PLACING "GENDER" IN DISINFORMATION

APC
ASSOCIATION FOR PROGRESSIVE COMMUNICATIONS

Placing “gender” in disinformation

Research and text: Paula Martins, with chapter-specific inputs from Amalia Toledo and Anja Kovacs

Review: Erika Smith, Hija Kamran, Karla Velasco and Namita Aavriti Malhotra

This brief was drafted by Paula Martins, APC’s Social and Environmental Justice Programme manager, based on a year-long project carried out in partnership with the UN Special Rapporteur on freedom of opinion and expression, Irene Khan, during which regional consultations and a series of seminars and workshops were organised to discuss freedom of expression and gender justice, with a focus on the different types of censorship, threats and attacks faced by women and gender non-conforming individuals in their interactions with technology. Regional and thematic workshops and the collection of case studies were organised with TEDIC and IFEX (Latin America), SMEX (MENA) and Pollicy (Africa).

Copy editing: Zara Ismail
Proofreading: Lori Nordstrom and Lynn Wellburn
Design and layout: Jaspreet Singh
Cover design: Cathy Chen

For more information on APC’s work on disinformation see: https://www.apc.org/en/project/placing-gendered-disinformation

This project and publication were possible due to the support of FCDO, Wellspring Foundation and Oak Foundation.

Published by the Association for Progressive Communications (APC)
2024

Creative Commons Attribution 4.0 International (CC BY 4.0)
https://creativecommons.org/licenses/by/4.0/

ISBN 978-92-95113-69-5
APC-202404-SEJ-R-EN-DIGITAL-359

TABLE OF CONTENTS

INTRODUCTION: THE PROBLEM

In a recent survey carried out with women journalists by UNESCO and the International Center for Journalists (ICFJ), 73% of the respondents (n=456) who identified as women said they had experienced online violence in the course of their work.[1] The Troll Patrol project, a collaboration between Amnesty International and Element AI, surveyed millions of tweets received by 778 journalists and politicians from the UK and US over the period of one year and found that 7.1% of the tweets sent to these women were "problematic" or "abusive". This amounted to 1.1 million tweets mentioning 778 women across the entire year, or one every 30 seconds.[2] Plan International interviewed 14,000 girls across 31 countries in the study *Free to Be Online?* and concluded:

> Girls are targeted online just because they are young and female, and if they are politically outspoken, disabled, Black or identify as LGBTIQ+, it gets worse. [...] Like street harassment it is unremitting, often psychologically damaging and can lead to actual physical harm.[3]

The evidence of a grave problem is overwhelming. The Association for Progressive Communications (APC) has worked to render visible the impact of technology-facilitated gender-based violence (TFGBV) for more than two decades. We have worked with women's organisations and advocates to identify, monitor, analyse and combat uses of the internet and digital technologies that are harmful to women and marginalised communities, and with individual internet users to assist them in using technology to document and combat TFGBV and challenge harmful sexist online practices. We have also advocated for internet policies and regulations that enable the expression, protection and promotion of human rights, women's rights, and the rights of people of diverse sexualities to both states and private sector actors.

Over the past few years in particular, we have seen how online TFGBV has moved from a peripheral discussion in both the women's rights and internet rights communities to occupying a central space in conversations about a free and open internet.

TFGBV is part of the continuum of offline-online gender-based violence and, as such, occurs in all countries, contexts and settings; it is a pervasive violation of human rights, and is a "manifestation of the historically unequal power relations between women and men and systemic gender-based discrimination."[4] With advances in technology and our changing relationships with it, however, the

[1] Posetti, J., et al. (2021). *The Chilling: Global trends in online violence against women journalists.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377223

[2] Amnesty International. (2018, 18 December). Crowdsourced Twitter study reveals shocking scale of online abuse against women. https://www.amnesty.org/en/latest/press-release/2018/12/crowdsourced-twitter-study-reveals-shocking-scale-of-online-abuse-against-women

[3] Plan International. (2020). *The State of the World's Girls 2020: Free to Be Online?* https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf

[4] Prevention of violence against women and girls: Report of the Secretary-General. Commission on the Status of Women, Fifty-seventh session, 4-15 March 2013. https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=E/CN.6/2013/4&Lang=E

concrete manifestations of TFGBV have also evolved. One of these manifestations that deserves special attention today, in APC's view, is gendered disinformation.

In order to better capture the variety of problems women and gender-diverse people face when expressing their views and opinions, APC co-organised a series of consultations during 2023 to collect further information on their lived experiences, in different cultures and geographies. These in-depth conversations pointed to the fact that gender-based violence, hate speech and disinformation are different challenges that sometimes overlap, and that intersecting areas have been abused and used as an excuse to limit expression that is legitimate or legally protected – including expression on gender-related issues. At the same time, cases that amount to incitement to hatred, discrimination and violence have not been duly investigated and punished.

The relationship between these concepts – online gender-based violence, disinformation and hate speech – is complex and multilayered. One reason for this is that, more than mere theoretical concepts, they reflect the many imbricated manifestations of discrimination, inequality and violence that women and gender-diverse individuals experience day after day. Despite these compounding layers of complexity, however, it is important to recognise gendered disinformation as a specific phenomenon that lies at the intersections of different concepts and that should be made explicit if better responses to address it are to be developed.

There is no one definition of gendered disinformation that is broadly agreed on or commonly accepted. The term is often used interchangeably with TFGBV, or is referred to as the gender dimensions of disinformation. Disinformation itself remains an expression that lacks a single agreed upon definition and is too often conflated with other concepts, such as propaganda and advocacy to incite discrimination, violence and hostility.[5] The UN Special Rapporteur on freedom of expression and opinion has expressed concern with the growing use of manipulation, deception and the distortion of information, aimed at creating confusion.[6] These conceptual challenges testify to the fact that disinformation is a multifaceted phenomenon.

APC considers it important to characterise gendered disinformation because it relates to a specific type of violation of women's and gender-diverse people's rights, in particular their freedom of expression, which is not properly encapsulated by other concepts. By failing to talk about gendered disinformation and trying to clarify its meaning, we may be contributing to the invisibilisation of particular situations of abuse that require specific and targeted responses.

[5] A/77/288, paragraph 12.

[6] See, for example, A/77/288, paragraph 13.

Disinformation can serve various objectives. The ultimate goal of gender-related and identity-based disinformation is to discourage the exercise of freedom of expression by women, gender-diverse individuals and marginalised groups and to manipulate the information ecosystem. Harm to specific individuals is in general a secondary outcome.[7] It is the combination of radical narratives opposed to gender equity and the phenomenon of disinformation that characterises gendered disinformation.[8]

In her 2023 report on this issue, the United Nations Special Rapporteur on freedom of expression clarified that gendered disinformation is *gendered* "because it targets women and gender non-conforming individuals, because of the gendered nature of the attacks and their gendered impact, and, very importantly, because it reinforces prejudices, bias and structural and systemic barriers that stand in the way of gender equality and gender justice."[9]

[7] Jankowicz, N., et al. (2021). *Malign Creativity: How Gender, Sex, and Lies Are Weaponized Against Women Online.* Wilson Center. https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online

[8] InternetLab. (2023, 7 July). InternetLab submits contribution to the UN on gendered disinformation. https://internetlab.org.br/en/news/internetlab-submits-contribution-to-the-un-on-gendered-disinformation

[9] Khan, I. (2023). *Gendered disinformation and its implications for the right to freedom of expression – Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.* A/78/288. https://www.ohchr.org/en/documents/thematic-reports/a78288-gendered-disinformation-and-its-implications-right-freedom.

CHARACTERISING GENDERED DISINFORMATION

There is ample evidence that demonstrates that women[10] and gender non-conforming people are major targets of disinformation campaigns.[11] Not only that, but the impact they suffer is differentiated. The contours of the most common attacks they face tend to follow specific patterns, and maintaining the exclusion of these groups from public debate and cultural spaces has been identified as a common underlying goal.

However, because there are no broadly accepted definitions of either TFGBV or disinformation, and because these challenges tend to take place in conjunction with one another, studies in the area map or refer to a variety of human rights violations and harmful behaviours that relate to the gender dimensions of disinformation, but also relate to "neighbouring" and sometimes overlapping challenges. As a consequence of this common conflation of concepts, and other methodological factors,[12] it is difficult to find comprehensive data on the various aspects of gendered disinformation that is specific and isolated from broader TFGBV. This leads to serious challenges relating to monitoring, documenting and defining the clear boundaries of gendered disinformation.

While these borders are difficult to delineate, the analysis of existing data and cases allows us to identify a number of common elements that point to the most common profiles of targeted individuals, perpetrators, as well as recurrent vectors, narratives and resulting harms. These elements will be briefly presented below. To do so, this report will refer to literature on disinformation, online abuse, harassment directed at women generally, abuse directed at women active in public life, state-backed operations with gendered dimensions, and disinformation that employs gendered or sexual language. The data and cases collected from scholarship have all been contrasted and compared with the testimonies directly collected by APC from women human rights defenders, women politicians and women journalists in regional consultations.

[10] For example, using data analytics from the non-partisan firm Marvelous AI, a 2019 study concluded that accounts considered low in credibility, including bots and trolls, attacked female candidates in the U.S. Democratic presidential primary at higher rates than their male counterparts. See: Di Meco, L., & Wilfore, K. (2021, 8 March). Gendered disinformation is a national security problem. *Brookings*. https://www.brookings.edu/techstream/gendered-disinformation-is-a-national-security-problem

[11] See, for example: Dunn, S., Vaillancourt, T., & Brittain, H. (2023). *Supporting Safer Digital Spaces*. Centre for International Governance Innovation. https://www.cigionline.org/publications/supporting-safer-digital-spaces

[12] The absence of agreed-upon definitions of online gender-based violence and methodologies for measurement coupled with widespread underreporting make it a challenge to understand the true extent of the problem globally, as well as to identify regional variations.

2.1 Intersectionality: The identity of the survivor is key

The more outspoken and visible, the more they are attacked. The main targets of disinformation are women engaging politically or culturally – women in politics, journalists and human rights defenders. This does not mean, however, that a public role is a defining trait of disinformation survivors, since women of any background strongly expressing their views can be subjected to disinformation,[13] especially if these views refer to issues relating to women's rights, sexual and reproductive issues, the violence against LGBTQI+ persons and other gender-related themes.

There has actually been disproportionate attention paid to high-impact cases involving public female figures,[14] and there is, therefore, a need to broaden this focus to learn more about attacks on trans and non-binary persons and women rights defenders.

Existing evidence already confirms that the identity of an individual plays an important role in why and how they are targeted. Intersectionality is a key aspect of gendered disinformation. Recent studies demonstrate the widespread nature of online harms and the greater negative impact of such harms on women and LGBTQI+ people. The data available indicates that the subgroups of women that are at a heightened risk of offline GBV are also more at risk of facing online risks.[15]

As highlighted by different studies, disinformation exploits existing social divides and tension points, targeting groups already in a situation of marginalisation.[16] The compounding challenges imposed by their identities heightens both their vulnerability to attacks and the harms described below.

A study by Amnesty UK, for example, found that gendered disinformation affects Black, Asian and minority ethnic (BAME) women members of parliament (MPs) far more than their white colleagues. The 20 BAME MPs in this study received almost half (41%) of the abusive tweets recorded, despite there being almost eight times as many white MPs in the study.[17] The 2017 Troll Patrol project by Amnesty International and Element AI, which studied journalists and politicians from the UK and US, found that women of colour (Black, Asian, Latinx and mixed-race women) were 34% more likely to be mentioned in abusive or problematic tweets than white women; and Black women, in particular, were disproportionately targeted, being 84% more likely than white women to be mentioned in abusive or problematic tweets.[18]

[13] See, for example, the case of Safoora Zargar in India, at: Lalwani, V. (2021, 8 March). 'I kept feeling it was a nightmare': Safoora Zargar on surviving 38 days in solitary confinement. *Scroll.in*. https://scroll.in/article/988844/i-kept-feeling-it-was-a-nightmare-safoora-zargar-on-surviving-38-days-in-solitary-confinement

[14] https://counteringdisinformation.org/topics/gender/3-current-approaches-countering-gendered-disinformation-and-addressing-gender

[15] APC. (2023, 1 September). Overview of the manifestations and impacts of technology-facilitated gender-based violence and the need for safety by design. https://www.apc.org/en/pubs/overview-manifestations-and-impacts-technology-facilitated-gender-based-violence-and-need

[16] Polletta, F., & Callahan, J. (2017). Deep Stories, Nostalgia Narratives, and Fake News: Storytelling in the Trump Era. *American Journal of Cultural Sociology*, 5, 392-408. https://doi.org/10.1057/s41290-017-0037-7

[17] https://www.amnesty.org.uk/online-violence-women-mps

[18] Amnesty International. (2018, 18 December). Op. cit.

Well-known cases of women running for elections, including to high posts such as vice president candidates, confirm those numbers. This was the case, for example, with Francia Márquez in Colombia in 2022,[19] and Kamala Harris in the US.[20] In 2020, Harris faced a number of coordinated attacks, including the promotion of the narrative that she was constitutionally ineligible to be vice president because she was the daughter of immigrants (despite Harris being born in the United States) and many questioning her identity as a Black woman.[21]

In India, in 2022, a disinformation campaign was organised against a Muslim girl after a video showing her standing up to a mob of saffron-clad boys chanting “Jai Shri Ram” went viral[22] amidst protests concerning the banning of the use of the burqa in some schools.[23]

A 2019 study by the National Democratic Institute in Indonesia, Colombia and Kenya confirmed the importance of paying attention to minority communities and intersecting identities. In Colombia, for example:

> [F]emale representatives from the deaf community shared that the violence they faced was not in text, but through the uploading of violent GIFs and/or video clips in sign-language. It was explained that this delivery mechanism was particularly effective in conveying threat and insecurity because, for the majority of the members of the deaf community in Colombia, sign language is their first language, and the targeting was therefore unmistakable. Understanding that the kinds of threats and modes of online violence can differ substantially when targeting different marginalized communities indicates that further work is required to create relevant lexicons.[24]

[19] El Espectador. (2022, 9 Mayo). Candidatas 2022: víctimas de violencia machista, según observatorio de violencias. *El Espectador*. https://www.elespectador.com/politica/elecciones-colombia-2022/elecciones-2022-candidatas-2022-victimas-de-violencia-machista-segun-observatorio-de-violencias/

[20] Tumulty, K., Woodsome, K., & Peçanha, S. (2020, 7 October). How sexist, racist attacks on Kamala Harris have spread online – a case study. Washington Post. https://www.washingtonpost.com/opinions/2020/10/07/kamala-harris-sexist-racist-attacks-spread-online

[21] Ibid.

[22] Khadka, S. (2022, 20 May). Gendered Disinformation On Social Media: The Future India Foundation’s Report Calls For Accountability. *Feminism in India*. https://feminisminindia.com/2022/05/20/gendered-disnformation-on-social-media-the-future-india-foundation-report/

[23] The Quint. (2022, 9 February). Burqa-Clad Student Confronts Saffron-Clad Mob in K’taka, Earns Owaisi’s Praise. *The Quint*. https://www.thequint.com/news/education/burqa-clad-student-confronts-saffron-scarved-mob-in-karnataka-amid-hijab-row

[24] National Democratic Institute. (2019). Tweets That Chill: *Analyzing Online Violence Against Women in Politics*. https://www.ndi.org/sites/default/files/NDI%20Tweets%20That%20Chill%20Report.pdf

2.2 Content

With regard to its content, gendered disinformation relates to messages that represent direct attacks on women or gender-diverse individuals, generally based on gender bias, stereotypes and expectations. These individuals are targeted for occupying domains normally considered to be male spaces, for speaking out against gendered and other kinds of inequality, or for behaving in ways considered to be in contravention of "proper" and dominant cultural, religious or moral standards.

A study of these narratives has also revealed a trend of perpetrators seeking to question whether the targets of said violence are legitimate voices worthy of attention. Most perpetrators do not concern themselves with questioning the logic of their targets' opinions, but rather the individual's "standing in society" and whether they are "capable" of expressing valuable views and ideas. Too often, this violence is sexualised in nature or seeks to impact its targets' family life.

An important aspect of the content of gendered disinformation campaigns is that it not only seeks to set upon individuals, but also ideologies and struggles. Feminism, gender rights and LGBTQI+ awareness, for example, are among the gender-related concepts and terminologies particularly subject to coordinated debunking and delegitimisation attempts. As seen in the attacks on individuals, these ideas are considered "less", "crude", "baseless", "corruptive" or, in a nutshell, as unworthy as those that disseminate them through words or actions.

The content used in disinformation campaigns may be produced by the perpetrators or may involve a devious reusing of content produced by third parties. This can include content that is not false, but that matches the false narrative being promoted.[25]

Sobieraj proposes three main overlapping strategies used by those who seek to silence women or limit their impact in the digital publics: intimidation, shaming and discrediting.[26] "Women are often reduced to their roles as mothers, daughters and caregivers rather than seen as legitimate political and economic actors," she notes. "They are labeled 'bad mothers', 'difficult', 'loose', 'loud', 'nasty' or 'witches'. They are cast as 'unbelievers', 'atheists', 'guerrillas', 'separatists', 'the enemy within', 'traitors', 'anti-nationalists' or 'terrorists'."[27]

[25] Hindman, M., & Barash, V. (2018). *Disinformation*, 'Fake News' and Influence Campaigns on Twitter. Knight Foundation. https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter

[26] Sobieraj, S. (2018). Bitch, slut, skank, cunt: patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, 21(11), 1700-1714. https://doi.org/10.1080/1369118X.2017.1348535

[27] Ibid.

2.3 Malicious actors and their drivers

Disinformation can serve various objectives. Generally speaking, disinformation is better understood when seen as a symptom of a broader information disorder, mainly fuelled by social media. Malicious actors benefit from this disorder for purposes that are varied and sometimes interrelated. Most campaigns are fluid and can be found at the crossroads of political, lucrative, and issue-based disinformation.[28] As for gender-related and identity-based disinformation, its ultimate goal is to discourage the exercise of freedom of expression and manipulate the information ecosystem against specific groups. Harm to individuals is in general a secondary outcome.[29]

Actors behind gendered disinformation are often motivated by ideology or the intention to undermine social cohesion. In extreme cases, gendered disinformation campaigns may seek to incite violence. The goal is not only to threaten democracy itself, or impact electoral results, but to create mistrust in information, particularly gender-related information. Financial and political gains are common, but ideological victories are, at times, the longer-term goals.

Moving beyond the isolated actions of misogynistic and anti-LGBTQI+ individuals, gendered disinformation campaigns often indicate coordination and, in some cases, centralisation and funding. Coordinated disinformation campaigns are repeatedly organised by extremist groups and state-aligned collectives.[30] Both national and foreign agents have reportedly been active in promoting gender-related disinformation.[31]

An extensive investigative piece by the fact-checking group Chequeado, for example, reported on the network of anti-rights groups promoting disinformation on gender issues in the Americas. According to them, the network provides coordination in areas such as messaging, the channelling of funds, strategic alliances between organisations, training scholarships and international events. Misinformation can range from listing the false adverse effects of abortions to population control conspiracy theories, and follows similar strategies from country to country.[32]

[28] EU DisinfoLab. (2020). The Few Faces of Disinformation. https://www.disinfo.eu/wp-content/uploads/2020/05/20200512_The-Few-Faces-of-Disinformation.pdf

[29] Jankowicz, N., et al. (2021). Op. cit.

[30] Krasodomski-Jones, A., et al. (2019). *Warring Songs: Information Operations in the Digital Age*. Demos. https://demos.co.uk/wp-content/uploads/2019/10/Warring-Songs-final-1.pdf

[31] See, for example: Krasodomski-Jones, A., et al. (2020). *Engendering Hate: The contours of state-aligned gendered disinformation online*. Demos. https://demos.co.uk/research/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online and Di Meco, L., & Wilfore, K. (2021, 8 March). Op. cit.

[32] Sohr, O. (2023, 29 June). Desinformación de género: cómo se articulan los grupos que difunden falsedades sobre el tema en América Latina. *Chequeado*. https://chequeado.com/nota/desinformacion-de-genero-como-se-articulan-los-grupos-que-difunden-falsedades-sobre-el-tema-en-america-latina

Gendered disinformation campaigns have been triggered by state leaders and high-level politicians, and then escalated by the ranks of their followers. In some cases, these actors utilise their media assets, control of the information environment and state-backed troll farms to spread disinformation about women politicians, policy makers, journalists and activists, and even policies targeted at women.

This type of coordinated public shaming is an effective tactic because it alienates women, often turning family members, colleagues and neighbours against them. In cases where families and communities are their primary source of protection, this tactic can leave women defenders vulnerable to physical attacks and psychological harm. In the Middle East and Northern Africa (MENA) region, participants in a regional consultation organised by APC highlighted the aforementioned point, also speaking about the role male family members played in relation to disinformation campaigns and related violence.[33]

At different levels, the aggressors feed off each other and form an "ecosystem", each with different roles. While individual posts may not seem too problematic in isolation, when coordination and amplification take place, a "virtual mob" launches an operation that, over time, can lead to serious threats and aggression.

2.4 Formats and vectors

Gendered disinformation campaigns are often characterised by the deployment of coordination and malign intention. These campaigns include false or harmful content that exploits gender inequalities or weaponises gender stereotypes.

Gendered disinformation campaigns are, as a rule, launched on various mediums, both online and offline, at different times, and rely on a combination of human engagement and automation.[34] According to a recent survey carried out by the Centre for International Governance Innovation (CIGI) – and supported by APC – in 18 countries, among participants who had experienced at least one type of online harm, 71% identified social media as the platform where it occurred.[35]

[33] For example, Qandeel Baloch was a Pakistani social media celebrity who was constantly trolled on the internet and then killed in 2016 by her own brother in an "honour killing" after her real identity was revealed. See, for more information: Boone, J. (2017, 22 September). 'She feared no one': the life and death of Qandeel Baloch. *The Guardian*. https://www.theguardian.com/world/2017/sep/22/qandeel-baloch-feared-no-one-life-and-death

[34] Plan International. (2021). *The State of the World's Girls 2021: The Truth Gap*. https://plan-international.org/uploads/2022/02/sotwgr2021-commsreport-en.pdf

[35] APC. (2023, 1 September). Op. cit.

Studies have identified Facebook[36] and X (formerly Twitter)[37] as the top locations for incidents of online gender-based violence. However, there is also evidence of the use of search engines[38] and messaging services.[39] Perpetrators have developed techniques of "attention hacking" through the strategic use of memes, graphic comments, explicit images, manipulated texts or contexts and manipulated images.[40] Deceptive behaviour also often includes manufacturing virality (using bots, cyborgs or fake accounts) to artificially increase the reach and popularity of certain content for a greater perceived impact. Audiences can be misled by mimicking organic engagement or masking the sponsors of messages ("astroturfing") and giving the impression of spontaneous action or support by grassroots participants.[41] Bots are commonly used and then journalists, bloggers and influencers are targeted to help spread content.[42]

Deepfake technology is predominately being used to create sexual videos of women without their consent.[43] In 2019, for example, DeepTrace Labs identified 14,678 deepfake videos across a number of streaming platforms and pornographic sites. Of these, 96% portrayed women.[44] A 2020 report by Sensity AI "found that 96 percent of deepfakes were non-consensual sexual deepfakes, and of those, 99 percent were made of women."[45] Well-known cases include Indian journalist Rana Ayyub[46] and activist Noelle Martin.[47]

[36] A survey conducted by Pollicy found that the majority (71.2%) of all the incidents of online gender-based violence against respondents in Africa occurred on Facebook. In Kenya, Uganda, Senegal and South Africa, this violence happens primarily on Facebook and WhatsApp. In Ethiopia, Facebook and additionally Telegram were the main platforms where women experienced online violence. See: Iyer, N., Nyamwire, B., & Nabulega, S. (2020). *Alternate Realities, Alternate Internets: Feminist Research for a Feminist Internet*. Pollicy. https://www.apc.org/sites/default/files/Report_FINAL.pdf; see also: Hicks, J. (2021). *Global evidence on the prevalence and impact of online gender-based violence*. Institute of Development Studies. https://www.ids.ac.uk/publications/global-evidence-on-the-prevalence-and-impact-of-online-gender-based-violence-ogbv

[37] Jankowicz, N., et al. (2021). Op. cit.

[38] Taylor, E., et al. (2020). Follow the Money: *How the Online Advertising Ecosystem Funds COVID-19 Junk News and Disinformation*. Oxford Internet Institute. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2020/08/Follow-the-Money-3-Aug.pdf

[39] Sessa, M. G., Willaert, T., & Van Soest, J. (2022). *The disinformative ecosystem: Link sharing practices on Telegram as evidence of cross-platform amplification*. VUB & EU DisinfoLab. https://belux-edmo.s3-accelerate.amazonaws.com/wp-content/uploads/2023/03/EDMOBELUX_The-disinformative-ecosystem-FINAL-updated.pdf

[40] Collins-Dexter, B. (2020, 16 October). Butterfly Attack: Operation Blaxit. *Media Manipulation Casebook*. https://mediamanipulation.org/case-studies/butterfly-attack-operation-blaxit

[41] Haas, J. (2022). *A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation*. University of Innsbruck. https://diglib.uibk.ac.at/ulbtiroloa/content/titleinfo/7710207/full.pdf

[42] Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society. https://perma.cc/AHD8-SPXQ

[43] See, for example: Romano, A. (2019, 7 October). Deepfakes are a real political threat. For now, though, they're mainly used to degrade women. *Vox*. https://www.vox.com/2019/10/7/20902215/deepfakes-usage-youtube-2019-deeptrace-research-report

[44] Ibid.

[45] Dunn, S. (2021, 3 March). Women, Not Politicians, Are Targeted Most Often by Deepfake Videos. *Centre for International Governance Innovation*. https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos

[46] Ayyub, R. (2018, 21 November). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *Huffington Post*. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

[47] https://www.ted.com/speakers/noelle_martin

Disinformation has an intricately entwined relationship with online partisan media, both responding to and setting its issue agenda.[48] The ecosystem formed by perpetrators often makes use of a network of proxies, fake accounts, carve-outs and cut-outs to make identification hard, "dissimulating the geographical area from where they operate, for ideological reasons or financial motivation."[49] Groups of pages or people work together to mislead others about who they are or what they are doing.

Schafer and Meleshevich argue:

> Just as ill-gotten money needs to be moved from an illegitimate source into an established financial institution, disinformation is most powerful when a façade of legitimacy is created through "information laundering".[50]

Misogynistic content moves across services and platforms, including legacy media, to gain legitimacy. This means that radicalised speech may move from the fringes to more popular and mainstream spaces – this is an aspect that requires further research, but clearly indicates that an analysis of the vectors of disinformation should adopt an ecosystem approach.

2.5 Impact

Gendered disinformation reinforces harmful patriarchal and heteronormative institutional and cultural structures.

Gendered disinformation and other forms of online gender-based violence may push women and gender non-conforming people away from public and cultural spaces, reducing the diversity of voices and worldviews in such spaces.[51] Gendered disinformation, like other forms of disinformation, undermines democracy and good governance, increases political polarisation, and expands social cleavages.

A 2019 study in Finland "found that 28 percent of municipal officials targeted with hate speech said they were less willing to participate in decision-making as a result."[52]

[48] Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society, 20*(5), 2028-2049. https://doi.org/10.1177/1461444817712086

[49] Grégoire, A. (2021, 8 April). The Coordination Assessment. *EU Disinfo Lab*. https://www.disinfo.eu/publications/cib-detection-tree1

[50] Schafer, B., & Meleshevich, K. (2018, 9 January). Online Information Laundering: The Role of Social Media. *German Marshall Fund*. https://www.gmfus.org/news/online-information-laundering-role-social-media

[51] See, for example: Perraudin, F., & Murphy, S. (2019, 31 October). Alarm over number of female MPs stepping down after abuse. *The Guardian*. https://www.theguardian.com/politics/2019/oct/31/alarm-over-number-female-mps-stepping-down-after-abuse

[52] Cater, L. (2021, 17 March). Finland's women-led government targeted by online harassment. *POLITICO*. https://www.politico.eu/article/sanna-marin-finland-online-harassment-women-government-targeted

Growing evidence shows that online gender-based violence facilitates offline violence and creates "climates of unsafety" within society.[53]

At the individual level, an observation of the cases of online gender-based violence point to the conclusion that perpetrators intended to harm women and LGBTQI+ individuals socially, psychologically, economically and physically.[54] It was found that they sought to:

- Endanger the integrity of the information ecosystem.
- Put women targeted at risk of (further) violence (including physical).
- Promote or reinforce gender inequality and discrimination, deepening social cleavages.

And that such violence:

- Has a chilling effect on women's willingness to participate in public spaces and on their many types of self-expression.
- May result in negative economic consequences for targeted subjects.
- Has a compounded effect, disproportionately affecting women (they are more targeted and subject to "more" effects).

2.6 Global reach

Gendered disinformation is a global phenomenon. However, few country- or region-specific studies are currently available and there is limited systematic documentation of gendered disinformation in the global South.

Women and gender-diverse individuals in the global South face specific circumstances and challenges related to gendered disinformation and any responses to it should be attentive to context, including language and cultural specificities. Countries in the global South may have fewer resources to both monitor and address technology-facilitated violence and face other challenges concerning rule of law and the local representation of technology companies, as will be addressed below in the chapter on responses. More research and attention to the realities of women, girls and gender non-conforming individuals in the global South is urgently needed.

This has been confirmed, for example, by APC's members, such as Ugandan NGO Pollicy, which states:

[53] Hicks, J. (2021). Op. cit.

[54] Testimonies collected during the regional consultations.

> There is a major gap in data on the prevalence of all types of online violence against women and girls in low and middle-income countries. Furthermore, where this evidence is available, the data is not gender-disaggregated or does not take into account the intersectional impact on class, women with disabilities, refugee situations or traditionally marginalized areas. [55]

2.7 Context: Locating toxic speech

Gendered disinformation flourishes in societies where women's freedom of expression is constrained. Understanding the context (and language) is key to understanding the messaging. Although gendered disinformation is a global issue, there is a need for it to be understood and addressed from regional and hyperlocal perspectives.

In the regional consultations carried out by APC, the common messaging in the background was mostly the same, reverberating the narratives already discussed in the section titled "Content" above. In each region, however, these messages were dressed up to fit regional narratives built on top of salient cultural and religious beliefs.[56]

In all consultations the role of religious narratives was highlighted by participants. Different dynamics between majority religious groups and minority ones are exploited in disinformation campaigns, evidencing power relations that move from religious identity to broader identity politics.

The regional consultations served to demonstrate that, although they present common elements, disinformation campaigns are strongly contextual. In South Asia, much disinformation lies at the intersections of gender, religion and caste. In Africa, the narrative of family protection was reportedly strong, and women politicians, journalists and human rights defenders were portrayed as attacking family values with their ideas or simply by virtue of their professions. The anti-colonialism narrative was also indicated as a trend, with many actors distorting anti-colonialism concerns to weaponise them against women activists and gender-diverse people, accusing them and their ideas of being a form of the undue influence of "Western values". In Latin America, women and gender non-conforming people were often targeted in the midst of narratives preaching the dangers of "gender ideology". In Eastern Europe, LGBTQI+ groups highlighted the links between geopolitics and gendered disinformation. For example, this region saw campaigns that centred the risks posed by "Western values", which were portrayed as being backed by Western states and as a threat to the nation and national security. Activists were accused of being traitors and spies.

[55] Iyer, N., Nyamwire, B., & Nabulega, S. (2020). Op. cit.

[56] Sess\a, M. G. (2022, 26 January). What is Gendered Disinformation? *Heinrich Böll Stiftung*. https://il.boell.org/en/2022/01/26/what-gendered-disinformation

LOOKING AT DISINFORMATION FROM A GENDER LENS: KEY LEARNINGS

Disinformation is a challenge related to freedom of speech that places information and its manipulation at the centre of any analysis. TFGBV is a broad phenomenon that, in many cases, involves harmful behaviour that is not related to speech and expression. There is, however, a space where they intersect and where we can locate gendered disinformation.

During 2023, APC carried out a series of activities that sought to characterise gendered disinformation. The results were presented above, and in the sections below we highlight key learnings arising from the analysis of the information, data and cases collected.

First, we will explore how gendered disinformation is a symptom of information ecosystems marked by gender inequality. We will then look at how misleading and confrontational narratives, more than falsity, tend to characterise gendered disinformation campaigns, along with coordinated behaviour and emotional content. The intrinsic relation between gendered disinformation and violence will be assessed. The final sections will review the contextual factors that facilitate the flourishing of disinformation narratives and the complicated characterisation of gender activists as simultaneous targets and agents of disinformation.

3.1 Gendered disinformation as a symptom of information ecosystems marked by gender inequality

Disinformation tends to thrive where human rights are constrained, where the public information regime is not robust, and where diverse, verifiable sources of information are lacking.[57] As demonstrated in the 2021 report by the UN Special Rapporteur on freedom of opinion and expression on the issue of gender justice, this is exactly the freedom of expression context when assessed through a gender justice lens: women are censored in different ways; they pay a disproportionate cost for speaking out; public morals are weaponised against them; their access to information and participation is restricted (in particular as they relate to sexual and reproductive issues); and platforms have failed to respond adequately to the risks and dangers that women confront in digital spaces.[58]

Academics have characterised disinformation as an "information disorder". This is an interesting proposition, for it allows the application of a holistic and interconnected approach to the problem, encouraging a multidimensional, varied and contextualised interpretation that positions disinformation as an aspect of information systems analysis. Assessing the health of information systems allows

[57] A/HRC/47/25, paragraph 4.

[58] A/76/258

us to verify aspects of a given system that may be conducive to the "success" of disinformation narratives, such as a lack of information, a lack of media freedom, restrictive civic space, structural inequality and discrimination, among others.

3.2 Falsity versus misleading narratives

The Global Disinformation Index (GDI) views disinformation through the lens of adversarial narrative conflict:

> Anywhere someone intentionally peddles a misleading narrative, often implicit and constructed using a mix of cherry picked elements of fact combined with fabrications, that is adversarial in nature against an at-risk group or institution, and most importantly, creates a risk of harm, they are engaging in disinformation.[59]

This definition shifts away from a focus on the falsity of specific posts or statements to concentrate on the overarching narrative. This conceptualisation of disinformation is particularly interesting for analysing identity-based disinformation campaigns. First, because as seen below, a lot of gendered disinformation content does not relate to lies or false facts, but to heated opinions and emotional content aimed at inflaming people and exploiting biases and prejudices, thus manipulating facts to create a broader narrative of hatred and opprobrium. Second, and as further clarified by the GDI:

> [I]t illustrates the role that algorithmic recommender systems play in exacerbating the problem, since adversarial narratives exploit our human tendency toward negative content and thus disproportionately drive engagement on algorithmically-driven platforms. That engagement results in more ad sales.[60]

Algorithmic news feeds craft automatically generated, highly personalised adversarial content streams that keep users engaged, on platform, and monetised, and in the end corrupt the entire global information ecosystem.[61]

[59] Rogers, D. (2022, 22 June). Disinformation as Adversarial Narrative Conflict. *Global Disinformation Index*. https://www.disinformationindex.org/blog/2022-06-22-disinformation-as-adversarial-narrative-conflict

[60] Ibid.

[61] Ibid.

3.3 Coordination

Another important aspect of disinformation is the issue of coordination. The risk of harm arising from disinformation, particularly when online, results mainly from the power of amplification that is a fallout of that coordination. As discussed above, perpetrators feed off each other, forming a complex ecosystem. While individual posts may not seem too problematic in isolation, when coordination and amplification take place, a "virtual mob" launches an operation that, over time, can lead to serious threats.[62] As mentioned earlier, disinformation campaigns are, as a rule, launched on various mediums, both online and offline, at different times, and rely on a combination of human engagement and automation. Coordination – and "coordinated inauthentic behavior" in particular – is becoming an increasingly important proxy indicator of disinformation campaigns.[63]

3.4 Emotional content

Studies have shown that many disinformation campaigns pertaining to gender are based on emotional content that does not actually refer to facts. Or when they refer to facts, they manipulate context, dates or other elements of the narrative, making it difficult to clearly or easily spot "lies" (and differentiate disinformation from malinformation or misinformation). The main objective of gendered disinformation is to make people angry, and confuse and influence their perceptions and opinions about women and gender-diverse individuals and their role in society, or reinforce or even validate patriarchal perceptions. "Fake news" bills tend to fail to address the main risk of disinformation campaigns, which is not simply to make people believe in untrue facts, but to create doubt, suspicion and fear.

The subtlety of why people believe in what they do, and how these beliefs are targeted and manipulated, can hardly be addressed through restrictive regulatory frameworks aimed at content. Other holistic responses are needed, including those that address the coordinated amplification of this type of speech, and how monetisation plays a key role in this.

[62] Beck, I., Alcaraz, F., & Rodríguez, P. (2022). *Violencia de género en línea hacia mujeres con voz pública. Impacto en la libertad de expresión*. Alianza Regional por la Libre Expresión e Información & ONU Mujeres. https://lac.unwomen.org/sites/default/files/2023-03/Informe_ViolenciaEnLinea-16Mar23.pdf

[63] Jankowicz, N., et al. (2021). Op. cit.

3.5 The intrinsic relation to TFGBV

Disinformation is a phenomenon difficult to isolate. The use of violence as an element of disinformation campaigns has been identified in relation to disinformation broadly considered. Donovan, Dreyfuss, Lim and Friedberg, for example, affirm that based on their research and domain expertise, disinformation violates the right to freedom of expression and the right to information and truth in the following ways:

- It makes it harder to access timely, relevant, and accurate information.
- It takes advantage of algorithmic amplification to intentionally mislead.
- It silences its target victims through harassment, incitement of fear, and by crowding out their words, opinions and other forms of expression.[64]

APC is convinced that in relation to gendered disinformation, however, the correlation between disinformation campaigns and the use of violence is much more present and marked.

As stated by the Center for Democracy and Technology (CDT):

> Disinformation flows from the same heteronormative patriarchal context in which people experience online GBV, and in some cases there may be an overlap between gendered disinformation and online GBV. One way to think of the difference between the two is that gendered disinformation involves intentionally spreading false information about persons or groups based on their gender identity, and online GBV involves *targeting* and abusing individuals based on their gender identity.[65]

Disinformation can be regarded as a strategy that can serve various objectives. The ultimate goal of gender-related and identity-based disinformation is to discourage the exercise of freedom of expression and manipulate the information ecosystem. Harm to individuals is in general a secondary outcome.

That said, violence or the threat of violence of a gendered nature is often applied as part of gendered disinformation campaigns. So although different concepts, they as a rule take place concomitantly and with the goal of reinforcing each other.

As a result, gendered disinformation and technology-facilitated gender-based violence have often been lumped together in research, analysis and responses, with the result that inadequate attention has been paid to the distinct characteristics and specific impacts of gendered disinformation.

[64] Donovan, J., Dreyfuss, E., Lim, G., & Friedberg, B. (2022). *Submission to the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Technology and Social Change Project. https://mediamanipulation.org/research/submission-un-special-rapporteur-promotion-and-protection-right-freedom-opinion-and

[65] Thakur, D., & Hankerson, D. L. (2021). *Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender*. Center for Democracy & Technology. 2021-02-10-CDT-Research-Report-on-Disinfo-Race-and-Gender-FINAL.pdf

3.6 The many factors that open the way to disinformation of a gendered nature

The profound and recurrent attacks on rights that women and gender non-conforming people face results from a blatant failure of states to properly promote gender justice. "Gender justice" conveys the need for transformative changes encompassing equity (an equal distribution of resources, access and opportunity) and equality (equal outcomes) to break the structural and systemic barriers holding women back.

Gendered disinformation tends to flourish in authoritative countries and societies where women and gender-diverse people's rights are not seriously upheld.[66] In all societies, but in these in particular, power imbalances feed sexism, gender-based discrimination and misogyny, and constrain women's enjoyment of their freedom of expression. Their voices are suppressed, controlled or punished explicitly by laws, policies and discriminatory practices and implicitly by social attitudes, cultural norms and patriarchal values. Growing trends of populism, authoritarianism, nationalism and fundamentalism worldwide have accentuated patriarchy and misogyny and enhanced discrimination against women, as well as the suppression of their ability to express themselves.

These power imbalances manifest through a number of situations, which include women's limited access to information, limited access to meaningful connectivity, stereotyped portrayals by the media, gender data gaps and the weaponisation of public morals, among others. For example:

- An important part of states' omissions in relation to gendered disinformation is their failure to produce qualitative information and data on gender-related themes, in particular sexual rights, disaggregated data relating to socio-economic indicators, and violence against women and LGBTQI+ people. The information gaps and data deficits created by this omission generate a vacuum where disinformation can thrive.[67]
- National laws and judicial decisions often cite the protection of public morals as a reason to criminalise or seek the removal of content deemed to be improper, indecent, obscene or immodest. In a number of countries, such laws have been used to police the online social behaviour of women and remove content relating to sexual expression, sexual orientation or gender identity.
- Many countries criminalise not only homosexuality and "transgender behaviour" but also LGBTQI+ information on grounds of morals, traditional values and child

[66] See, for example: Bardall, G. (2019, 30 October). Autocrats use feminism to undermine democracy. *Institute for Research on Public Policy*. https://policyoptions.irpp.org/magazines/october-2019/autocrats-use-feminism-to-undermine-democracy

[67] See, for example: Golebiewski, M., & Boyd, D. (2018). *Data Voids: Where Missing Data Can Easily Be Exploited. Data & Society*. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3-1.pdf

protection. Evidence shows that such an approach fosters intolerance, stigmatisation and violence, and deprives people of access to accurate information.

- Legal solutions to gender-based violence often adopt a protectionist and punitive approach to violence too. Instead of focusing on its root causes, these norms isolate accountability, targeting specific perpetrators despite the fact that responsibility for gender-based violence is diffuse, multi layered and systemic, as pointed out above.
- Protectionism views women, girls and other marginalised individuals as inherently vulnerable and in need of state and patriarchal protection. However, these policies often sacrifice autonomy and freedom as a result. Many states wash their hands of their broader obligations concerning the promotion of gender equality and non-discrimination, and fail to provide women and gender-diverse people with holistic policies, access to justice and access to public services, while focusing solely on punishing individual perpetrators.

All these profound links between (a lack of) gender justice and its impact on women's enjoyment of the freedom of expression and information cannot be set aside when assessing why gendered disinformation has the impact it has and the structural responses it requires.

3.7 Targets and agents of disinformation

Partly as a result of some of the issues addressed in the preceding section, gender activists and those outspoken in relation to gender issues have been major targets of disinformation, while at the same time being accused of being major agents responsible for the spreading of disinformation.

As a result of these perverse dynamics, some poorly calibrated responses to disinformation end up hurting the same groups they allegedly seek to protect, as will be discussed below.

TECHNOLOGY AND GENDERED DISINFORMATION: BIG TECH'S ROLE

As mentioned earlier, according to a recent survey carried out by CIGI and supported by APC in 18 countries, among participants who had experienced at least one type of online harm, 71% identified social media as the platform where it occurred.

Companies have a responsibility to respect human rights, including the right to gender equality and to freedom of opinion and expression. In line with the Guiding Principles on Business and Human Rights' "Protect, Respect and Remedy" Framework, they are expected to exercise due diligence and conduct regular human rights assessments of their products, operations and policies with a view to identifying, preventing or mitigating actual or potential adverse impacts on human rights and providing remediation. Some jurisdictions have adopted national regulations imposing similar obligations on companies.

Despite such legal standards, online and platform content that amounts to gendered disinformation and may cause harm is widespread and requires much further attention and action from companies. As has been confirmed by leaked internal reports and communications and ratified by employees' accounts, companies have been prioritising profit and engagement over women and gender-diverse people's safety. Their algorithms amplify harmful content and narratives, and facilitate their virality through recommender systems that are built to maximise attention and features that speed up widespread distribution.

Some key concerns in relation to platforms' responses to gendered disinformation include:

- Common content moderation practices relating to online harassment and hate speech observed by most larger platforms present some important limitations concerning their focus and theoretical application to gendered disinformation. Additionally, they also present serious shortcomings in terms of practical implementation.
- Tentative solutions developed to specifically address disinformation are gender-blind.
- Issues pertaining to the business model behind a platform's operation, including so-called "attention economics" and automated advertising systems, have the impact of amplifying algorithmically problematic content, including gendered disinformation, and should be urgently addressed.
- Companies' business models, based on the expropriation of personal data, also fragilise privacy and personal data protection, rendering women and gender-diverse individuals more vulnerable to data breaches and other attacks on their privacy.
- The lack of transparency and data concerning the operation of platforms, in particular social media, runs against efforts to expand research, knowledge and understanding of the measures, practices and policies that allow for the proliferation of gendered disinformation online.

4.1 Content moderation: Limitations and challenges

Given the common overlap between gendered disinformation and TFGBV, gendered disinformation cases are often addressed in content moderation through the application of policies that were originally designed for different manifestations of technology-facilitated violence. Facebook, Twitter, YouTube and many other major companies ban hate speech, harassment, the promotion of violence, and abuse. Most platforms remove offensive content, and remove users who repeatedly violate their terms of service or community guidelines. However, each company has a unique, platform-specific user code of conduct.[68]

To identify abusive content, social media companies use a combination of proactive detection via automation and human moderation and reactive detection via user reporting, which is then adjudicated by automated systems or human moderators.[69] Researchers have called attention to the fact that decision making, in practice, relies on internal documents and not on publicly available policies.[70]

Although human moderators are in theory better equipped to take the nuances of language into account, as well as the cultural and socio-political context, they normally work in poor labour conditions[71] and need to reconcile contradictory instructions.[72] They are also often traumatised by the kind of content they view every day, with no mental health support from companies, and this may also impact the quality/effectiveness of content moderation. Moreover, content moderators lack gender-sensitive training, in addition to often being alien to the local context, culture and language.[73]

The uneven application of community standards and the lack of legal representation of platforms in global South jurisdictions are problems reported to the Special Rapporteur during regional consultations in Africa, Asia, MENA and Latin America and the Caribbean (LAC).[74]

Reporting mechanisms, when available, are cumbersome, sometimes confusing, and often force users to attribute their experiences to predetermined categories that fail to capture the multifaceted nature of the abuse faced, in particular when

[68] Jankowicz, N., et al. (2021). Op. cit.

[69] PEN America. (2021). *No Excuse For Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users*. https://pen.org/report/no-excuse-for-abuse

[70] Hopkins, N. (2017, 21 May). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

[71] Perrigo, B. (2022, 14 February). Inside Facebook's African Sweatshop. *TIME*. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment

[72] Newton, C. (2019, 25 February). The trauma floor: The secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[73] Web Foundation. (2020, 25 November). The impact of online gender-based violence on women in public life. https://webfoundation.org/2020/11/the-impact-of-online-gender-based-violence-on-women-in-public-life

[74] Association for Progressive Communications submission to the call for inputs by the UN Special Rapporteur on the right to freedom of opinion and expression on the gendered dimensions of disinformation, July 2023.

TFGBV intersects with disinformation. A 2021 survey of US adults done by the Pew Research Center found that near 80% of respondents considered that social media companies are doing an only fair or poor job addressing online harassment.[75] Furthermore, most reporting mechanisms require the targets themselves to file a report. In addition to putting the onus on the harassed, these mechanisms force them to re-experience the abuse suffered and may lead to re-traumatisation. A recent study by Pen America refers to the "deep frustration, exasperation, and harm caused by the reporting mechanisms themselves." [76]

In terms of enforcement, those reporting abuses complain about the lack of feedback concerning their cases, the obscurity of processes and decisions that too often conclude that their experiences did not violate the platform's policies. The processes followed once a case is reported are not only unclear, but uneven across geographies and demographies. Moderation practices for content in any other language are not nearly as advanced or robust as they are for English-language content.[77]

Another point of concern raised by researchers and activists refers to the fact that content moderation practices and policies fail to account for coordination and the resulting dogpiling. While individual posts or messages may not reach the threshold of harmful content, their compounded volume does lead to a differentiated type of harm that is not captured by individual complaints – which are the only manner in which targets can report them. At the same time, when receiving individual complaints, content moderators are unable to consolidate individual reports.[78]

4.2 Specific responses to disinformation are gender-blind

Zakem, Wainscott and Arnaudo affirm that tech companies have been responding to disinformation through a wide range of measures that "vary widely in character and efficacy," but that can be classified as falling under one of three categories:

- Policies, product interventions and enforcement measures to limit the spread of disinformation.
- Policies and product features to provide users with greater access to authoritative information, data or context.

[75] Pew Research Center. (2021, 13 January). The State of Online Harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment

[76] Vilk, V., & Lo, K. (2023). *Shouting into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It*. PEN America. https://pen.org/report/shouting-into-the-void

[77] See, for example: Rogers, K., & Longoria, J. (2020, 20 October). Why A Gamer Started A Web Of Disinformation Sites Aimed At Latino Americans. *FiveThirtyEight*. https://fivethirtyeight.com/features/why-a-gamer-started-a-web-of-disinformation-sites-aimed-at-latino-americans and Paul, K. (2022, 6 October). Disinformation in Spanish is prolific on Facebook, Twitter and YouTube despite vows to act. *The Guardian*. https://www.theguardian.com/media/2022/oct/06/disinformation-in-spanish-facebook-twitter-youtube

[78] Jankowicz, N., et al. (2021). Op. cit.

- Efforts to promote a stronger community response and societal resilience, including digital literacy and internet access, to disinformation and misinformation.[79]

An assessment of the disinformation policies of Meta (Facebook/Instagram), Twitter, TikTok and Discord[80] evidenced that all four platforms prefer the conceptual approach of misleading information rather than disinformation. One possible reason for this may rest in an intentional attempt to move away from the motive or intention usually assumed for disinformation. To avoid judging a post's accuracy, most platforms have partnered with external third parties – fact-checking organisations – to counteract misinformation. With this action, they have developed human and technological review processes: human fact-checkers whose verification and rating are translated into a technical application, ranging from labelling and issuing warnings to limiting the amplification or visibility of content, among other actions.

The case of Twitter, however, is unique, as it is testing a geographically limited programme – Community Notes[81] – so that community members themselves can give context and rate the accuracy of a claim. It does this through an open and transparent process in which contributors leave notes on any given tweet, and if a sufficient number of contributors with different points of view rate that note as useful, the note will be publicly displayed on the tweet. Platforms' policies and tools to counter misinformation are gender-blind. At best, we find an intersection between these policies with other guidelines that directly or indirectly address gender-based violence on these platforms. The most obvious connections are related to community guidelines on violent speech, violent behaviour, harassment and bullying, and the non-consensual dissemination of intimate images. Policies to counter misinformation on platforms are generally aimed at addressing information that may contribute to endangering public safety, civic and electoral processes or public health. Some platforms also include other aspects in their anti-disinformation strategies. For example, Twitter[82] and Meta[83] include a prohibition against synthetic, manipulated or out-of-context media that can mislead and lead to harm. Twitter also has a specific policy to address disinformation in crisis cases.[84]

The platforms use a combination of technological tools to identify fake content through behavioural patterns and word lists, among other things, and human ones through teams that monitor the platform, partnerships with fact-checking organisations, and user reporting processes. All four platforms recognise that actions on content deemed misleading or false may be mistaken. Therefore, they

[79] https://counteringdisinformation.org/topics/platforms/0-overview-platforms

[80] Association for Progressive Communications submission to the call for inputs by the UN Special Rapporteur on the right to freedom of opinion and expression on the gendered dimensions of disinformation, July 2023.

[81] https://help.twitter.com/en/using-twitter/community-notes

[82] https://help.twitter.com/en/rules-and-policies/manipulated-media

[83] https://transparency.fb.com/policies/community-standards/misinformation#policy-details

[84] https://blog.x.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

offer appeal processes if a user understands that the action taken on a piece of content or account was wrong. The algorithms and human moderators whose job it is to identify and slow down disinformation are normally trained using English-language content.[85]

Companies have adopted a number of product features and technical interventions intended to help limit the spread of disinformation. One of the biggest issues these platforms have tried to address across the board is virality – the speed at which information travels on these platforms. As the Countering Disinformation Guide notes, "When virality is combined with algorithmic bias, it can lead to coordinated disinformation campaigns, civil unrest, and violent harm."[86] Facebook and Twitter, for example, have implemented interventions and features that work either to suppress the virality of disinformation and alert users to its presence or create *friction* that impacts user behaviour to slow the spread of false information within and across networks.[87]

These measures include, for example, the use of content labels and warnings; the use of algorithmic strategies to "down-rank" false or disputed information, decreasing the content's visibility; distribution limits placed on entire pages and websites that repeatedly share false news; and notifications to users who have engaged with certain misinformation and disinformation.

The CDT points out that some of the responses adopted for general disinformation may not be effective for disinformation targeted at groups based on race, gender and other categories. This is the case, for example, with changes made within a user interface, such as nudging or labelling:

> Is there a difference, for example, between the effectiveness of nudging to combat false information about "voting by mail", and the false narrative that Kamala Harris is not a U.S. citizen? Similarly, where gendered disinformation may include a combination of false information and abuse, how can nudging address both problems, one about veracity and the other about violence?[88]

Both Twitter and Facebook utilise automation to detect certain types of misinformation and disinformation and to enforce content policies. These systems played a more prevalent role during the pandemic as public health concerns required human content moderators to disperse from offices. The companies similarly employ technical tools to assist in the detection of inauthentic activity on their platforms.[89]

[85] Paul, K. (2022, 6 October). Op. cit.
[86] https://counteringdisinformation.org/topics/platforms/0-overview-platforms
[87] Ibid.
[88] Thakur, D., & Hankerson, D. L. (2021). Op. cit.
[89] https://counteringdisinformation.org/topics/platforms/0-overview-platforms

The lack of transparency with regard to the algorithms used, however, has been pointed out as a key concern.

Moreover, “the use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection” have been identified by researchers as a critical challenge, for they make automated detection hard, and often require very specific situational knowledge or in-depth cultural or language analysis to allow understanding.[90]

Platforms have also deployed strategies to promote authoritative content. These strategies have, to date, included labelling content that may be misleading or harmful to users, directing users to official information sources on important topics like voting or public health, and providing researchers and civil society observers with access to tools and data to better understand the information environment across various digital services. Despite being more common in relation to some topics, this practice is still rarely utilised in relation to gendered disinformation.

4.3 Attention economics, content curation, automated advertising and the amplification of gendered disinformation

Disinformation can be a profitable business to both creators and disseminators (through the monetisation of misleading content) and to the platforms that provide advertising and automated targeting infrastructure. The economic logic behind platforms – advertising and content curation built on attention economics – is the same logic followed by disinformation. Responses that seek to address the root causes of disinformation must, therefore, address this economic model.

The economics behind platforms is such that they collect in-depth data about how to influence our decisions, then sell that influence to the highest bidder. The more time one spends scrolling and clicking, the more data they can collect and the more ads they can sell. The practice of cultivating attention influences what is prioritised in people’s content and advertising feeds, as well as what they are recommended.[91] This has the effect of distorting how we, as users of these platforms, “see” the world in those spaces. The Council of Europe, for example, recognised that platforms’ prioritisation of certain values over others shapes the contexts in which individuals access and process information, and come to conclusions and decisions.[92]

[90] Jankowicz, N., et al. (2021). Op. cit.

[91] Tech Transparency Project. (2022, 10 August). Facebook Profits from White Supremacist Groups. https://www.techtransparencyproject.org/articles/facebook-profits-white-supremacist-groups

[92] Council of Europe. (2019). Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes. https://search.coe.int/cm/pages/result_details.aspx?ObjectId=090000168092dd4b

Content curation builds on the profiling and micro-targeting of individuals, with the ultimate goal of serving platforms' advertising purposes. While hosting and distributing user-generated content may be the more apparent service, the main revenue of large intermediaries stems from buying, selling and marketing for advertising.[93]

Targeted advertising results in algorithms being customised to cultivate attention and engagement. In order to retain attention, a specific type of message and language is commonly used. Because social media apps are caught in a race for our attention, they tend to promote more provocative, attention-grabbing content.[94] A study by Carrasco-Farré shows that misinformation, on average, is easier to process in terms of cognitive effort (3% easier to read and 15% less lexically diverse) and more emotional (10 times more reliant on negative sentiment and 37% more appealing to morality).[95]

As pointed out by Ryan et al., anyone can become an attention vendor. Attention vendors enter the market and distract or draw patrons or customers into monetised business models. This is often accomplished by monetising attention through advertising. They use complex statistical models to predict and maximize engagement with content.[96] Those behind the market of disinformation may have different goals, as indicated above, including simple profit – spreading disinformation is a lucrative and growing market.[97]

As clarified by Llanso et al.:

> Wittingly or not, platforms may actively contribute to the amplification of incendiary, controversial and divisive (dis)information as it directly aligns with the commercial and technological infrastructures of their recommendation systems that are optimized for user engagement. However, blaming recommendation systems alone ignores the fact that these infrastructures work in conjunction with users' own biased content and behavior, and are furthermore used and strategically exploited by sophisticated actors with more resources and experience than the average user, who can accordingly work the system and gain more political influence.[98]

[93] Norwegian Consumer Council. (2021). *Time to Ban Surveillance-Based Advertising: The case against commercial surveillance online*. https://storage02.forbrukerradet.no/media/2021/06/20210622-final-report-time-to-ban-surveillance-based-advertising.pdf

[94] Brady, W. J., et al. (2017). Emotion shapes the diffusion of moralized content in social networks. *Psychological and Cognitive Sciences, 114*(28), 7313-7318. https://www.pnas.org/doi/full/10.1073/pnas.1618923114

[95] Carrasco-Farré, C. (2022). The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications, 9*, 162. https://rdcu.be/dzMdg

[96] Ryan, C. D., et al. (2020). Monetizing disinformation in the attention economy: The case of genetically modified organisms (GMOs). *European Management Journal, 38*(1), 7-18. https://doi.org/10.1016/j.emj.2019.11.002

[97] Miller, C. (2018, 10 November). Meeting Kosovo's clickbait merchants. *BBC*. https://www.bbc.com/news/technology-46136513

[98] Llanso, E., et al. (2020). *Artificial Intelligence, Content Moderation, and Freedom of Expression*. Transatlantic Working Group. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

The GDI has carried out research to verify the role of ad-funded content that promotes misogynistic disinformation. It concluded that ad tech policy gaps and inadequate enforcement continue to facilitate the monetisation of disinforming content that perpetuates and spreads adversarial narratives directed at marginalised and at-risk groups – ad tech vendors are monetising content related to misogyny, showing clear supply-quality gaps and an inadequate enforcement of existing policies.[99]

4.4 Privacy and data protection

The point raised in the preceding section is also related to the fragilisation of privacy and data protection. Although platform companies also acquire people's personal data from external brokers, their ability to directly collect and mine first-party data is a major component of the attention economics model. This mining of data for advertising applies even to services that do not depend on "attention economics", such as search engines and search functionalities on other platforms.

Increasing user engagement does not only result in more advertising but also provides access to even more behavioural data and, consequently, improved targeted advertisement and increased profit.[100]

In her 2021 report on disinformation, the UN Special Rapporteur on freedom of opinion and expression highlighted that the spread of disinformation – with the gratuitous data collection and profiling techniques utilised by the online advertising industry – increasingly impacts our right to privacy. Data protection is key to re-orient the ad-driven business model of the digital economy.

Running on an automated basis that privileges transactions particularly enables micro-targeting with potentially harmful paid-for content. The system also facilitates advertising revenues' ability to flow not only to the platform concerned, but also sometimes to the producers of such problematic content, including via link-based traffic sent to their websites.

[99] Global Disinformation Index. (2023). *Ad-Funded Disinformation: Misogyny*. https://www.disinformationindex.org/files/gdi_misogyny-disinfo-ads-deck_jan-23.pdf

[100] Llanso, E., et al. (2020). Op. cit.; see also: Coalition to Fight Digital Deception. (2021). *Trained for Deception: How Artificial Intelligence Fuels Online Disinformation*. https://assets.mofoprod.net/network/documents/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation_T2pk9Wj.pdf

4.5 The need for improved transparency

Internet platforms, and in particular social networking services, need to increase transparency within their operations to allow researchers and activists a better understanding of the scope, dynamics and nature of disinformation.

The fact that social media is a privileged space for gendered disinformation also means that these platforms' access to privileged and proprietary data and metadata often uniquely positions them to understand the challenges posed by gendered disinformation.

Social media policies and enforcement actions are constantly evolving as the threat landscape constantly changes. The publication of transparency reports that update users and researchers about these changes and the provision of data are, therefore, key responses to allow more effective responses to gendered disinformation. The reports and data currently made available by platforms, however, are considered too limited and incomplete.

According to Ranking Digital Rights, for example, only five out of the 14 major platforms publish any data about actions taken to restrict advertising that violates their policies.[101]

Transparency about recommendation systems and the use of algorithms can help hold these systems accountable and enable more evidence-based policy making. Ad transparency is key – given the granularity with which advertisers are able to target users, the companies must provide much more information about why users are seeing a particular ad.

As highlighted by UNESCO:

> [A] considerable amount of information about companies' policies and outcomes is available in the public domain. However, while there are areas of overlap, particularly regarding questions of content removal, each company reports on different issues and in different ways, making simple comparisons impossible. Information about actual practices, not only of moderation, but especially of curation, the approach to trade-offs between rights, and the role of company interests, is usually less forthcoming.[102]

[101] https://rankingdigitalrights.org/bts22/indicators/F4b

[102] Puddephatt, A. (2021). *Letting the Sun Shine In: Transparency and Accountability in the Digital Age*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377231.locale=en

4.6 Consequences for perpetrators

One repeated concern shared by survivors with APC and the UN Special Rapporteur on freedom of opinion and expression during consultations, is that they feel that the consequences for perpetrators are feeble at best, or inexistent as a rule. This sense of a lack of redress was very present in all consultations and is disempowering for survivors.

Research has also shown that malicious actors, once having their accounts closed, quickly move to open new fake ones or move to other platforms.[103] Even serious responses such as "deplatforming" are at times seen as not having significant impact. This may be different for high level disseminators of disinformation – a small number of key accounts are too often behind large volumes of false content – and their deplatforming, in extreme cases and with the necessary due process and freedom of expression guarantees, may have an important impact.[104] But this remains a field open for study and verification.[105]

As has been largely pointed out by researchers and practitioners, some smaller platforms build their online presence precisely by selling their network as "niche", thus attracting more radical groups;[106] these spaces are largely unmoderated.[107]

[103] Morris, R. (2022, 30 December). Researchers warn of rise in extremism online after Covid. *BBC*. https://www.bbc.com/news/uk-politics-61106191

[104] Center for Countering Digital Hate. (2022). *The Disinformation Dozen: Why platforms must act on twelve leading online anti-vaxxers*. https://counterhate.com/wp-content/uploads/2022/05/210324-The-Disinformation-Dozen.pdf

[105] Center for Countering Digital Hate. (2021). *Malgorithm: How Instagram's algorithm publishes misinformation and hate to millions during a pandemic*. https://counterhate.com/wp-content/uploads/2022/05/210309-Malgorithm-Report.pdf

[106] Heilweil, P. (2021, 11 January). Parler, the "free speech" social network, explained. *Vox*. https://www.vox.com/recode/2020/11/24/21579357/parler-app-trump-twitter-facebook-censorship

[107] Nicas, J., & Alba, D. (2021, 10 January). How Parler, a Chosen App of Trump Fans, Became a Test of Free Speech. *The New York Times*. https://www.nytimes.com/2021/01/10/technology/parler-app-trump-free-speech.html

STATES AND STATE ACTORS

Globally, states have reacted in different ways to gendered disinformation. Most responses could be positioned along a spectrum that ranges from inaction to failed regulatory attempts and, in some concerning cases, the active use of public resources to promote gendered disinformation campaigns.

Bateman and Jackson stress that even "when leaders know what they want to achieve in countering disinformation, they struggle to make an impact and often don't realize how little is known about the effectiveness of policies commonly recommended by experts. Policymakers also sometimes fixate on a few pieces of the disinformation puzzle."[108]

Below, some key trends observed are presented and discussed. These trends have been organised under four areas of concern:

- Inaction or lack of attention paid to the distinct features of gendered disinformation;
- A focus on restrictive legal frameworks and/or criminalisation as a response to disinformation.
- Failure to protect gender-related speech as a carte blanche to gendered disinformation.
- Direct state action cracking down on women and gender-diverse people through the deployment of disinformation campaigns.

5.1 Inaction or lack of attention paid to the distinct features of gendered disinformation

APC has identified three trends in this regard: (i) some states simply ignore the different types of gendered attacks facilitated by technology that women and gender-diverse individuals face; (ii) states seek to address disinformation, but fail to consider its gendered dimensions; and (iii) states consider that other efforts aiming at addressing TFGBV are well suited to also address gendered disinformation.

(i) Inaction and failure to protect women and gender-diverse individuals

In relation to state inaction, APC collected numerous testimonies that accounted for the difficulties women faced in reporting the violence and attacks they were subject to. In most cases, activists and survivors were concerned with the lack of legal recognition of the fact that online threats, harassment and disinformation have real and actual impacts on their lives. Too often, according to reports received,

[108] Bateman, J., & Jackson, D. (2024). *Countering Disinformation Effectively: An Evidence-Based Policy Guide. Carnegie Endowment for International Peace.* https://carnegieendowment.org/files/Carnegie_Countering_Disinformation_Effectively.pdf

law enforcement officials would refuse to register the cases for a lack of applicable legal frameworks. Women and gender-diverse individuals also shared the difficulties faced even when legal frameworks were in place – they narrated stories of bias and prejudice in law enforcement officials, who questioned their experiences or sought to expose them to ridicule. Therefore, in addition to a lack of formal recognition of the specific challenges faced by women online, poor implementation of existing frameworks led to state inaction and omission, and a failure to comply with a state's international obligation to protect women against different types of violence.

(ii) Measures addressing disinformation are gender-blind

A second set of cases observed by APC were those referring to states that have shown concern in relation to disinformation and have sought to address it through different measures, some of which will be discussed below. However, in view of the conceptual challenges and conflation of concepts described above in Chapter 2, many of the attacks faced by women and gender-diverse people have been lumped together in research, analysis and, consequently, in the responses to different phenomena, with the result that inadequate attention has been paid to the distinct characteristics and impact of gendered disinformation. By not sufficiently examining its features, states may be missing the point of disinformation campaigns, which are often intentionally designed to exploit existing forms of discrimination.

APC has been troubled by some of the legal responses adopted by states to address disinformation, as pointed out below. Whenever states are addressing disinformation, such as through policies aimed at digital and information literacy and social media regulation, including promoting due diligence, addressing virality and promoting the demonetisation of illegal content, states should always pay specific attention to the gendered dimensions of disinformation and ensure that any anti-disinformation measures not only abide by international freedom of expression standards, but also apply a gender lens to their design, planning and implementation.

(iii) Equating gendered disinformation with TFGBV

Finally, we have observed that some states are indeed concerned with supporting women in relation to violence and attacks facilitated by technology, but they have not addressed the particularities of gendered disinformation or have sought simply to equate it with TFGBV. However, some of these efforts, despite their merits, may fail to address issues pertaining to coordination, malign intent and the very particular harms created by the overarching anti-gender narratives that gendered disinformation seeks to promote. In view of their commitment to women's rights, APC recommends that these states engage in further monitoring and research and establish partnerships and cooperation with civil society and academia to develop improved data and knowledge on the specificities of gendered disinformation and the required responses to it.

5.2 A focus on restrictive legal frameworks and/or criminalisation as a response to disinformation

Legislative and regulatory solutions are critically important, but they are also fraught, complex and hard to get right without further undermining the safety and free speech of individuals and communities already struggling to be heard online. On top of that, it is important to acknowledge that there are sub-optimum levels of governmental knowledge in many countries about the complexities of platforms' operations, and this may impact the adoption of well-calibrated laws, regulations and policies.

In practice, APC has observed that most responses to the challenges of online harms have resorted to overly restrictive legal frameworks and the use of criminalisation. The clearest examples are restrictive platform regulations and the passing of so-called "fake news" bills. These alleged solutions may end up imposing serious restrictions on freedom of expression.

Most proposed legislation will not pass the three-part test according to which interferences with freedom of expression are legitimate only if they (a) are prescribed by law; (b) pursue a legitimate aim; and (c) are "necessary in a democratic society".[109] In particular, regulatory proposals in this area tend to be disproportionate and ignore the institutional reality (conservative courts and their composition, poorly trained and non-gender-aware law enforcement agents) in many countries.

(i) Criminalisation of speech

States are hastily putting forward policy and legislative proposals that they affirm aim at addressing disinformation, and this includes the passing of criminal provisions.[110] However, without the comprehensive kinds of evidence that activists and researchers call for, these solutions may fall short and could likely harm the same communities they aim to protect.

[109] In accordance with Article 19(1) of the International Covenant on Civil and Political Rights (ICCPR), freedom of opinion may not be subject to any interference. Article 19(2) defines freedom of expression as multidirectional ("seek, receive and impart"), unlimited by viewpoint ("information and ideas of all kinds"), without boundaries ("regardless of frontiers"), and open-ended in form ("or through any other media"). Article 19(3) provides narrow grounds on which governments may restrict the freedom of expression, requiring that any limitation be provided by law and be necessary for respect of the rights or reputations of others, or for the protection of national security or public order, or of public health or morals. That is, such limitations must meet the tests of necessity and proportionality and be aimed only towards a legitimate objective.

[110] For measures against disinformation across the world, see, for example: Funke, D., & Flamini, D. (n/d). A guide to anti-misinformation actions around the world. Poynter. At the EU level, lawmakers have been working on a directive that could criminalise some types of gender-based violence: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0105

A number of these measures seek to impose penalties on or create crimes related to expression and speech. However, the use of criminal provisions in this regard is only accepted in very extreme and limited situations under international human rights law.

In practice, women and gender-diverse people have shared their fear that these same provisions could be used against them. As explained in the preceding chapters, too often activists and outspoken individuals on gender-related themes are accused of spreading disinformation by anti-rights groups or by conservative government officials. These activists are repeatedly accused of promoting gender ideology, Western values and anti-family principles, and strong anti-disinformation provisions could be used to silence them. The use of criminal law to address disinformation is ineffective and risky and may lead the way to abuse. In 2019, in Poland, for example, hostile attitudes toward LGBT persons led regions and municipalities to begin declaring themselves "LGBT Ideology Free" or to join a government-supported Family Charter that called for the exclusion of LGBT people from Polish society. [111]

This is evidenced by the use of other types of existing legislation to persecute or criminalise women and gender-diverse people who express their views on gender-related issues or react against disinformation campaigns. We have seen this in relation to blasphemy laws, defamation laws, and the use of civil liability lawsuits against feminists, LGBTQI+ groups, journalists or women denouncing gender discrimination and violence.

A few of the many examples of this are:

- Dina Smailova, head of the NeMolchi ("Do Not Be Silent") Foundation in Kazakhstan, who faced criminal charges for defamation in 2020 after she publicly criticised a well-known Kazakh blogger and former parliamentarian Tanirbergen Berdongarov about his public vilification of a gang-rape survivor.[112]
- Lim writes that in Malaysia, in July 2014, a fatwa was issued by the religious authorities in the state of Selangor, among others, declaring Sisters In Islam (SIS) – a non-profit women's rights organisation in Malaysia – as an organisation that "practices liberal ideas and religious pluralism" and is therefore a "deviant organisation". The fatwa further declared that any publications by SIS must be banned, and directed the Malaysian Communications and Multimedia Commission to block "any social media sites that contravene Islamic teachings and Syariah principles." [113]

[111] Human Rights Watch. (2022, 15 December). Poland: Rule of Law Erosion Harms Women, LGBT People. https://www.hrw.org/news/2022/12/15/poland-rule-law-erosion-harms-women-lgbt-people

[112] Equality Now. (2020, 26 January). Women's human rights defender faces defamation charges in Kazakhstan. https://www.equalitynow.org/news_and_insights/womens_rights_lawyer_faces_defamation_charges_in_kazakhstan

[113] Lim, S. (2019, 3 October). We are Sisters in Islam. *GenderIT.org*. https://genderit.org/feminist-talk/we-are-sisters-islam

- The advocacy group Know Your IX,[114] which lobbies on behalf of student survivors of sexual violence, says that 23% of students who make Title IX complaints are threatened with defamation suits by their alleged abusers. [115]

(ii) Overly restrictive platform regulations

There is currently an impetus at the national level to regulate internet platforms. This has been seen as the solution to several diverse challenges relating to the online context, many of which have been blamed on a perceived lack of accountability of internet platforms for human rights violations facilitated by the technologies they provide and disseminate. These are indeed legitimate concerns. However, many of the solutions developed have been problematic.

Most regulatory proposals are overly restrictive and normally lead to one or some of the following problems:

- Poor definitions of what constitutes unlawful or harmful content.
- Outsourcing regulatory functions to companies.
- Over-emphasis on take downs and the imposition of unrealistic timelines.
- Over-reliance on artificial intelligence (AI) mechanisms.

Many regulatory proposals are also narrow and short-sighted, focusing solely on the objective of controlling content and disregarding the adoption of other legal provisions that could be directed at addressing broader aspects of the governance of the digital space and digital market. These could include measures aimed at altering the current enabling environment where gendered disinformation is promoted and amplified in practice and by design. Such provisions could tackle issues such as market concentration; business models that rest on the exploitation of data and sensationalist content; a lack of transparency relating to the operation of platforms, including the use of algorithms; as well as issues of participation in and user engagement with the creation of community rules. Some of these issues will be further discussed in the section on companies' responses below.

As clarified by the former UN Special Rapporteur on freedom of opinion and expression David Kaye:

> Smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums. […]

[114] Nesbitt, S., & Carson, S. (2021). *The Cost of Reporting: Perpetrator Retaliation, Institutional Betrayal, and Student Survivor Pushout*. Advocates for Youth. https://knowyourix.org/wp-content/uploads/2021/03/Know-Your-IX-2021-Report-Final-Copy.pdf

[115] https://twitter.com/knowyourIX/status/1532122403534430208?ref_src=twsrc%5Etfw

> Companies must embark on radically different approaches to transparency at all stages of their operations, from rule-making to implementation and development of “case law” framing the interpretation of private rules.[116]

An interesting alternative in terms of state-led efforts to regulate internet platforms has been the European Commission’s Digital Services Act (DSA). This regulation seeks to be attentive to some of the concerns raised above. The DSA includes the removal of illegal goods, services and content, advertising transparency measures and obligations for large platforms to take action against the abuse of their systems. Tech companies could face severe fines for noncompliance, with very large online platforms (VLOPs) facing fines of up to 6% of their global revenue for a serious breach of the rules.

The DSA establishes that VLOPs shall conduct comprehensive assessments of systemic risks to fundamental rights from their services (Article 34), develop and implement mitigation measures (Article 35) and be subjected to independent audits to assess their efforts (Article 37). Negative consequences in relation to TFGBV are explicitly mentioned as one of these specific systemic risks. These mandatory due diligence obligations, therefore, require platforms to annually assess and mitigate the risk of TFGBV in their operation. This provision is especially important to address forms of behaviour that do not amount to illegal acts (since illegal acts will be covered by a specific EU Directive under construction).[117]

Allen suggests:

> [These] risk assessments should be envisaged [as] Human Rights Impact Assessments (HRIAs), which are extensive, cyclical processes of identifying, understanding, assessing and addressing the adverse effects of the business project or activities on the human rights enjoyment of impacted rights-holders. This process will not only identify specific impacts but their severity and how they may *intersect* with other fundamental rights violations.[118]

The task is not simple. Striking a balance between the protection of free expression, addressing illegal content and creating a safe online environment will be challenging. It is important to stress that mandatory due diligence obligations must be accompanied by effective accountability mechanisms to ensure that online platforms comply with their responsibilities. Additionally, consultation with civil society is key, as well as the availability of data for researchers and activists to develop evidence-based policy recommendations.

[116] Kaye, D. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. https://digitallibrary.un.org/record/1631686?ln=en

[117] Martins, P. (2024). *How Can Impact Assessments Improve Protection from TFGBV?* Centre for International Governance Innovation. https://www.cigionline.org/publications/how-can-impact-assessments-improve-protection-from-tfgbv/

[118] Allan, A. (2022, 1 November). An Intersectional Lens on Online Gender Based Violence and the Digital Services Act. *Verfassungsblog*. https://verfassungsblog.de/dsa-intersectional

Governments across the world, ranging from democracies to totalitarian regimes, are eager to regulate the online space and the DSA will probably be referenced as a standard to be replicated. APC believes that lawmakers should carefully consider context-specific challenges and consult with local civil society and researchers before simply importing the DSA framework as a solution to gendered disinformation and TFGBV. These challenges, as seen above, are highly context specific and require local expertise. Attention should be paid to monitoring the implementation of the DSA and its effectiveness in upcoming years.

5.3 Failure to protect gender speech as a carte blanche to gendered disinformation

Expression is not free for many women or gender non-conforming people. As stressed by the UN Special Rapporteur on freedom of opinion and expression in her 2021 report freedom of expression and gender justice, while the international human rights system has largely focused on censorship as repressive action by the state, non-state and private actors – whether social, cultural, religious or commercial – often play a leading and visible role in gendered censorship alongside the state, using various social mechanisms that "mute women's voices, deny validity to their experience, and exclude them from the political discourse."[119]

Interpretations of culture, religion and tradition that subordinate women within patriarchal systems and structures are often used to justify discriminatory laws, institutions, rules and regulations. They disempower women and undermine their agency to express themselves or define their own culture, religion and traditions, while at the same time assigning them the role of preserving cultural traditions and values. It creates a form of structural silencing that leads women to self-censor. Many women fear the consequences of challenging existing norms and practices, or lack the support mechanisms needed to take action. In some contexts, the fact that a woman, especially a young woman, is expressing her views is enough for her ideas to be discredited, and for the speaker to be socially sanctioned.

States have an obligation not only to respect freedom of opinion and expression, but also to proactively remove the structural and systemic barriers to equality, including sexual and gender-based violence, which impede women's and gender-diverse people's full enjoyment of freedom of opinion and expression. When states fail to protect women's and gender-diverse people's right to expression, not only do they fail their international obligation to protect, they also provide a carte blanche to perpetrators, including malicious actors distributing gendered disinformation content.

[119] Manne, K. (2017). Down Girl: *The Logic of Misogyny*. Oxford University Press.

As highlighted above in Chapter 2, when states fail to uphold women and gender-diverse individuals' freedom of expression, this creates a conducive environment to discrimination, violence and the spread of disinformation against these same groups. More than simply proposing reactive solutions when responding to TFGBV and gendered disinformation, states should address the root causes of these social problems, committing themselves to gender justice and the needed structural, systemic and long terms changes it requires.

5.4 Direct state action cracking down on women and gender-diverse people through the deployment of disinformation campaigns

One of the most striking modalities of gendered disinformation reported by activists and researchers is that of state-sponsored and state-aligned disinformation. Two alarming aspects of this type of disinformation are: (i) that it diverts public resources that should be used to promote gender equality and anti-discrimination policies and actions to instead attack women, gender-diverse people and their struggles; and (ii) that the authority of those in posts of political power has great potential to mobilise and influence their electorate and followers, potentially leading to significant levels of amplification or even violence.

The use of state resources to harass individuals was clear in the case of the campaign orchestrated against journalist Ghada Oueiss, where in addition to disinformation, Oueiss was targeted with surveillance/spyware.[120] As highlighted by Jones, the case evidences the potential hacking of her phone using Pegasus spyware and the following mass dissemination of this hacked content on social media via a mix of pro-regime influencers and anonymous accounts. Jones writes:

> [B]y putting doxed malinformation on social media, the perpetrator deflects accusations that it might be a potentially state-orchestrated operation. Only governments can theoretically get hold of hacked content from tools such as Pegasus. Thus, the idea that the story broke via unknown anonymous accounts gives the story the veneer of grassroots legitimacy when, in fact, it could have been planted. This is compounded by the fact that the smear campaign was chiefly dominated by verified accounts based out of Saudi Arabia and the UAE, who at the time had severed diplomatic relations with Qatar, and had targeted Al Jazeera and its journalists through information wars and also spyware.[121]

[120] Shilad, J. (2021, 11 February). Al-Jazeera's Ghada Oueiss on hacking, harassment, and Jamal Khashoggi. *Committee to Protect Journalists*. https://cpj.org/2021/02/ghada-oueiss-hacking-harassment-jamal-khashoggi

[121] Jones, M. O. (2021). State-aligned misogynistic disinformation on Arabic Twitter: The attempted silencing of an Al Jazeera journalist. *Open Information Science, 5*(1), 278-297. https://doi.org/10.1515/opis-2020-0126

The case of Maria Ressa also exemplifies how campaigns of disinformation against women journalists have been triggered by high level politicians and their followers. According to a report by the National Endowment for Democracy, after the election of Rodrigo Duterte as president of the Philippines in 2016:

> [I]ndependent journalists and opposition politicians have been targeted by systematic campaigns of online harassment, and investigators have identified networks of pro-government bloggers and automated social media accounts engaged in a concerted effort to tarnish the credibility of the independent Philippine press and bolster support for President Duterte.[122]

The ICFJ conducted a forensic analysis of the torrential social media attacks on Maria Ressa over a five-year period (2016-2021). These attacks also created an enabling environment for Ressa's persecution and prosecution in the Philippines. According to the ICFJ:

> The attacks against Ressa are fueled by Duterte, who has publicly condemned her – while musing that journalists are not exempt from assassination.
> His government has also employed a number of the key actors who have targeted Ressa online. And the worst attacks against her appear to have been orchestrated.[123]

A study looking at state-sponsored or state-supported gendered disinformation campaigns in Poland and the Philippines observed that campaigns aimed to push the narrative that women are not good political leaders. The majority of gendered disinformation examined through the course of this research was state-aligned in content – that is, aligned with state interests, attacking critics of the state and so forth. In the Philippines, for example, supporters of Duterte were directly involved in disseminating and amplifying that content.[124]

Research by Bradshaw and Henle also shows how gendered disinformation campaigns against feminism and women's rights were orchestrated by state-sponsored accounts from Iran, Russia and Venezuela, with high-profile feminists being commonly targeted.[125] Bradshaw and Henle concluded:

[122] https://www.ned.org/wp-content/uploads/2018/02/Maria-Ressa-on-Digital-Disinformation-and-Philippine-Democracy-in-the-Balance.pdf

[123] Posetti, J., Maynard, D, & Bontcheva, K. (2021). *Maria Ressa: Fighting an Onslaught of Online Violence*. International Center for Journalists. https://www.icfj.org/sites/default/files/2021-03/Maria%20Ressa-%20Fighting%20an%20Onslaught%20of%20Online%20Violence_0.pdf

[124] Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020). *Engendering Hate: The Contours of State-Aligned Gendered Disinformation Online*. Demos. https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online

[125] Bradshaw, S., & Henle, A. (2021). The Gender Dimensions of Foreign Influence Operations. *International Journal of Communication*, 15, 4596-4618. https://ijoc.org/index.php/ijoc/article/view/16332

> [F]oreign state actors co-opted intersectional critiques and countermovement narratives about feminism and female empowerment to demobilize civil society activists, spread progovernment propaganda, and generate virality around divisive political topics. [...] [A]mplifier accounts – particularly from the Russian IRA and GRU – drove more than one-third of the Twitter conversations about feminism and women's rights. [...] [H]igh-profile feminist politicians, activists, celebrities, and journalists were targeted with character attacks by the Russian GRU. These attacks happened indirectly, reinforcing a culture of hate rather than attempting to stifle or suppress the expression of rights through threats or harassment.[126]

Indian Prime Minister Narendra Modi has followed Twitter accounts responsible for rape and death threats against female politicians in his own government; his party has been accused of running a "troll army" that targets political opponents, especially prominent female figures, with online harassment, abuse and disinformation campaigns.[127] After publishing on the use of disinformation by Jair Bolsonaro in the 2018 presidential elections in Brazil, Patrícia Campos Mello was herself the target of an intense disinformation campaign on social media that led to serious offline threats: "People started calling my phone and sharing the public events I was going to with email lists of Bolsonaro supporters," she shared in an interview. "It got to the point where I couldn't go out of my house."[128]

Disinformation campaigns promoted by foreign sources are directly linked to national security and relate to the promotion of geo-political interests. In the case of Ukraine, for example, researcher Cori Fleser argues that:

The Kremlin has intentionally targeted and exploited societal gender fault lines through hybrid warfare as a reliable tactic for destabilizing cohesion and unity among populations throughout Europe. Though its hybrid campaigns focus on many issues, gender issues are some of the most divisive and polarizing for local populations, making them ripe for targeted disinformation. In the Ukraine conflict, for example, the Kremlin accused a woman who gave birth in the immediate aftermath of the Mariupol maternity hospital bombing of being an actress paid by Ukraine to sow uncertainty about the reality of its operations.[129]

[126] Ibid.
[127] Di Meco, L., & Wilfore, K. (2021, 8 March). Op. cit.
[128] Angwin, J. (2022, 11 June). Brazil on the Brink of a Disinformation Disaster. *The Markup*. https://themarkup.org/newsletter/hello-world/brazil-on-the-brink-of-a-disinformation-disaster
[129] Fraser, C. (2022, 15 August). Beyond munitions: A gender analysis for Ukrainian security assistance. *Atlantic Council*. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/beyond-munitions-a-gender-analysis-for-ukrainian-security-assistance/#introduction

COMMUNITY RESPONSES

As with disinformation in general, gendered campaigns need to be understood as an ecosystem of cross-platform interaction and contamination, thus requiring articulated and systematic counter-disinformation initiatives.[130] In addition to responses by states and companies, it is crucial that community responses are also supported, promoted and funded.

Scott emphasises that most existing responses to gendered disinformation tend to be reactive, rather than proactive, and more ad hoc than systematic:

> Reactive interventions, such as content tagging or removal and fact-checking, myth-busting, or otherwise correcting the record in response to direct attacks, are generally insufficient to reverse the harms caused by gendered disinformation, from reputational damage and self-censorship to withdrawal from public and digital spaces and sowing seeds of distrust and discord.[131]

APC has also collected the same feedback during consultations – most states' and companies' current responses to gendered disinformation are reactive. A look into community responses allows us to further observe proactive responses that not only seek more systematic and holistic approaches to gendered disinformation, but also demonstrate that women are not victims, but rather autonomous, creative and courageous actors fighting against this and other attacks against gender justice.

As per a systematisation suggested by APC,[132] there are eight main categories of responses that have been promoted by different kinds of groups, including civil society, social movements, collectives of artists and communicators, and researchers, among others, which provide examples of important and inspiring manners in which women are standing up against gendered disinformation:

- Counterspeech initiatives
- Support for survivors and targets
- Resources for targets and for bystanders who want to support them
- Social media monitoring
- Early warning systems
- Awareness raising, media literacy and capacity building
- Research and evidence gathering.

[130] Sessa, M. G. (2022, 14 November). The disinformative ecosystem: Link sharing practices on Telegram as evidence of cross-platform amplification. *EU DisinfoLab*. https://www.disinfo.eu/publications/the-disinformative-ecosystem-link-sharing-practices-on-telegram-as-evidence-of-cross-platform-amplification

[131] https://counteringdisinformation.org/topics/gender/3-current-approaches-countering-gendered-disinformation-and-addressing-gender

[132] Association for Progressive Communications submission to the call for inputs by the UN Special Rapporteur on the right to freedom of opinion and expression on the gendered dimensions of disinformation, July 2023.

Counterspeech initiatives

The first set of responses identified seeks to challenge the underlying messages promoted in gendered disinformation campaigns, to mock stereotypes, denounce hatred and speak out against falsity.

In some cases, those targeted may engage in such counterspeech themselves. In the Philippines, Leni Robredo has been a target of sexualised distortion campaigns, including ones accusing her of having a sexual relationship with a married lawmaker, her hairdresser, and two government officials. False news and social media stories about her "boyfriends" and a supposed pregnancy led her to create a series of Facebook videos addressing each false claim against her, debunking them one by one.[133]

In other cases, however, groups of allies join together in countering disinformation. For example, after a singer-songwriter in the Philippines called out a male TV host for his sexist remarks, #HijaAko trended as women and girls used the tag to share their stories and call out misogyny.[134] When people associated with #GamerGate attempted to hijack the #takebackthetech hashtag by posting false claims about APC and engaging in the online abuse of feminist internet rights activists associated with the campaign, the organisations and individuals involved in the campaign joined forces with their allies to reclaim the narrative in a tweet storm. In addition, APC issued a statement correcting the false claims made about its work.[135]

There have also been efforts to automate some of this work. In Canada, Areto Labs – a women-owned technology company – launched the software platform Areto,[136] which tracks online abuse, displays key trends on a personalised dashboard and provides alerts and creative counteractions. This tracking takes place on Twitter and Instagram in a range of languages (including Arabic, Hindi, and Indonesian). The platform evolved out of an earlier initiative by the entrepreneurs, ParityBOT, which detected "problematic tweets about women candidates" in elections and responded with "positive messages". It thus served "both as a monitoring mechanism and a counterbalancing tool."[137]

In situations where hate narratives and disinformation are rampant, correct information may sometimes be difficult to come by. Some organisations have taken the initiative to fill these gaps in the information ecosystem. For example, in

[133] https://counteringdisinformation.org/topics/gender/4-promising-approaches-gender-sensitive-counter-disinformation-programming

[134] Bautista, N. (n/d). #HijaAko: Why Filipinas Are Sick Of Victim Blaming In The Philippines. *Cambio & Co.* https://www.shopcambio.co/blogs/news/hijaako-these-filipinas-are-brave-young-and-sick-of-the-victim-blaming

[135] APCNews. (2015, 11 October). Facts on #TakeBacktheTech. *Association for Progressive Communications*. https://www.apc.org/en/pubs/facts-takebackthetech

[136] https://www.aretolabs.com

[137] Di Meco, L., & Brechenmacher, S. (2020, 30 November). Tackling Online Abuse and Disinformation Targeting Women in Politics. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/2020/11/30/tackling-online-abuse-and-disinformation-targeting-women-in-politics-pub-83331

Indonesia, in 2017, Qbukatabu was born in precisely such a context, "with a belief that access to information and services with a feminist and queer perspective should be available in online spaces."[138] Also in Indonesia, because information about transmen remained limited, "Transmen Indonesia collaborated with Transmen Talk Indonesia to make an educational video about female-to-male transgender people. The video aims to bring awareness to everyone who want to know further about transmen."[139] Both Qbukatabu and Transmen Indonesia have also "tried to be strategic in creating hashtags for posting content on social media."[140] However, these hashtags do attract negative comments as well.

Support for survivors and targets

Support for those targeted by disinformation can take a number of forms.

(i) Creating online communities

Some initiatives "have established online communities of supporters who are ready to support the targets of these attacks with counterspeech efforts," as well as "other supportive services such as monitoring the digital space where the attack is taking place and assisting the target of the attack in reporting the incident."[141]

For example, TrollBusters[142] is an at-the-ready US-based network of supporters who respond to women journalists' reports of online harassment by monitoring the targets' social media accounts for continued attacks, sending continued, positive counter-messaging, helping report content to the platforms and authorities and other supportive services.

Right To Be[143] (earlier known as Hollaback!) runs HeartMob, a storytelling platform and safe space where those targeted can "share their harassment story, get support, and help others experiencing harassment." Journalists who want to use the platform "can create a special 'journalist account,' as well as tag their stories as journalist stories, allowing for more specific and attentive support" from the HeartMob community and other colleagues. This latter feature was developed in collaboration with the International Women's Media Foundation (IWMF) and the Coalition Against Online Violence. Right To Be also provides training and resources on how "to respond to, intervene in, and heal from harassment" for those targeted, bystanders, and their institutions and communities.

[138] Rizqy, R., & Andriyanti, Y. D. (2022, 22 March). We Rise, We Heal, We Resist. *GenderIT.org*. https://genderit.org/feminist-talk/we-rise-we-heal-we-resist
[139] Ibid.
[140] Ibid.
[141] https://counteringdisinformation.org/topics/gender/4-promising-approaches-gender-sensitive-counter-disinformation-programming
[142] http://www.troll-busters.com
[143] https://stories.righttobe.org

Some organisations, such as HER Internet[144] in Uganda, have proactively worked to build their own "alliances and networks as support systems for [the] mitigation of impact and countering false narratives."[145]

For companies and organisations that want to be proactive in keeping their employees safe, Tall Poppy[146] provides "accessible, step-by-step guidance to digital safety" for all employees. This includes digital security awareness training, step-by-step guidance on setting up safeguards, and incidence response and follow-up support if an employee is attacked nevertheless. They also provide proactive risk management services, including digital footprint assessments and social media and hate site monitoring.

(ii) Raising awareness about disinformation incidents and campaigns

Support can also take the form of awareness raising about disinformation incidents and campaigns, and letters of support from allies. For example, Rizqy and Andriyanti write:

> The LGBT movements in Indonesia have to take few steps back because the spaces for freedom of expression and assembly have been taken away by people and groups who act arbitrarily in the name of morality and religion. However, supports poured in from various individuals and groups such as human rights organisations, at the local and national level as well as various statements of support put out by national human rights institutions that gave hope for the continuation of the struggle for LGBT rights in Indonesia. The issuance of security guarantees for LGBT activists to convey their aspirations across the country from the National Human Rights Commission, made them (the LGBT communities) feel somewhat safer.[147]

Such efforts can also develop at the global level. Thus, there is value in "bringing instances of gendered disinformation to the attention of the United Nations and the international community who can defend the targets of campaigns through public statements of support."[148] Examples of such interventions abound.

For instance, in February 2022, two UN Special Rapporteurs issued a statement "calling on India to end relentless misogynistic and sectarian attacks against an investigative journalist," Rana Ayyub, by far-right Hindu nationalist groups.[149]

[144] https://www.herinternet.org

[145] Kyogabirwe, L. (2023, 6 February). Pushing Back Against Gendered Disinformation in Uganda. *CIPESA*. https://cipesa.org/2023/02/pushing-back-against-gendered-disinformation-in-uganda

[146] https://www.tallpoppy.com

[147] Rizqy, R., & Andriyanti, Y. D. (2022). Op. cit.

[148] EU DisinfoLab. (2021, 20 October). Gender-Based Disinformation: Advancing Our Understanding and Response. https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response

[149] UN News. (2022, 21 February). Halt All Retaliation Attacks against Indian Journalist Rana Ayyub – UN experts. https://news.un.org/en/story/2022/02/1112362

(iii) Alerting international media

Alerting international media can also be of value, "as this is one of the fastest ways to motivate platforms to put in place protection measures or take action on disinformation campaigns."[150] In the case of Chinese-American journalist Leta Hong Fincher, for example, Twitter took action on some of the content of a disinformation campaign targeted against her, and verified her profile, after an email exchange and a public awareness campaign led by the Coalition for Women in Journalism.[151]

Resources for targets and bystanders who want to provide support to them

A variety of civil society organisations have started to develop trainings and toolkits to support those targeted by gendered disinformation in particular or by online gender-based violence more broadly. They frequently cover a variety of topics. Two areas that have attracted specific attention are digital safety and security and effective counterspeech. Other topics that may receive attention include how to document and report disinformation and harassment and how to provide support as a bystander. Most trainings and toolkits have been developed by or under the leadership of organisations based in the global North.

Here are some examples:

- #ShePersisted has developed a Digital Resilience Toolkit for Women in Politics, meant specifically for women elected leaders, activists, election consultants working with women's campaigns, journalists, and the broader international women's community. It covers digital security, incident reporting, and counterspeech. The toolkit also includes a detailed guide for how to counter disinformation related to sexual and reproductive health and rights (SRHR) specifically.[152]
- The Coalition Against Online Violence's Online Response Hub has resources for journalists facing online violence, as well as for newsrooms and for those seeking to support others. It lists a large number of studies, toolkits and guides. This includes an entire section on newsroom protocols. It also includes "Know Your Trolls", a course designed to "help journalists identify the abuse they are receiving online and who may be behind it" as well as to "offer some key strategies that may help journalists to be better prepared." The course

[150] EU DisinfoLab. (2021, 20 October). Op. cit.

[151] Jankowicz, N., et al. (2021). Op. cit.

[152] Wilfore, K. (2022). *A Digital Resilience Toolkit for Women In Politics: Persisting and Fighting Back Against Misogyny and Digital Platforms' Failures*. #ShePersisted. https://r2g26a.n3cdn1.secureserver.net/wp-content/uploads/2022/06/ShePersisted_Digital_Resilience_Toolkit.pdf

was developed by the International Women's Media Foundation (IWMF) and is available in several languages. The Online Response Hub is a project of the IWMF with the International Center for Journalists (ICFJ), under the banner of the Coalition Against Online Violence.[153]

- Glitch provides a range of resources: various strategies women in public life can use to stay safe online (ranging from digital security to how to build a positive online campaign), how to document online abuse and how to be an active bystander, among others. Some of these resources are developed specifically to support Black women.[154]
- PEN America's Online Harassment Field Manual is a training guide for journalists, writers, activists and artists who identify as women, Black, Indigenous and/or people of colour (BIPOC) and/or LGBTQIA+, as well as for those who are bystanders to the harassment of these people, and for their employers. It includes sections on safety and security, on building a community of supporters and developing counterspeech messages, and on how employers can support staff experiencing online harassment, including through counterspeech, among a range of other topics.[155]
- Equality Labs offers digital security trainings to "help activists learn how to protect devices, identities, networks, and organizations," consultations to "provide rapid response and organizational support to organizations, collectives, and individuals," and audit and technical support which consists of "running analytics and asking the right questions, to support organizations to adopt a holistic feminist approach that gives people the tools they need to be safe." Equality Lab's work centres the leadership of South Asian caste-oppressed, queer, and religious minority communities.[156]
- Tactical Tech offers the Gendersec Curricula, "a resource that introduces a holistic, feminist perspective to privacy and digital security trainings, informed by years of working with women and trans activists around the world." It includes, among other things, a Hacking Hate Speech workshop on how to set up an online support network, create textual and visual counterspeech content and deploy a counterspeech campaign.[157]

[153] https://onlineviolenceresponsehub.org
[154] https://glitchcharity.co.uk/resources
[155] https://onlineharassmentfieldmanual.pen.org
[156] https://www.equalitylabs.org/work/digital-security
[157] https://en.gendersec.train.tacticaltech.org

Social media monitoring

Researchers, practitioners and civil society actors are increasingly engaging in social media monitoring activities to inform their understanding of gendered disinformation, to identify entry points to disrupt gendered disinformation, viral misinformation and hate speech, and to advocate for laws or regulations that are responsive to the growing challenges of online gender-based violence and the spread of harmful gendered content online.[158]

In 2021 and 2022, for example, Panos South Asia (PSA) monitored the Nepali media sphere "with special focus on hate speech targeted at politically active women on social media." It explained:

> With the year 2022 being the election year in Nepal, PSA has been studying trends in gendered online violence against women in politics and women aspiring to join politics. [...] A team of Panos media monitors studied the social media sphere with special focus on Facebook and Twitter for misogynistic content and hate speech.

On the basis of this research, PSA released two quarterly reports with its findings, documenting where gendered disinformation emerges, what type(s) of content it contains and who the actors that create it are.[159]

The initiative is aimed at identifying the phenomena of misinformation and disinformation, using a gender sensitive lens to locate examples of misogyny in the online public sphere, helping to better understand the latest trends and techniques of online manipulation as well as the means to tackle it so that citizens can receive the necessary facts to make more informed political choices.

The overall aim is to reduce the level of harm caused due to the spread of deliberate lies and hate speech, strengthen the awareness of stakeholders on the multiple negative impacts of dis/misinformation and hate speech around elections and the role that political parties can play in countering such phenomena, and to promote greater accountability and transparency in public life.[160]

[158] https://counteringdisinformation.org/topics/gender/4-promising-approaches-gender-sensitive-counter-disinformation-programming

[159] South Asia Check. (2022, 31 July). Panos releases second media monitoring report on online gendered violence against women. https://southasiacheck.org/in-public-interest/panos-releases-second-media-monitoring-report-on-online-gendered-violence-against-women

[160] Panos South Asia. (2022). *Analysis of Gendered Violence in Social Media against Women in Politics in Nepal*. Panos. https://southasiacheck.org/wp-content/uploads/2022/07/Analysis-of-Gendered_web.pdf

Early warning systems

Consistent social media monitoring can enable the development of early warning systems.

In the EU for example, Maldita.es, a fact-checker organisation based in Spain, is partnering with Citibeats, an ethical social understanding service, to create DEWARD (Disinformation Early Warning Data Tool). MediaFutures describes DEWARD as "a citizen-informed data-driven service that provides fact-checking organisations, researchers, and other stakeholders with timely information about emerging and potential disinformation campaigns relating to gender, migration, and climate change".

Monitoring social conversations on the Citibeats platform and cross-referencing them with the citizen disinformation tips included in Maldita.es's Disinformation Management System (DMS), DEWARD, will provide DMS users with warnings about "emerging misinformative content or social conversations that have the potential to produce disinformation." Meanwhile, a disinformation tagging tool will be integrated into the Citibeats platform "to visibilise misinformation use in social discussions, and facilitate investigation of disinformation's role in broader public debates."[161]

The International Center for Journalists (ICJ) is developing an online violence early warning system, in partnership with computer scientists from the University of Sheffield. The project is "designed to identify key indicators and metrics signaling escalation of online violence against women journalists. They are studying the two-way trajectory between online and offline attacks and developing open-source digital tools to detect, monitor and alert key responders to high-risk cases". ICJ explicitly recognises that online violence often operates at the intersections of multiple forms of discrimination and disinformation.[162]

Awareness raising, media literacy and capacity building

The Foundation for Media Alternatives in the Philippines worked with illustrator Mariam Hukom to develop a comic strip on gendered disinformation, as part of a broader series on disinformation, media literacy and the importance of critical thinking.[163]

In 2022, HER Internet "implemented a project to create awareness and understanding of gendered disinformation including its effects and perpetrators in Uganda." The project focused on sexual minorities and sex workers. Kyogabirwe explains:

[161] https://mediafutures.eu/projects/disinformation-early-warning-data-tool-deward/

[162] https://www.icfj.org/our-work/online-violence-early-warning-system

[163] https://fma.ph/marian-hukom

> HER Internet convened an interactive dialogue in Uganda's capital Kampala to share real life experiences as well as strategies on how to avert the negative effects of gendered disinformation. Targeting 20 individuals from communities of structurally marginalised women, the dialogue also covered aspects of fact-checking and safety online. [...] The dialogue called for non-discriminatory enforcement of current cyber laws and the need for diverse narratives to eliminate biased reporting, amongst other measures. In addition to the dialogue, Her Internet also conducted a campaign on its social media platforms on the key concepts of gendered disinformation, its manifestations and counter strategies. The project also compiled and disseminated a handbook on understanding gendered disinformation as a go-to guide for communities to understand and further engage beyond the campaign and dialogues.[164]

Building on the findings of its social media monitoring efforts (see the relevant section above), PSA raised awareness about online gendered disinformation in Nepal in media literacy workshops organised with representatives from the media, civil society and political parties around the country.[165] The workshops were aimed at building the capacity of local journalists, civil society members, youth and women politicians to counter gendered disinformation: "participants would learn about gendered violence in the media, how mis-and disinformation spreads and ways to tackle it."[166]

As part of the project, PSA also produced a handbook titled How to Identify and Counter Online Gendered Disinformation (the handbook itself does not seem to be available online, though excerpts have been shared on the South Asia Check Twitter account).[167]

In 2022, the US-based advocacy group UltraViolet released a guide titled Reporting in an Era of Disinformation: Fairness Media Guide for Covering Women and People of Color without Bias. This guide is intended to "help journalists, reporters, and social media platforms identify and avoid unintentional sexist and racist bias or disinformation when interviewing, writing about, or moderating content about women and people of color, particularly Black women."[168]

On the issue of violence against women in elections in particular, a number of bigger public awareness campaigns have been organised around the world, including the global #NotTheCost campaign[169] and the #BetterThanThis campaign in Kenya.[170]

[164] Kyogabirwe, L. (2023, 6 February). Op. cit.

[165] South Asia Check. (2022, 9 June). Panos media monitoring initiative looks at trends in gendered online violence. https://southasiacheck.org/in-public-interest/panos-media-monitoring-initiative-looks-at-trends-in-gendered-online-violence

[166] https://twitter.com/SouthAsiaCheck/status/1516678518641360896

[167] https://twitter.com/SouthAsiaCheck/status/1564513713096634368

[168] https://weareultraviolet.org/fairness-guide

[169] https://www.ndi.org/not-the-cost

[170] International Foundation for Electoral Systems. (2017, 19 July). Kenyans Say "We are #BetterThanThis," Aiming to Support Women's Participation in Elections. https://www.ifes.org/news/kenyans-say-we-are-betterthanthis-aiming-support-womens-participation-elections

CONCLUSION

Falsehoods
?!?!?!
Hoax
GENDER
#
Myth
Fake
accounts
Propaganda
Fake

It is imperative to make digital spaces safe for women and gender non-conforming individuals. The interdependence of human rights requires that there can be no trade-off between people's right to be free from violence and the right to freedom of opinion and expression. Preserving that freedom while also protecting women and gender non-conforming individuals from violence and hate requires a three-fold approach: firstly, the use of a gendered lens when interpreting the right to freedom of expression and opinion; secondly, a calibrated approach that ensures that responses to violations are aligned with the level of harm or threat; and thirdly, a clear recognition of technology-facilitated gender-based violence and its impact on individual women, gender non-conforming people and society at large.

Disinformation is a complex, multifaceted phenomenon with serious consequences. Sitting at the intersection between disinformation and TFGBV, it destroys people's trust in democratic institutions and democracy itself, and promotes suspicion in relation to gender issues and gender activists. It thrives where public information regimes are weak, independent investigative journalism is constrained and women's rights are disregarded. It disempowers individuals, robbing them of their autonomy to search, receive and share information and form opinions.

The onus of responding to and preventing gendered disinformation should not fall on the shoulders of the targets of gendered digital attacks, nor on those targeted or manipulated as consumers of false or problematic content. States must regulate the operation of digital platforms through smart regulations that focus on safety and privacy by design, due diligence and transparency. Companies must review their business models, halting practices based on the exploitation of private data and "attention economics", placing the safety of women and gender-diverse individuals before profit.

WHEN PROTECTION BECOMES AN EXCUSE FOR
CRIMINALISATION: GENDER CONSIDERATIONS
ON CYBERCRIME FRAMEWORKS