaborations with a community driven approach should be at the center of disinformation responses with gender dimensions.

The complexity of addressing online gendered disinformation lies in a web of challenges. These include defining it, its interrelationship with online and offline violence that puts gendered disinformation in a larger pool of violence that may dilute strategies, opaque measures of social media companies to tackle multi-platform spread of disinformation, limited involvement of communities in understanding the cultural and contextual nuances of gendered disinformation.

To address these, there needs to be a concerted effort along with contextual solutions and greater reliance on the expertise and lived experiences of civil society to inform more robust technical systems, more effective strategies and more inclusive policies.

[33] We would like to call attention to Meedan's ongoing project to document gendered disinformation in South Asia along with our community partners - Chambal Media, Digital Rights Foundation and The Quint and supported by the Sexual Violence Research Initiative . As gendered disinformation has not been studied in detail in these contexts, it presents a challenge in identifying specific examples that can be developed into datasets and used for more effective monitoring and disinformation response. We will tackle this by bringing together our local, community knowledge with state-of-the-art machine learning techniques pioneered at Meedan, and developed in collaboration with disinformation researchers and computational social scientists from around the world. The objective of the project is to collaboratively define, identify, document and annotate a high quality dataset of gender based disinformation in online spaces for better understanding and countering of the issue in the region and apply machine learning techniques to further detect gendered disinformation online.