unesco

The Chilling:
Assessing Big Tech's Response to Online Violence Against Women Journalists

AUTHORS
Julie Posetti,
Kalina Bontcheva and
Nabeelah Shabbir

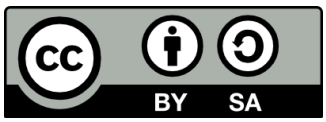
About this Publication

This is an extracted chapter of a wider UNESCO-commissioned global study on online violence against women journalists produced by the International Center for Journalists (ICFJ). The full-length study will published in 2022. The chapter identifies the role of big tech companies and especially social media platforms, as vectors and facilitators of gender-based online violence targeting women journalists. And it assesses the responses of these companies to the problem, making 23 recommendations for more effective countermeasures.

EDITORS: Julie Posetti and Nabeelah Shabbir

CONTENT WARNING: This document includes graphic content that illustrates the severity of online violence against women journalists, including references to sexual violence and gendered profanities. This content is not included gratuitously. It is essential to enable the analysis of the types, methods and patterns of online violence.

DISCLAIMER: The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views and opinions expressed in this document are those of the authors and should not be attributed to UNESCO.



CI-2022/FEJ/PI/1B

ICFJ ONLINE VIOLENCE PROJECT LEADERSHIP TEAM

LEAD RESEARCHER/AUTHOR: Dr. Julie Posetti. Global Director of Research, International Center for Journalists (ICFJ); Senior Researcher, Centre for Freedom of the Media (CFOM), University of Sheffield; Research Associate, Reuters Institute for the Study of Journalism (RISJ), University of Oxford

SENIOR RESEARCHERS: Prof. Kalina Bontcheva (CFOM); Prof. Jackie Harrison (CFOM); Dr. Diana Maynard (CFOM); Nabeelah Shabbir (ICFJ); Dr. Sara Torsner, University of Sheffield (CFOM)

RESEARCH ASSOCIATE: Nermine Aboulez, ICFJ researcher and University of Oregon PhD candidate

REGIONAL RESEARCH TEAMS

AFRICA: Assoc. Prof. Glenda Daniels (Regional Lead); Fiona Chawana; Dr. Omega Douglas; Dr. Julie Posetti; Nabeelah Shabbir; Alexandra Willis

ARAB STATES: Nermine Aboulez (Regional Lead); Dr. Julie Posetti; Nabeelah Shabbir;

ASIA AND THE PACIFIC: Assoc. Prof. Fiona Martin (Regional Lead); Liana Barcia; Dr. Ayesha Jehangir; Nirasha Piyawadani; Dr. Julie Posetti; Dr. Jenna Price

CENTRAL AND EASTERN EUROPE: Dr. Greta Gober (Regional Lead); Jen Adams; Bojana Kostić; Nabeelah Shabbir

EUROPE AND NORTH AMERICA: Dr. Julie Posetti (Regional Lead); Dr. Greta Gober; Prof. Jackie Harrison; Nabeelah Shabbir; Dr. Sara Torsner; Prof. Silvio Waisbord

LATIN AMERICA AND THE CARIBBEAN: Dr. Luisa Ortiz Pérez and Dr. Yennue Zárate Valderrama (Regional Leads); Dr. Kate Kingsford; Carolina Oms; Dr. Julie Posetti; Nabeelah Shabbir; Kennia Velázquez; and Prof. Silvio Waisbord

SPECIALIST RESEARCHERS: Becky Gardiner and Angelique Lu

UNESCO EDITORIAL COORDINATION: Saorla McCabe and Theresa Chorbacher.

PROJECT SUPPORT: Jen Adams; Fatima Bahja; Heloise Hakimi Le Grand; Mark A. Greenwood; Mora Devi S. Hav; Senka Korać; Juan Mayorga; Eunice Remondini; Erin Stock; Joanna Wright; Mengyang Zheng (ICFJ/CFOM); Johann Bihr; Sara Bonyadi; Annina Claesson; Lou Palin; Dana Muresan; Antonia Eser-Ruperti (UNESCO).

ICFJ PROJECT PARTNERS: Centre for Freedom of the Media (CFOM), University of Sheffield; Dart Asia Pacific; Ethical Journalism Network (EJN); International Association of Women in Radio and Television (IAWRT). This project has received financial support from UNESCO's Multi-Donor Programme on Freedom of Expression and Safety of Journalists and the Swedish Postcode Foundation.

Published May 2022

Introduction

The role of internet communications companies in online attacks against women journalists cannot be underestimated. They operate in an era of digital journalism, networked disinformation, online conspiracy communities, and political actors weaponising social media and misogyny as tools to attack women journalists. Their claim that they are simply operating as passive 'platforms' for third party use distracts from their role as vectors and enablers of gendered online violence. Firstly, they have an obligation to provide services that are safe to use, and to act against users who perpetrate online violence against others. Secondly, these companies should address their content recommendation algorithms, which are aimed at maximising user engagement and serve to escalate abuse through the promotion of misogynistic content and groups engaged in online harassment and abuse (Spring, 2021).

For many women journalists around the world, Facebook (along with the company's other assets WhatsApp, Messenger and Instagram, which are now grouped under the new brand Meta), Twitter, YouTube and other services are essential tools for newsgathering, content distribution and audience engagement.[1] But the necessity to work in these spaces has resulted in a double bind: women journalists are heavily reliant on the very same services which are most likely to expose them to online violence. This tension is a feature of news organisations' dependent integration with big tech companies, a feature of what has been termed 'platform capture' (Posetti, Simon, and Shabbir, 2019), and it has been exacerbated by the COVID-19 crisis, which has made journalists even more reliant upon these technologies. This development may help explain why many journalists, including those interviewed for this study, said they had experienced "much worse" online violence in the context of the pandemic (Posetti, Bell and Brown, 2020).

The failure of these companies to ensure safe environments for many users is widely recognised. For women journalists, this curbs their ability to research stories, share journalism, and engage safely with audiences. But it also reinforces a climate of impunity for crimes against them - online and offline. For example, Al Jazeera's principal Arabic presenter Ghada Oueiss raised concerns with the researchers about threats to her life which have been made with impunity on social media platforms. One person posted on Facebook that he would give US$50,000 to anyone who would kidnap or kill her. Oueiss called the police and the perpetrator was arrested, however the menacing threat remained on Facebook, she said, increasing the physical danger she faced.

Nick Pickles, Twitter's US-based Director of Policy, Strategy, Development & Partnerships, acknowledged that more needs to be done to deal proactively with serious threats against women journalists on the platform, including those emanating from organised crime figures and cartels. However, a sense of impunity is emboldened by the platforms' failure to take action against the content and perpetrators involved in gendered online violence.

Almost without exception, the women journalists interviewed for this study complained about the companies' unresponsiveness, inaction, ineffective action and convoluted and cumbersome processes for reporting and escalating incidents. Some said that all this compounded the effects of abuse they endured on the platforms themselves. Getting the companies to deal with perpetrators of online violence against women journalists is "like trying to talk to God...pulling out a tooth from a child is easier", Catherine Gicheru from the

1 Journalists interviewed shared their experiences of online harassment on other apps including Telegram, Clubhouse, Google Ads/Google Voice/Google Play and Discord servers amongst others. See more in the detailed country case studies for US, UK and Serbia which will be published separately by ICFJ.

African Women Journalism Project said. Additionally, the research participants were highly critical of the companies' perceived failure to recognise and adequately respond to the role of misogyny in attacks against women journalists on their platforms, especially at the intersection of racism, religious bigotry, homophobia and transphobia. Some interviewees described these US-based companies' incapacity to deal with diverse cultures and linguistic variations as particularly problematic. South African journalist and editor Ferial Haffajee said: "They treat their users in Africa like a colonial outpost".

Arbitrary features of content moderation, opacity of processes and responses, and corporate resistance to scrutiny and accountability for their role in violence against women journalists, were also heavily criticised by the research participants. This underscores the "distinct lack of clarity about what platforms are currently doing to combat abuse" (Dragiewicz et al., 2018), which represents a major impediment to assessing the relative effectiveness of their responses to the problem (Suzor, van Geelen and Myers West, 2019).

Another key dimension that needs to be taken into account is the quickly evolving nature of online abuse tactics. As this study confirms, gender-based online violence against women journalists now occurs at the nexus of viral disinformation, digital misogyny, online conspiracy communities, and political populism and extremism, and it is also increasingly cross-platform. This demands sophisticated and collaborative responses to the problem. Ellen Tordesillas, President of Vera Files and columnist for ABS-CBN News in the Philippines, said: "When [the platforms] come up with preventive measures, it will not take long for [abusers] to come up with another way to circumvent. It's a continuous battle." For example, key among this study's findings is the shift by abusers to more subtle, less easily detectable, and less actionable forms of abuse, which nonetheless can be cumulatively devastating.

This chapter draws on over 714 women-identifying survey respondents, 15 country case studies (Kenya, Nigeria, South Africa, Pakistan, The Philippines, Sri Lanka, Lebanon, Tunisia, Poland, Serbia, Brazil, Mexico, the UK, the US, and Sweden) produced by the regional research teams attached to this study, and 182 long form interviews with journalists, editors, digital safety practitioners and experts conducted between July 2020 & October 2021.[2] The chapter analyses the UNESCO-ICFJ survey responses and interview data collected as it pertains to women journalists' experiences of working on, and dealing with, the internet companies in the context of online violence. Research interviews were sought with senior representatives from Twitter, Facebook and Google to discuss the measures they were taking to address gender-based online violence. Twitter's Nick Pickles agreed, and he is therefore quoted in this chapter. Facebook declined the request, and Google said it could not identify an appropriate interviewee in time to meet the deadline set. In August 2021, UNESCO received a formal response from Facebook to a discussion paper previewing the study[3], in which company representatives said they recognised the importance of the recommendations formulated in the paper.[4]

1. Not safe for work

The top five platforms or apps most frequently used for work by the 714 women journalists who responded to the survey were ranked as follows: Facebook (77%; n=550); Twitter (74%; n=528); WhatsApp[5] (57%; n=407); YouTube (49%; n=350); and Instagram[6] (46%; n=328). Although Twitter was used almost as heavily as Facebook by the respondents in the course of their work, Facebook was

2 These interviews were conducted by a number of international research team members identified on the inside cover, including the authors of this chapter.
3 The Chilling: Global trends in online violence against women journalists Research Discussion Paper elicited 26 preliminary findings on pages 12-18 https://en.unesco.org/sites/default/files/the-chilling.pdf
4 Internal Facebook communication to UNESCO received August 19, 2021.
5 WhatsApp is owned by Facebook/Meta
6 Instagram is owned by Facebook/Meta

disproportionately identified as the service to which respondents most frequently reported online attacks (39%; n=279), with Twitter attracting complaints at the rate of 26% (n=186). Sixteen percent (n=114) had reported instances of online violence to Instagram, while 50 women respondents had referred complaints to YouTube, and 43 to WhatsApp.

Interviewees in all 15 countries studied conveyed the same sense of futility as the survey respondents when it came to reporting online violence to the companies. "I feel nothing will be done, so I don't bother [reporting incidents to the platforms] anymore," Nigerian journalist Kiki Mordi said, echoing comments made by many other women interviewed. Fatigue and frustration is further illuminated by the survey data identifying levels of dissatisfaction among the women respondents who had reported online violence to the services which they use in the course of their work. 17% (n=122) of survey respondents said they were "very dissatisfied" by Facebook's response. That was almost twice the rate of respondents who said they were "very dissatisfied" with Twitter's response to incidents they had reported to that company. Instagram was ranked third in the dissatisfaction stakes, followed by YouTube and WhatsApp. Facebook was also identified as the least safe of the high-use platforms globally among women journalists surveyed, with 12% (n=86) rating it "very unsafe" - almost double the number who rated Twitter "very unsafe".

2. Inaction, inadequacy, and ineffectiveness

On International Women's Day in 2021, Director General of Sweden's Sveriges Radio Cilla Benkö wrote about the need to shift the onus for responding to gendered online violence to encompass an increased role for big techs: "[D]igital platforms need to take more responsibility for removing hatred and threats from their platforms. Threats targeting journalists as a group must also be taken more seriously. The level of measures in place and feedback have been far too little for far too long" (Benkö 2021).

The women journalist interviewees largely regarded reporting incidents to the companies to be an act of futility. This assessment is the result of frequently frustrated attempts to get the companies to flag or remove misogynistic, racist, threatening and libellous posts, comments, memes, pictures or videos. The most frequently reported response they said they received from the companies when they did report abuse was that the material reported was not in breach of corporate policies and therefore unable to be addressed. Some also described waiting weeks, or even months, for threatening and abusive content to be removed – and some assessed that this might only happen after a certain (unknown) threshold of complaints is received about accounts or posts. Many said that they had reported incidents which were never even acknowledged.

While attitudes to the companies varied among the 113 women journalists interviewed for this study, they were almost universally critical of Facebook, and expressed little faith in the company's announcements and initiatives regarding online harassment and abuse. Many were also scathing about Twitter, but several journalists referenced what they perceived to be recent improvements in Twitter's reporting tools and abuse minimisation efforts, saying that they now feel "safer" or "less exposed" on the platform.

Overall, the dissatisfaction reported by the interviewees covers eight main areas of concern:

- Inadequate and cumbersome abuse reporting processes.
- The absence of human-centred points of contact and response mechanisms.
- Poor, unidentifiable or inaccessible processes for escalation.

- Unresponsiveness - including non-responsiveness, and poor and inconsistent responses - to incidents reported.

- Concerns about an absence of gender-sensitivity and a lack of awareness about intersectional threats.

- Failure to develop moderation capabilities suitable for linguistically and culturally diverse communities of users.

- Failure to recognise the specific press freedom and journalism safety risks entailed, and respond appropriately on the basis of recognition that freedom from online violence and freedom of expression are not mutually exclusive.

- A lack of transparency and accountability demonstrated in policies and official responses to the problem of gendered online violence on their services.

These identified concerns are elaborated upon below.

Shortcomings of abuse reporting tools and processes

The interviewees and survey respondents shared a common sense of frustration that they were left to block or mute abusive users themselves due to the failure of automated platform-based systems. Many also expressed concern that repeat perpetrators were often able to act with impunity. Further, when incident reports did elicit a response (usually after a significant time lag), the journalists said that their requests for flagging, muting or deleting offensive content or accounts were most often rejected.

Al Jazeera's Ghada Oueiss, who is the target of coordinated cross-platform disinformation campaigns allegedly involving State actors, said she had "lost count" of reports she had made to Facebook, Twitter, YouTube and Google. She described YouTube and Google Search as the sites of some of the worst abuse she has experienced. "You can never know who Ghada Oueiss is for her journalism. You only see attacks, attacks, attacks…you would think that either I'm a terrorist, or I'm a whore," she said of the smears prevalent on the sites. Oueiss also said Twitter was very slow to deal with tens of thousands of tweets sharing stolen and altered pictures of her that were part of a coordinated smear campaign.

Brazilian journalist Patricia Campos Mello said the companies' response to the online abuse triggered among their supporters was meagre. When she reported doctored images of her in 2018, she found Twitter to be "more agile" while Facebook "just ignored it". Campos Mello said she has virtually given up on using standard platform-based reporting systems.

In order to force action, Swedish magazine editor Susanna Skarrie enlisted help from an external consultant to liaise with Google and Facebook about removal of abusive content and the need to reduce traffic to websites targeting her, her family, and her colleagues. She said Google promised to stop search engine optimisation for websites publishing false information about her and her colleagues. However, this did not halt the websites coming back online: "Every time a new subpage emerges, we have to contact Google again," she said. Facebook told Skarrie that the abusive accounts she reported were "not illegal", but it did remove some. However, the accounts kept regenerating, "and are used to slander me and my family and other journalists who have investigated them," Skarrie said.

Doxxing and other digital security breaches which expose women journalists to increased offline threats are also rarely dealt with swiftly enough by the platforms. For example, when Serbian journalist Jovana Gligorijević of Vreme said

she was doxxed in YouTube comments in 2019 (Strika, 2019), she reported that her personal information was only taken down by the Google-owned company after the breach was reported more than 30 times.

Many interviewees also expressed concern about the companies only taking action after the prospect of immediate physical harm had become apparent. Facebook's standard for 'credible' violence, for example, requires language that incites or facilitates serious violence be judged "a genuine threat of physical harm" before moderators will act to remove content - although its requirements for assessing the seriousness of threats are not apparent (Facebook, 2020). One example was provided by Guardian US investigative journalist Julia Carrie Wong regarding the 'Unite the Right' rally in Charlottesville, Virginia, which started as a Facebook event and resulted in the death of a woman. A few weeks before the event, Wong said she sent Facebook a spreadsheet with links to 175 neo-Nazi, white nationalist and neo-Confederate hate groups that were using the service to recruit and organise. She said Facebook had declined to take any action against the vast majority of them until after the woman, Heather Heyer, was killed during the rally. She said this chain of events raises questions about Facebook's role in facilitating and amplifying hate.

Various studies have made complementary points. An assessment of barriers to more effective responses from the internet companies by PEN America criticised Facebook's "byzantine" settings for allowing users to make profile or cover photos private, unlike the 'one-click' systems Instagram and Twitter have (Vilk, Vialle and Bailey, 2021). This functionality could help mitigate the problem of stolen images of women journalists being used by imposter accounts, or in the production of deepfakes. PEN America has also highlighted that purely reactive measures such as blocking and muting can mitigate online abuse once it is underway, but do not proactively shield targets. These points resonate with the argument that the companies need to "centre the voices of those who are directly impacted by the outcome of the design process" (Costanza-Chock, 2020).

Many interviewees and survey respondents said that the companies need to create rapid response units - staffed by multilingual employees with expertise in both press freedom and gender-based violence. They also wanted to be able to hand over their accounts to trusted colleagues for monitoring in serious cases. A series of practical product design solutions proposed by PEN America includes a privacy-preserving facility to allow trusted contacts to assume control of a journalist's accounts when she is under attack. They also recommended an 'SOS button' for those facing severe threats or large-scale pile-ons, which would allow them to access a support hotline and instantaneous in-platform protections (Vilk et al., 2021). The World Wide Web Foundation's Tech Policy Design Lab: Online Gender-Based Violence and Abuse report recommended a similar measure:

> **Users could assign different roles to trusted contacts, giving them authority to upload or delete content, or restrict and delete comments. Posts uploaded by trusted contacts could have the option to be marked with a verified 'trusted contact' badge....[However, it is vital] that companies do not shift too much responsibility onto people who volunteer to help others manage abuse (Dhrodia et al., 2021).**

There was evidence from this study's research subjects that Twitter monitors some highly vulnerable journalists and intervenes technically to choke online abuse in real-time, sometimes making direct human contact to warn a target of impending threats. Julia Carrie Wong said: "I actually got alerted by a Twitter staffer who told me that [they] had seen some kind of threats or discussion...[So] I started to lock down quite seriously on my digital footprint to try to minimise things, and started using a Delete Me[7] service."

Former *The New York Times'* journalist Taylor Lorenz[8] pointed to recent technical improvements on Instagram, such as comment filters and action against imposter accounts, which, while not being nearly enough to address the problem of violence against women journalists on the platform, are "a very good step in the right direction," she said. Lorenz has also tried contacting the companies' PR representatives via Direct Messages on Twitter, TikTok, Clubhouse, Substack and Instagram as a work-around to deal with the companies' non-responsiveness – although she said that they usually respond that "there's nothing they can do".

The need for human points of contact and proactive detection

The concerns expressed above underpinned the widespread call among the interviewees and survey respondents for the companies to employ many more human moderators and policy specialists with training in human rights, particularly in the areas of gender equality and press freedom, and in countries where the offline risks associated with online violence are most severe.

In the US, Al Jazeera's White House Correspondent Kimberly Halkett, who was doxxed and received death threats in 2020 across multiple platforms, articulated the research participants' shared demands for change. She said special response units which prioritise human contact are needed to deal with online violence complaints involving women journalists:

> When you've been a target of sexual violence or any sort of gender-based violence of that magnitude, it needs a different phone number. It needs a different email address and it needs a quick reply. [But] these platforms make it very challenging to reach a human voice, and they do that deliberately. This insulates them from having to deal with the hassles of people like me.

Halkett suggested responses to address these deficits, such as a dedicated email address and phone number with human contact points, along with a dedicated ombudsperson for special categories of users (e.g. journalists and human rights defenders) to ensure appropriate action in cases of gender-based violence.

7 At least a dozen interviewees based in the US or working for US organisations said their newsrooms had paid for them to use Delete Me, or they paid for it themselves. Founded by US company Abine in 2011, this tool removes personal data from websites, or 'scrubs the internet' of an individual's personal information.
8 Lorenz moved to *The Washington Post* in early 2022. Her interview was conducted when she was still at *The New York Times.*

Overall, this research revealed the need for social media companies to respond proactively and pre-emptively to acts of online violence against women journalists. Instead of placing the onus for managing the abuse and harassment on those women targeted, many interviewees said that the companies needed to work harder to prevent such attacks at their point of origin, not wait for them to make complaints on a scale deemed necessary to trigger action.

When interviewed in June 2021, Twitter's Nick Pickles acknowledged that a shift was required by the platforms in responding to online violence incidents to avoid further burdening those targeted. He said the company was going in the right direction: "We definitely hear the feedback that the burden is way too much on victims. And that's something that we're working to change now in real time... Now we're at a point where more than half of all the content we removed is detected proactively by us".

Company policy and human rights

While a State cannot legitimately mute a citizen permanently without disproportionate violation of the right to expression, a corporation can remove a user from its service. Such an individual can use, or set up, alternative channels to impart their opinions. The companies have no legal constraint to tolerate users who are routinely abusing the terms of use, injuring other citizens' rights in the process, and threatening democracy (Posetti and Bontcheva, 2021). By de-platforming abusers (based on due process of consumer rights, such as providing tiered warnings where appropriate, and appeal options), a company can end the prevalent impunity on its service, putting an end to a situation where online violence can continue to be committed without consequences for the perpetrators within this space. This depends in part on how the company interprets respect for human rights in general and freedom of expression in particular. Within several US-based internet communications companies, there is a framed tension between allowing "free speech" (seen basically as a right to unfettered speech), and protecting other rights. But freedom of expression rights as enshrined in international human rights law do not uphold the right of a person to use online violence to limit the speech (or personal safety) of another person, especially not journalists - whose public interest service merits particular protection.

Even in the US, while the authorities are constitutionally restrained from restricting much speech, private sector entities are free to adopt restrictions that reflect the limits of what they permit in their realm. In addition, due to the limited liability offered them by the Communications Decency Act, they only need worry about legal consequences of carrying third-party speech that crosses the threshold of illegality, if and when such content is drawn to their attention and they take no corresponding action.

In this context, the companies have routinely deflected demands for action against much online violence on the basis that this is simply part of legitimate 'free speech'. This has led to numerous women journalists suffering the violation of their own freedom of expression - both as citizens and professionals. However, according to Brandy Zadrozny: "freedom to does not overtake freedom from [during] a harassment campaign. I say to the tech platforms 'why does this person's rights usurp this user's?'". This is a point echoed by PEN America's Viktorya Vilk:

Women journalists, journalists of colour, LGBTQ journalists are getting forced out of public discourse, which is increasingly taking place online, and sometimes getting forced out of their professions altogether. And so this whole idea that we can't do anything about online abuse, because if we do, we will damage people's free expression rights is wrong, because the online abuse itself is actually what's limiting the free expression rights of so many folks who are marginalised…already. Now they're getting marginalised online and in their professions to boot.

This perspective underlines the recommendation by the former Special Rapporteur for Freedom of Expression and Opinion, David Kaye, that companies align their conceptualisations of freedom of expression with international human rights laws and norms. This situation would then recognise that protecting against online violence is a legitimate restraint on speech. However, it would require much more investment by the companies if they were to accept and to follow such a voluntary commitment in practice.

There is a UN-level human rights framework for corporations that should guide their conduct. The UN-commissioned Ruggie principles (OHCHR, 2011c) are designed to prevent corporations from undermining human rights. Called the "UN Guiding Principles on Business and Human Rights" (OHCHR, 2011b), these require corporations to "avoid infringing on the human rights of others and... address adverse human rights impacts with which they are involved", while "taking appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication". Meanwhile, the Rabat Plan of Action (OHCHR, 2012) is a UN operational framework that can serve companies seeking to balance freedom of expression rights against the need to curtail incitement to hatred, violence, hostility and discrimination. To date, there is no evidence of the companies agreeing to be held accountable in terms of commitments in this area.

Other initiatives include the B-Tech Project, facilitated by the UN's Office of the High Commissioner for Human Rights, which is designed to synthesise guidelines and tools for practical application to human rights questions in connection with business and technology (OHCHR, 2019b). The UNESCO-led UN Plan for the Safety of Journalists is a particularly relevant instrument (UNESCO, 2012) which social media companies could factor into their efforts to respond to gendered online violence against journalists. Further work by UNESCO seeks to promote transparency by the internet companies as a means towards accountability for respect for human rights, and this will be followed in 2022 by specific focus on access to data concerning journalistic safety issues.

The 250 recommendations of the Information and Democracy Forum's Working Group on Infodemics[9] for structural reform to improve the platforms' governance,

9 Lead author Julie Posetti was a member of the Working Group.

transparency and accountability could apply to the problem of online violence. Key among them: "Platforms should follow a set of Human Rights Principles for Content Moderation based on international human rights law: legality, necessity and proportionality, legitimacy, equality and non discrimination" (Forum on Information and Democracy, 2020). However, Brazilian journalist Patricia Campos Mello drew an effective comparison between the distinctions in the policy regarding disinformation and gendered online violence, whereas substantially more progress is being made with regard to dealing with disinformation in the context of the pandemic compared to gender-based online violence.

Additionally, there has been effective civil society work around content moderation standards such as the Santa Clara Principles of 2018, developed by a collective of human rights organisations, advocates, and academics (Electronic Frontier Foundation, 2021). The principles were updated in 2021 with an emphasis on respect for human rights frameworks and human oversight (ibid.). Facebook published its first Corporate Human Rights Policy in March 2021 - 17 years after the company was founded (Sissons, 2021; Facebook, 2021a). It makes an explicit commitment to the safety of journalists by promising to protect "professional and citizen journalists" (under the umbrella of 'human rights defenders') from online attacks. Twitter was working on a similar policy in late-2021.

In July 2021, Facebook, Google, Twitter and TikTok signed a World Wide Web Foundation (WWWF) pledge to tackle gender-based online violence (WWWF, 2021). The companies committed via the pledge to build better ways for women to curate their safety online by:

- Offering more granular settings (e.g. who can see, share, comment or reply to posts);
- Using more simple and accessible language throughout the user experience;
- Providing easy navigation and access to safety tools;
- Reducing the burden on women by proactively reducing the amount of abuse they see.

They also committed to implement improvements to reporting systems by:

- Offering users the ability to track and manage their reports;
- Enabling greater capacity to address context and/or language;
- Providing more policy and product guidance when reporting abuse;
- Establishing additional ways for women to access help and support during the reporting process.

NBC News-MSNBC's Brandy Zadrozny has argued that a more radical transformation is required, with a need to de-platform propagators of online violence at first strike in serious cases in order to combat recidivism: "If [the platforms] were to adequately enforce their own policies against harassment, they'd lose half their users… The companies need to do a lot of soul searching and then come away with a vigorous plan. Their commitment to freedom from harassment needs to be larger than their commitment to say whatever the hell you want on the internet."

There is also the issue of whether companies will address recommendation algorithms for content, users and groups, which have been found to exacerbate the problem in some cases and to promote misogyny (see 'Gap 5' below for details).

Inconsistent application of standards

Even though there are the above-mentioned human rights standards and principles, and even though the companies have policies for dealing with online abuse (e.g. Twitter Help Center, 2021a), when women journalists report online attacks, these companies also often fail to adequately enforce their own rules (see also: Amnesty International, 2018).

The unevenness of the social media companies' policy implementation and enforcement across different countries and languages worldwide is another significant challenge. One example is the investigation by the Balkan Investigative Reporting Network (BIRN) into how Facebook and Twitter deal with content violations in the region. This established that half of the posts they reported as hate speech, threatening violence or harassment remained online - even though they were in clear violation of platform policies (Jeremić & Stojanovic, 2021).

Another issue is response times. In Germany, where the Network Enforcement Act sets time limits for the removal of hate speech, the companies react much faster, and more decisively. Penalties for failing to act in a timely manner include steep fines and the companies are required to report on the local personnel they employ, and the actions they take to remove hate-related content they assess to be evidently unlawful.

A further issue is double standards within community guidelines. For example, the deputy editor-in-chief of a women-oriented magazine in Poland, Monika Tutak, expressed frustration with the way her publication's content is censored by social media companies due to breaches of policies regarding nudity, while gender-based hate speech against her staff is not deemed to meet the requirements for removal. Tutak referred to the symptomatic 'ban of the nipple' that Facebook adheres to, resulting in their content being deemed inappropriate and removed or 'shadow banned' by Facebook on a number of occasions:

> We don't trust this company...I very often report to them hate speech and they don't react. I am afraid of the opposite situation, when our journalists are being blocked by Facebook. For example, we had a topic about period poverty, and there was an illustration of a woman with a stained skirt and Facebook blocked our journalist for 24 hours for this story... Facebook sometimes cuts the reach [of our posts] and it mostly applies to content about the female body.

Facebook has been accused of censoring the accounts of journalists in a number of other countries, and de-platforming them without providing justification. In Tunisia, 60 journalists and activists had their Facebook accounts deleted without warning or explanation in 2020. Anti-corruption watchdog Iwatch managed to get 14 deleted accounts reactivated after lodging complaints (Cordall, 2020). This pattern also highlights the double-edged sword associated with

blunt content moderation policies, and the need to balance responses to gendered online violence against journalists with broader freedom of expression considerations.

Content removal is often requested in response to online violence cases, yet content removal can also work against women journalists unless there is oversight by teams with freedom of expression expertise and local contextual knowledge. Several interviewees criticised the social media companies for frequently failing in regard to balancing the need to protect women journalists against the need to respect freedom of expression. They pointed to arbitrary censorship, a lack of transparency and 'shadow banning'.[10]

The extent of transparency is also at stake. According to the transparency reports of Twitter, Facebook and Google, between 2017 and 2020 the Mexican authorities made more than 38,659 thousand requests for the removal of content. In 95% of these cases, no information exists about the nature of the content, and there is no accountability associated with its removal (ARTICLE 19, 2021b). According to ARTICLE 19, however, only 6% of the requests for content removal to Google made by public officials in Mexico were granted.

Available tools

There have been many policy announcements from the platforms regarding online abuse and harassment. Most interviewees dismissed these efforts as 'PR exercises', while also welcoming a number of platform initiatives. For example, Twitter's 2021 rollout of features such as allowing users to limit the ability of non-followers to reply to tweets and the option to remove followers was welcomed by a number of journalists, who said they now feel "safer" on the platform.

Figure 1: After being targeted, BBC Investigations reporter Rianna Croxford adjusted her Twitter settings so as not to be notified about all tags. "My [Twitter] notifications are now set in a way now that...if people comment and I don't know them, or I don't follow them, then I don't see it." [11]

Another Twitter feature that the interviewees found helpful in reporting and documenting online abuse is the ability to attach multiple tweets to a single report. This allows users to flag additional context and makes it easier and faster to provide proof that a particular account is being used in abusive ways, instead of requiring those targeted to submit a list of the attacks as they occur (Tang, 2016). Taylor Lorenz, formerly at the *The New York Times,* described this approach as extremely important for tracking and tracing online abuse in real-time under extreme stress: "If we don't have the screenshotted receipts, it's like it never happened". NBC News-MSNBC reporter Brandy Zadrozny said she documented her own online abuse on a spreadsheet, specifically with regard to Google Voice messages, recurring email addresses, or phone numbers calling to abuse her.

10 Poland's Panopticon Foundation perceives shadow banning as a form of censorship where either users or the reach of their content can be blocked by social networks in a way that they are unaware. A user who is officially blocked or removed, at least theoretically can appeal the decision, while shadow banning is arbitrary and there is no question of transparency here (Obem and Głowacka, 2019a).
11 Rianna Croxford tweet from June 2021: https://twitter.com/The_Crox/status/1404491531843670017; interviewed 03.03.21.

There are several third-party documentation facilities specifically designed for recording abuse in various stages of development. These include JSafe[12] and DocuSAFE[13]. Both of these apps still require users to manually track and enter data, but they offer a single place to store and organise it. Google's Jigsaw is also experimenting with documentation and reporting tools (including a 'harassment manager' which was still in Beta development in December 2021), leveraging their machine learning system Perspective API. This system, used by over 200 partners including *The New York Times*, detects toxic language to help targets of online violence take action in a more streamlined manner (RE•WORK 2021, Jigsaw 2021). Tune[14] is another Jigsaw tool developed to address online toxicity, while Jumbo[15], Sentropy Protect[16], Tall Poppy[17] and BodyGuard[18] are other offerings that help users change their settings on social media platforms. PEN America has recommended an abuse documentation facility that captures and aggregates screenshots, hyperlinks, and other publicly available data "automatically or with one click" (Vilk et al., 2021).

Such tools are important to aid police investigations and legal action against perpetrators, but the need for them highlights the difficulty of navigating privacy and security within the platforms themselves. To be really effective, tools like this need to have in-built facilities to instantly submit abuse reports directly to the platforms for action and escalation. To date, the companies have resisted such recommendations on the basis that they do not have capacity to deal with alerts at scale.

Twitter's Nick Pickles says the company takes enforcement actions based on their hateful conduct policy, which was extended to prohibit language which dehumanises others on the basis of religious affiliation, caste, age, disability, disease, race, ethnicity, or national origin, sexual orientation, gender, gender identity, or serious disease in 2009 (Twitter, 2020). Tweet or account removal is carried out by human moderators. This follows either reporting from the abuse target (Twitter Help Center, 2021c), other platform users, or through proactive detection through machine learning tools that flag posts for moderator review due to their similarity to known violating content. Pickles said the company is also experimenting with 'nudges': "...to say to people 'this reply may well be seen as abusive. Are you sure you want to post that?' [But] machine learning is definitely not a perfect science. And so one of the things that we just have to be really careful with is, for example, catching counterspeech, we don't want to be catching people who are quoting people in these filters".

From a user perspective, some of the women journalists interviewed reported that they still appreciated the 'block' option on Twitter,[19] which - despite placing the onus on the journalist to deal with the abuse - can serve as a means to stem online harassment. Most recently, Twitter introduced a silent block option[20]. This means that products and policies that empower users to mitigate the impacts of abuse, while also relieving those under attack from the onus of responding could be optimally effective. Taylor Lorenz uses the proactive tool

12 JSafe is a mobile app in beta developed by the Reynolds Journalism Institute at the University of Missouri with the Coalition For Women in Journalism https://womeninjournalism.org/jsafe.
13 DocuSAFE is a free app created by the National Network to End Domestic Violence in the US: https://www.techsafety.org/docusafe.
14 Tune is a Chrome extension created in 2019 that employs machine learning to allow users to "control" the volume of the conversation they see, for example in"customising" toxicity in comments https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?hl=en
15 Jumbo is a third-party app using a subscription-based model allowing users to connect to various platforms and controlling their search history, messaging and other data https://techcrunch.com/2020/06/24/privacy-assistant-jumbo-raises-8-million-and-releases-major-update/
16 Sentropy uses natural language processing and machine learning to "protect users and their brands from abuse, harassment, and malicious content". It was bought by Discord, the online chat platform, in July 2021 https://techcrunch.com/2021/07/13/discord-buys-sentropy/
17 Tall Poppy is described by Canadian founder Leigh Honeywell as a 'digital public health nurse' - a platform helping individuals to take charge of their cybersecurity, also via incident response, and the service is sold to companies https://www.youtube.com/watch?v=xpcb2n6uvzE
18 BodyGuard is an ad-free, free to use mobile application for individuals using contextual, algorithmic analysis to detect and avoid toxicity online. It also offers businesses API services, including digital media companies https://www.lefigaro.fr/secteur/high-tech/haine-en-ligne-bodyguard-s-attaque-au-secteur-du-jeu-video-20211006
19 Details of account blocking on Twitter are here.
20 In October 2021, Twitter added this feature which offers people the option to remove followers without them noticing them or needing to block them.

Block Party[21] which automatically mutes people at scale, as well as MegaBlock[22] which blocks a person and every person that has 'liked' their tweet.

On Twitter, another effective feature which users have reported protects them from spam and abuse enables users to control who can reply to their tweets. There are also options to filter notifications from people not followed by the user. In March 2020, Twitter started a dedicated gender-based violence search prompt for hotlines and support in local languages in partnership with local NGOs, government agencies and UN Women. First launched in Mexico, the prompt later became available in 27 countries and 20 languages. In August 2021, a Safety Mode feature was released by Twitter which auto blocks accounts with potential harmful content. This has been followed by an experimental feature which flags heated conversations and prompts users to be 'respectful and truthful'.

The comment lock option on public Facebook profiles also gave some women journalists interviewed temporary respite, but the downside is that it also blocks audiences from contributing useful information or providing support. Similarly, Facebook Messenger's filter feature (that sends messages from non-'friends' to a separate folder) can stop journalists from receiving messages from genuine sources and contacts, with adverse implications for their journalism practice. Serbian journalist Jovana Gligorijević, who had to resort to filtering her Facebook account so that everything goes to spam unless from close friends, recognises that this excludes people asking her to investigate stories. "I miss the chance to help someone, but I have to do it to protect myself," she said.

A Facebook policy change from mid-2020 involves the ability for its users to register as journalists, to justify stronger security protections. The feature was initially only available to journalists in the United States, Mexico, Brazil and the Philippines (Facebook Journalism Project, 2020). In late February 2021 it was rolled out in 19 additional countries. Due to limited publicly available information about it, it is also not yet clear how effective this new feature is, or what the uptake rate has been. It was also not possible at the time of writing to determine if this system entails human contact points and rapid escalation of reports lodged by the registered journalists Most of those interviewed in the first four countries where it was introduced were not aware of the feature. Several who were aware nevertheless expressed scepticism, saying that they did not trust Facebook with the process of registration for a range of reasons, including data privacy concerns. However, a Filipino interviewee highlighted Facebook's takedown of pages for what the company calls Coordinated Inauthentic Behaviour (CIB) as a "good start".[23] A 'Journalists' Safety Guide' developed as part of the Facebook Journalism Project (Facebook for Media, 2021) was also noted by one interviewee as useful in providing tips on how to stay secure and report abusive behaviour on the platform.

Dealing with the large-scale problem of online abuse via anonymous and fake accounts is also an ongoing challenge. At one point, Al Jazeera's Ghada Oueiss said there were dozens of fake Facebook pages in her name which were used to shame and defame her or to spread fake news and disinformation. She said that Facebook advised her through an Al Jazeera colleague that the only way to counteract this would be to set up her own professional page - since she is a "public figure" - which they would then verify. Thus, they put the onus of verification on her, not the abusive imposters, stating that whenever a fake account posted something in her name, she could prove that "this is not me".

21 Block Party is an anti-harassment paid subscription service founded in 2021 by software engineer Tracy Chou, who was stalked in real life by an online follower. The service currently works on Twitter by directing users to a "lockout folder". https://www.fastcompany.com/90686948/inside-the-life-of-a-tech-activist-abuse-gaslighting-but-ultimately-optimism

22 MegaBlock is a third-party service which allows you to block a tweet but also the accounts of anyone who liked that tweet. It is created on a Discord server called Gen Z Mafia, a community of young tech workers operating collaboratively https://megablock.xyz/, https://www.nytimes.com/2020/09/15/style/gen-z-tech-mafia.html

23 Under its Community Standards Facebook defines Coordinated Inauthentic Behaviour as "the use of multiple Facebook or Instagram assets, working in concert to engage in Inauthentic Behavior, where the use of fake accounts is central to the operation" https://www.facebook.com/communitystandards/inauthentic_behavior

In mid-2021 Oueiss' verified Facebook page had 2.2 million followers,[24] and the page remained a constant target for abusers.

User-based pressure applied to the platforms

In Tunisia, some interviewees said they had resorted to legal action in order to force social media companies to act against the online violence they reported. This is because abusive content is usually not removed by the companies until the court orders it to be removed, according to Ayoub El-Ghadamsy, a lawyer with the National Syndicate of Tunisian Journalists. While a number of other women journalists interviewed for this study were considering legal action against the platforms after their corporate incident reporting mechanisms failed, there are occasional examples of companies removing content associated with targeted online violence against journalists, even when the courts fall short.

Jessikka Aro, a reporter at Finland's public broadcaster YLE, has been a target of organised 'troll' campaigns linked to a foreign State actor for her reporting on disinformation networks since 2014 (Aro, 2016; Ireton and Posetti, 2018). In 2020, Aro lost an application for a restraining order against two individuals whom she described as "far-right YouTube livestream harassers", and whom she accused of stalking her (Pohjola, 2019). However, in this case, YouTube closed the offending video channel, which carried a message saying: "This account has been closed due to repeated or severe violations of YouTube's policy prohibiting intimidation, harassment, and harassing content" (ibid.).

In the experience of Al Jazeera's Ghada Oueiss, the companies' responses also depend on the amount of attention the journalist or incident is getting, the number of followers a journalist has, and the extent of involvement of international organisations. Oueiss said pressure does help to evoke a quicker and more definitive response from the platforms in her experience. Solidarity exercises can also include mass-reporting of abusive accounts. Two Tunisian interviewees - Najoua Hammami, investigative journalist and director of the Media Office at the Arab Institute for Human Rights, and Khaoula Boukrim, editor-in-chief of Kashf Media, said they deployed mass reporting tactics when they were under attack as a defensive measure. They recruited their friends and colleagues to report abuse against them to the companies en masse as a way of escalating their complaints. Similar tactics were described by Kenyan journalists.

However, automated content removal has been weaponised against women journalists, among others. Interviewees from around the world described being affected by manipulation of the platforms' automated reporting systems. This occurs through mass false reporting of legitimate content as abusive in coordinated campaigns designed to automatically 'de-platform' (i.e. have their accounts suspended) the targets and censor their journalism through blocking or platform takedowns.

Nigerian journalist Kiki Mordi said she had seen abusive users with a large follower base encourage mass-reporting of legitimate accounts and content they dislike, such as that shared by feminists like herself. Mordi noted: "Just one or two accounts have been brought down for harassing me, as opposed to the hundreds everyday that continually harass me".And even when an abusive account is blocked or suspended by a platform, the same user can make a new account under a new pseudo-identity or use a Virtual Private Network (VPN) connection to prevent tracing.

24 Ghada Oueiss' official Facebook page: https://www.facebook.com/ghadaoueiss14/

3. Policy gaps

Six core company policy gaps pertaining to gender-based online violence were identified in the course of research for this chapter. They are elaborated below with a view to informing responses.

Gap 1: Lack of prioritisation of gender-sensitivity in policy and implementation

A key reform needed to more effectively address online violence against women journalists is the development of gender-sensitive policies that recognise the increased risks that women are exposed to on social media platforms, along with the exponentially worse offline impacts. These include, but are not limited to, threats of sexual violence but they also extend to sexist attacks that portray women as sexually immoral and/or highly sexualised beings, non-consensual sexual imagery, and sexually explicit content (e.g. graphic images of male genitalia sent via direct message) that are all used to harass women journalists.

The 15 detailed country-level case studies produced in tandem with this report consistently demonstrated the need for comprehensive gender-sensitive policies designed to ensure that women, and in particular women journalists, can work safely on the platforms. In Brazil and Poland, for example, the research highlighted content monitoring and interviewees' accounts that suggest that companies may censor feminist posts more than hate speech and gender-based attacks (see also Martins et al., 2020). Other research shows that Facebook in Sri Lanka in effect allows a culture of casual sexism and misogyny, expressed through sexual harassment and non-consensual dissemination of images, including intimate pictures and videos with derogatory, abusive and violent captions (Perera and Wickrematunge, 2019).

Existing policies do not appear to proactively cover instances of targeted online violence. Neither do the companies appear to deal with pernicious hashtags like #Presstitute which are used to discredit women journalists personally and professionally in tandem, while also exposing them to increased risk in some contexts. Former CNN editor Inga Thordar said there was a need for the platforms to make misogyny and intersectional abuse much higher priorities, and that platforms "should be taking much stronger action against people who are persistent perpetrators of online harassment".

In addition to gender-aware policy improvements, Pakistani journalist Benazir Shah said the companies also needed to be more responsive to women users. She echoed many calls from interviewees around the world for gender-sensitive and in-country contact points to be made available for women journalists because of their particular exposure to risk in the course of their work on the services. When Facebook opened an operation centre in Nigeria, the company worked with the civil society organisation Paradigm Initiative to promote their online safety features, although not specifically in relation to online violence against women journalists.

In countries like Mexico, where there are high femicide rates (SSPC, 2021; UN Women, 2017), and extreme risks faced by journalists in general (RSF, 2021k), substantive attention needs to be paid by the companies to combating gender-based online violence. For example, columnist and political scientist Denise Dresser, who has been a victim of orchestrated online violence including death threats, said that she perceives a lack of responsibility and inaction associated with the social media companies' responses which increases the offline risks she faces.

With regard to closed groups and/or encrypted communications, there is also a policy gap concerning gendered online violence. For example, in 2019, Facebook made a strategic shift, moving away from a focus on public sharing and instead promoting closed or semi-closed Facebook Groups (Ingram, 2019a). Such an approach presents a twofold problem. Firstly, although women journalists do indeed create closed online communities to support one another during attacks, this encourages them to retreat from visibility in order to protect themselves. It amounts to putting the onus on targeted people (i.e. women) to withdraw in response to abuse and attacks, rather than addressing the business model and design failures at the core of the problem. Executive Director of the US-based Representation Project, Soraya Chemaly, told the researchers: "What does it say if on your own platform women are hiding away in special groups because they cannot speak openly in their own space?" Secondly, it is evident that closed Facebook groups and Facebook Messenger, as well as WhatsApp, are also subject to less scrutiny (external and internal) of online violence and disinformation content (Ingram, 2019b). This is one reason these channels are targeted by perpetrators. Since many of the research participants described receiving the worst gender-based threats and harassment via such closed channels, this points to a priority policy gap that needs to be addressed.

Finally, policy measures and processes also need to be more technologically sophisticated and transparent to respond to gender-based hate speech that takes the form of synthetic media attacks, such as doctored images and deepfake videos.

A short case study in systemic policy and process failures

Facebook's dealings with Northern Ireland's Patricia Devlin - a reporter with the *Sunday World* newspaper until early 2022, highlight serious gaps in the company's approach to dealing with online violence against women journalists. Devlin has received multiple death threats from figures associated with neo-Nazis and paramilitary extremism. She also received threats of sexual violence against her baby via Facebook Messenger after she had been doxxed. "I received a message via my personal Facebook account, and it said 'Don't go near your granny's house in Maghera, Tricia, or you'll watch your newborn get raped'. And it was signed off in the name of a neo-Nazi group called Combat 18, which in the past has had links to loyalist paramilitaries in Northern Ireland." This threat, sent in 2019, continued to be investigated by the Police Service Northern Ireland (PSNI) in late 2021. But Devlin received a renewed threat against her child (again via Facebook Messenger) in mid-2021.

Figure 2: The threat sent to Patricia Devlin in May 2021, promising to rape her baby, which Facebook said it had difficulty investigating. The message was still visible to Devlin in July 2021. The name and face of this user have been obscured in accordance with research ethics protocols.

Devlin, whose experiences were initially documented in a UNESCO-published discussion paper associated with this study in April 2021 (Posetti et al., 2021b), contacted the researchers for assistance regarding this threat, after she reported the profile to Facebook. However, a senior Facebook representative, who had earlier asked for the company's self-reported efforts to combat online violence against women journalists be included in this study, told the researchers and UNESCO representatives that they had no record of Devlin having reported the incidents to Facebook.

In an email sent to the lead author on May 13 2021, Devlin described the process of trying to get Facebook to take action to protect her by removing the threats and investigating the account as "absolutely staggering". She said that while she and others had reported the profile within the Messenger app multiple times, Facebook had not acknowledged the reports. As intermediaries, the researchers supplied Facebook with screen shots of the threat and confirmed details of Devlin's profile. But Facebook responded that screenshots did not help - they needed the researchers to supply the URL address for the Facebook profile of the person attacking Devlin because, they said, it was too difficult to trace the Facebook user through screenshots of their profile. This put the burden to document and investigate the threats on the journalist under attack, but the researchers supplied the URL required by Facebook

Facebook also suggested that Devlin apply for a "blue badge verification" to verify her profile, and said she should use Facebook Messenger's existing tools to protect against unwanted messages, such as disabling unsolicited messages (Facebook n.d.a), or block the user within Messenger (Facebook n.d.b). However, Devlin noted that these measures were not enough - blocking would not make the threat disappear. She wanted the incident investigated and the perpetrator de-platformed. In addition, disabling unsolicited messages was professionally unadvisable for Devlin because sources routinely contacted her via Facebook Messenger. Facebook also told the researchers to advise Devlin that she could use a "powerful FB anti-harassment feature" - controlling who can comment on your public feed (Facebook n.d.c) - but this particular threat did not come via her public feed.

Then, the Facebook representative told the researchers that they were in direct contact with Devlin. But the journalist denied this was the case in a message she sent to the lead author on May 24 2021: "They did respond to an email from my company where they AGAIN asked what I had done to report and block the profile; when I did it; how I did it; and claiming what I had already said, to you and my company, did not match their internal records."

This episode followed serious instances of Facebook-based online violence experienced by Devlin in 2020, which remained under police investigation in mid-2021. Police visited the homes of Devlin and a colleague to warn them of credible death threats following publication of false content about the journalists (including blaming Devlin for a bombing) on a Facebook page used by loyalist paramilitary groups. Devlin had reported the attack on 27 November 2020 to both Facebook and the Police Service of Northern Ireland (PSNI). After a second death threat, police again attended Devlin's home to inform her that they had received intelligence that she would be shot in the following 48 hours. Devlin made a statement to the police and provided the names of two individuals associated with the attack, but the PSNI could not go and speak to these people because Facebook said it was unable to trace the page. It was still active and running in late 2021, heightening the offline risks she faces. Devlin reported the posts separately to Facebook but she was told that they did not breach its 'community standards'.

Additionally, Devlin said she also reported a page to Facebook and the police which had accused her of being behind her own death threats - a classic disinformation tactic deployed in orchestrated online violence against women journalists.

This so-called 'community page' is being run by operatives linked to a paramilitary drugs gang, according to Devlin: "The individual who is involved with this page has been at the centre of the abuse and I have named him previously. He uses multiple fake profiles and 'anon' pages on Facebook to target me and I made numerous statements to police about this which are still being looked into".

Offline threats - including death threats taking the form of graffiti painted onto brick walls in Belfast - were also connected to these online attacks on Devlin.

Conclusion: In the specific case of Patricia Devlin, Facebook failed to respond effectively to some of the most serious examples of online violence surfaced by this study. Their response mechanisms were disjointed at best. The obfuscation, victim blaming and deflection that many research participants described as re-traumatising in the context of online violence attacks were also in evidence. The proprietary tools that they provided the researchers to share with Devlin were ineffective or inapplicable, and required her to do the labour of verifying and documenting the abuse. Even when the Police Service of Northern Ireland called on Facebook for assistance to trace the source of the threats, they were deflected. Despite its technical expertise, Facebook was apparently unable to trace the account of a user threatening to rape a journalist's baby despite being handed the account holder's name and screen grabs. When URLs connected to the profile were supplied by the researchers at Facebook's request, the company said it had no record of Devlin's multiple incident reports made via Facebook Messenger.

Facebook declined to provide a representative for a research interview for this study.

Gap 2: Inadequate responses to 'below-the-radar' abuse and linguistic and cultural content moderation challenges

This study points to inconsistencies in how the platforms detect and act on abuse across different languages and countries. While abuse moderation in languages spoken in the companies' major markets is somewhat better addressed, this is far from being the case in less prominent local languages, including in countries such as Kenya, Nigeria, Sri Lanka, the Philippines and Pakistan. At an intersectional abuse level, interviewees highlighted that being abused in languages such as Persian, Sinhala, Tagalog, Malay, Urdu or Tamil adds to the frustration of trying to report abuse to the companies which have very limited abuse moderating capability in these languages, and therefore they often do not respond appropriately.

The cultural and linguistic context of online violence is currently not captured by the platforms' algorithms and policy implementation workflows. This failure was strongly criticised by the women journalists and policy experts interviewed for this study. This reflects that "87% of Facebook's global budget for time spent on classifying misinformation goes towards the United States, while 13% is set aside for other countries — despite the fact that North American users make up just 10% of its daily users" (Popli, 2021). Languages such as Arabic are particularly poorly-served. For instance, according to one study, only six percent of Arabic-language hate content was detected on Instagram before it made its way onto the photo-sharing platform, and just 40 percent was proactively detected and taken down from Facebook (Scott, 2021b). Women and the LGBTQ community were primary targets (ibid.).

A key problem identified both by the interviewees and other recent recent research (Posetti et al., 2020; Popli, 2021; Scott, 2021b) is that the companies' have insufficient numbers of human content moderators and local policy staff who know and understand the nuances of international socio-cultural and political contexts, with linguistic capabilities that extend to minority languages and dialects. Women journalists interviewed said they found it hard to explain the

specific cultural implications of abusive language and the type and severity of the attacks against them in relation to universal platform community standards when they reported abuse to the platforms, resulting in delays or a lack of action. According to interviewees, in Pakistan, these companies are particularly weak in responding to hate speech and harassment against women journalists in Urdu - one of the country's two official languages. This was a point underscored by Sabahat Zakariya, who has reported for various news outlets in Pakistan (including BBC Urdu). She said the companies are not responsive to specific cultural problems and their repercussions:

> They don't sometimes understand how crucial it may be for somebody's life. There is no response at all, especially if it comes with [some kind of] a hashtag or abuse in the Urdu or Punjabi language, or another local language. You try to explain to them the nature of the abuse or threat, but they don't get it. They simply don't have [enough] human resources from these countries to understand the nuances of what is going on. So it's a very, very, one-size-fits-all approach, and it's really wrong.

The lack of an Afrocentric approach by social media also needs to be urgently addressed according to Kenyan editor Catherine Gicheru:

> Not all black women online face online harassment the same way...If you're African, you're black, you're a woman, you will experience online harassment differently than if you are an African American woman online...And just the fact that you're Zimbabwean might attract some kind of harassment.. And then it goes even further, it can go lower because you're from Bulawayo or because you are a Shona and not a Ndebele.

Broadcast journalist Cecilia Maundu pointed out the difficulty of reporting on the platforms using different languages in Kenya, where groups online are formed in vernacular languages. According to Catherine Gicheru of the Africa Women Journalism Project, much of the online harassment she witnesses, such as body shaming remarks or subtle threats, is not in English and so is not monitored or "understood" by the platforms' algorithms. The Media Council of Kenya has lobbied the social media companies - including Facebook and Twitter - to create tools that flag offensive content, aid the reporting of it, and enhance moderation in multiple languages, and with appropriate cultural context.

A particular example is the experience of the Al Jazeera journalist Ghada Oueiss, who reported to Twitter a fraudulently altered image purporting to show her naked in a jacuzzi, screengrabs from which she said were retweeted 40 thousand times. The image and a series of photos showing her eating a meal with colleagues were stolen from her phone, she alleges, as part of an orchestrated attack designed to discredit her[25]. They were distributed with messages alleging she was an alcoholic, drug-addicted prostitute. Twitter's initial perspective was that the content did not violate their policies. However, being a journalist working in the Arab States, the wide circulation of this stolen and manipulated material put Oueiss at risk of retribution and significant reputational damage.

Lawyer Caoilfhionn Gallagher QC said that getting the social media companies to take action against those perpetrating online violence against her BBC Persian service clients in Farsi was extremely difficult due to language issues. Similar points were made by women journalists in the Philippines, Sri Lanka, South Africa, Lebanon, Mexico and Brazil.

Twitter's Nick Pickles stated that building capability to moderate content and respond to threats in diverse cultural and linguistic context on a global scale is extremely challenging and accepted that human intervention would also likely be required to deal with more nuanced cases.

The failure to effectively address multilingual content moderation also has negative freedom of expression implications. There is the need to counterbalance such interventions against the potential for overreach and censorship justified on the grounds of curtailing abuse. In particular, linguistically deficient algorithms not only miss abuse, but also frequently delete abusive posts. For instance, in one study (Scott, 2021), 77 percent of deleted content on Facebook in the Arab States was found to be non-violent and legal, meaning the algorithms harmed people's ability to express themselves online, and limited the reporting of potential war crimes.

In Nigeria, journalist Kiki Mordi said algorithms need to be localised, "...because some of the offensive words that Nigerians use aren't picked up". Adeboye Adegoke, Senior Program Manager at Nigeria's Paradigm Initiative, said that understanding cultural context is important when designing algorithms so that cultural nuances, and the different ways abuse manifests itself around the world, may be better detected and prevented. For him, that means major tech companies must employ locals to ensure there is diversity of personnel and diversity of algorithmic engineering.

In Brazil, Gabi Coelho of daily newspaper *O Estado de S. Paulo* noted that black women are particularly at risk of hate speech on Twitter: "Twitter, the network I use the most, is very violent for black women, especially, so there is the gender issue and there is also the race issue … There was a week when I came to report more than 40 profiles."

In addition, the current algorithmic processes for content moderation and removal appear to be prone to bias and reinforcement of inequality while also being opaque. ARTICLE 19's Thiago Firbida noted that while people can see the impacts of online violence and its disproportionate effects on women, Black and Indigenous people, there is an "invisible structure that is built by white men, straight Europeans and Americans" which means that it is not possible to "see how the platform's algorithms facilitate the spread of attacks".

25 Ghada Oueiss said she knew she had been hacked when private photos on her phone were shared online, and her claim was confirmed by Forbidden Stories and Amnesty International. Leaked documents showed that Pegasus spyware, created by Israeli surveillance technology company NSO Group, had been installed on Oueiss' phone, turning it into a surveillance device (Solon, 2021).

The failure to remove misogynistic posts from public fora entails a risk of entrenching cultural, racist, sexist and abusive bias into the next generation of content moderation machine learning models, which are being trained on present-day abusive content (Hao, 2021). Twitter's Nick Pickles said: "The reason for investing in countries like India, like Ghana, is to build out that capacity where, being totally frank, we don't have the right diverse capacity that we need. And so those investments are explicitly intended to start building that capacity".

Failure to detect abuse 'below the radar'

As the companies' abuse prevention methods have begun to improve with regard to addressing highly explicit abuse and hate speech, abusive behaviour has started to shift towards harder to detect implicit cases. Such implicit abuse is often contextual in nature and grounded in cultural and political specifics, which makes it hard not only for the platforms' machine learning algorithms to flag, but also for employees who are not from the same culture and country to moderate. Additionally, rape and death threats go undetected when expressed via implicit wording, imagery and memes. One example is when the so-called Islamic State evaded moderation on Facebook via the use of local Arabic slang to spread hate speech (Scott, 2021).

One particular source of complaints from the interviewees was the companies' failure to address online violence that occurs via their private messaging services, such as Facebook Messenger, Direct Messages on Twitter (DMs) and Instagram Direct Messenger. These avenues are frequently used by perpetrators to deliver death threats, or sexual harassment through unsolicited sexually explicit messages and images (one of the top attack modes identified by the survey respondents). Where there is privacy preserving end-to-end encryption, this makes the task of monitoring and moderation challenging – although meta-data can reveal behavioural patterns and flows of content, which the company can control if it so decides.

The BBC's Marianna Spring described the frustrating process of trying to report threats and abuse on Facebook Messenger. This has involved her compiling and emailing evidence of the abuse (including links and screengrabs), then being asked to repeat the process of logging the abuse using Facebook's on-platform reporting tools (which rarely generate a response), and finally being told that nothing can be done because the content is "private". At this point she has escalated cases to the police for investigation, but meanwhile the reported accounts remain active on Facebook.

Vice's UK editor in chief Zing Tsjeng reported a user on Twitter Direct Messages who called her a "C-word who eats dogs" in May 2021, but Twitter did not take any immediate action against them. Twitter's Nick Pickles acknowledged that his company places the onus for monitoring online abuse sent via Direct Message on the receivers: "We do rely on users flagging to us things in their DMs that we should look at… This is also something that, potentially longer term, we can use technology for. But I think it's fair to say that the complexities of moderating private spaces, and the legal frameworks that go around them, is greater than the public space on Twitter." He added: "We expanded our policies to cover unwanted sexual advances, which I think for female journalists do often come through DMs. And if they're reported as a DM, we can investigate them, and look at them."

In Sri Lanka, women journalists also described voice-based harassment received via these closed channels that are even harder to police. In addition, according to our interviewees, some abusers use intimate verbal threats, calling reporters via Google Chat or Facebook Messenger and shouting "murderer" or "racist".

Twitter Spaces was also identified by one user in Pakistan as a place where she had experienced organised audio-based abuse.

A further area of online violence that is delivered 'below the radar' concerns subtle and coded online violence. In Sri Lanka, for example, discriminatory speech is often conveyed with 'humour', via memes and cartoons, in order to avoid moderation triggers for removal. Similarly, captions are often placed over images, to avoid text-based detection.

Freelance journalist Tulasi Muttulingam, who created a Facebook page called Humans of Northern Sri Lanka to document people's personal stories of war, said that social media perpetrators are "cunning" and use passive rather than active threats to avoid moderation. She gave these examples: "I hope you get hit by a bus and die. I hope a white van picks you up and you're tortured and raped and murdered, all that sort of nonsense." Another example is a death threat sent to journalist Maria Ressa in a tweet on February 21st, 2021 (Posetti, Maynard, & Bontcheva, 2021) where the text of the tweet itself was not abusive, and the text in the accompanying image is not easy to process by automated tools, but taken together the menacing underlying meaning is clear to human readers.

Other forms of more subtle abuse include body-shaming and framing women journalists as 'controversial' or 'divisive' figures as well as drawing attention to their employers in the hope of getting them fired. Nigerian documentarian Ruona Meyer, whose trolls targeted the BBC following the broadcast of a programme she made, said: "Talking about somebody's body parts isn't necessarily hate speech, but it's devastating." She feels Twitter's reporting processes lack nuance and therefore fail to recognise and respond to less overt forms of abuse. Twitter's Nick Pickles affirmed that online abuse tactics have morphed significantly with some forms of attack being harder to combat, such as when the attacks are designed to undermine credibility or trust in critical reporting. He added that "this kind of far more pernicious, subversive attack" on the credibility of journalism had been "particularly challenging when we've had public figures making these sorts of statements".

In Northern Ireland, when Patricia Devlin received an indirect death threat referencing an investigative journalist from her newspaper (*Sunday World*) who was assassinated in 2001 (RSF, 2020a),[26] the contextual nuance contained in the threat, along with the limited press freedom and journalism safety expertise at the platform, meant that it was not detected nor understood as a death threat until Devlin was able to get this account taken down. Lawyer Caoilfhionn Gallagher QC noted a similar incident involving the BBC Persian service journalists she represents (OHCHR, 2020f, BBC, 2021c), saying that the companies "... simply don't spot it when it's a pattern of threatening behaviour against women - a pattern of threatening behaviour which involves making subtle references to other cases. They don't get picked up at all… The key thing here is prevention is better than cure".

Another hard-to-address case is that of discriminatory abuse served through emojis with racist connotations, such as the monkey emoji which is weaponised in racist posts (MacInnes, 2021). Detecting contextualised racist cases automatically is currently beyond the platforms' technological capabilities. However, Twitter's Nick Pickles highlighted that reporting such contextual cases is helping them to improve their content moderation guidelines: "The challenge you've got is basically the scale of emojis on the platform. If you had some technology that surfaced every use of the monkey emoji, for example, you're going to get a huge amount of false positives. And so that's where I think the technology still has some refinement to do".

26 It read: 'You're going to end up like Marty O'Hagan'. Since the assassination of Martin O'Hagan near Belfast in 2001, the first killing of a journalist in the line of duty in the UK was Lyra McKee in Derry in 2019. No one has ever been convicted of O'Hagan's murder (RSF 2020a).

This shift from explicit threats and abuse towards more subtle kinds of abuse like networked gaslighting was also apparent in the findings of the big data analyses focused on Maria Ressa and Carole Cadwalladr, companion outputs from this study, snapshots from which were published in a 2021 UNESCO discussion paper. Long-range attacks are also associated with these less overt forms of online violence. However, sharpening anti-abuse policies to capture more subtle forms of online violence can risk catching strong opinions and legitimate critique in the net. Twitter's Nick Pickles: "One of the biggest challenges for any company in this space is trying to separate out harassment from very strong opinions. And that is something which I think, given the sort of polarisation we see around the world and the political controversy we see, has also made that job harder".

Gap 3: Need to address the cross-platform nature of online abuse

The companies' policies and processes also fail to account for the cross-platform nature of online abuse experienced by women journalists. The BBC's Marianna Spring described coordinated campaigns of abuse and harassment that begin on YouTube in comments that trigger escalating abuse, which then spread to other social media platforms: "It's always cross-platform...that YouTube link is sent to me on Instagram and on Facebook and on Twitter and [it floods] my mentions. It's never limited to a single platform and this ecosystem is very open".

In the US, former *New York Times* reporter Taylor Lorenz found that a one year-long episode of online violence she experienced may have been fuelled by tweeting a story from The Verge about a tech company CEO. A high-profile investor then incited harassment against Lorenz on the audio app Clubhouse, as well as other platforms:

> Once they...found me, they've never let go. And it's like all of these far-right actors have aligned themselves with the worst people in Silicon Valley. And then it's just been a year of harassment and abuse, on every platform...Instagram, TikTok. Twitter, Facebook Messenger... They were calling me and sending me crazy messages and stuff… Thousands of emails...thousands of direct messages.

Brandy Zadrozny of NBC News said she was "a big topic of discussion" in "white nationalist spaces" and it was mentally exhausting to monitor all the sites where she was being discussed and threatened. Julia Carrie Wong, Senior Reporter at Guardian US, described being targeted by 'Q', the reputed figurehead of the QAnon conspiracy (Wong and Collins, 2018), who "specifically sent people to my Twitter account in a 'Q-drop'" which was "overwhelming and unpleasant"[27].

27 QAnon is an internet conspiracy theory which began in October 2017 on 4chan claiming politician Hillary Clinton would be arrested. Q is the pseudonym of the anonymous posters and "a motivating factor for many of the insurrectionists who attacked the US Capitol on 6 January 2021". Twitter banned 70,000 related accounts after the event (Wong, 2021).

When Serbian journalist Jovana Gligorijević from Vreme tries to take 'mental breaks' from Twitter, abusive tweets are uploaded as screenshots on Facebook which she is alerted to via that platform's facial recognition technology. Brazilian researcher Claudia Lago described the manifestation of the cross-platform trend in Brazil: "Online attacks originate as fake news in the underworld of Whatsapp groups; when they reach Twitter they have already done real damage".

Another shortcoming is the predominant focus on responses by the "big four" online platforms (Facebook, Twitter, Instagram, and YouTube) which overlooks abuse targeting women journalists through other platforms, networks and apps. One example is a first-person blog post written about the BBC's Marianna Spring in her 'voice' on a fringe blog, which attracted around 3,000 unmoderated comments, including rape threats and other highly sexualised threats. And at the time of her interview for this study in March 2021, Taylor Lorenz, then still with *The New York Times*, received an alert of a "disgusting hate comment" in the comments section of her Substack newsletter.

Individually, the platforms have much more detailed information, compared with independent researchers and the targets of gendered online violence themselves, as to the origins of abuse (e.g. IP addresses from which a given user has posted abusive messages), possible coordination between users during pile-on abuse, and the deployment of 'sock puppets' and bots in orchestrated attacks. Collectively, they could work very powerfully to combat online violence incidents in real time, through information sharing and collaborative responses to coordinated attacks. They could also work jointly on developing more effective and gender-sensitive policies and tools to improve reporting of cross-platform instances of online violence against women journalists and human rights defenders.

Gap 4: Platform policies need to cover abuse from prominent political figures

A question raised by several interviewees, also evidenced in the companion big data case studies, is whether it is justifiable to exempt politicians and other political actors from content moderation policies and actions, given that some are prominent perpetrators of online threats, abuse and harassment.[28] Both women journalists and the platforms have acknowledged that in some cases prominent political figures (some of whom have had their social media accounts suspended for periods of time) and influencers play a major role as instigators of online violence. However, remedial actions from the social media companies in such cases tend to be more lenient due to policy exemptions on the grounds of "newsworthiness" and prominence (Reuters, 2021).

Based on her own experience, the BBC's Marianna Spring suggested that companies should better honour their duty of care to users by: "...breaking down these ecosystems, which generally have a few central figures at their heart, and as we've seen multiple times it tends to be pretty effective when those people disappear". Former UN Special Rapporteur David Kaye said: "It really does go back to the platforms and their treatment of political figures who engage in this generic kind of incitement against journalists. The platforms should consider that as they take action either to protect or promote political pages like Trump's."

In response to the recommendations of its own Oversight Board, in June 2021 Facebook announced changes to its policies and committed to acting quickly on posts by influential users which may lead to harm (Facebook, 2021b; Posetti and Bontcheva, 2021). As this policy change was still new at the time of writing, it was not possible to independently evaluate the extent to which it could help.

28 As of June 6 2021, the debate as to whether Facebook will end its policy of exempting political figures from content moderation rules was still ongoing, despite reports of an imminent change. https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy

However, it is important that all platforms adopt such policies and enforce them in a consistent manner.

Gap 5: Failures of algorithms for content recommendation and moderation

Algorithms for prioritising and recommending content, users and groups have been found to promote misogynistic hate.

Researchers have identified, for example, deficiencies in the metrics used by some companies to prioritise content: "...the top performing domains were those that surfaced in users' feeds over and over—including some highly partisan, polarising sites that effectively bombarded some Facebook users with content" (Faife, 2021). In 2020, the Wall Street Journal revealed a 2016 internal company presentation by Facebook researchers that stated "64% of all extremist group joins are due to our recommendation tools", showing that most of the activity came from the platform's "Groups You Should Join" and "Discover" algorithms. "Our recommendation systems grow the problem," according to the presentation.[29]

Due to the lack of transparency, it is hard for external researchers to establish the full role of these algorithms in online abuse. Therefore, a BBC documentary in 2021 reported by one of this study's research subjects, Marianna Spring, created a fake troll persona on YouTube, Facebook, Instagram, Twitter and TikTok to test algorithmic referrals (Spring, 2021). "Barry" - the invented persona - was interested in anti-vaccination and conspiracy content and "he" initially engaged with only a small amount of misogynistic content. After two weeks of following recommendations on each of the platforms, the top recommended pages to follow on both Facebook and Instagram were almost all misogynistic, whereas TikTok suggested no anti-women content and Twitter and YouTube only a small amount (ibid.). The fact that a small scale and brief study can elicit such striking results points to ongoing problems on some platforms, potentially also exacerbated by insufficient moderation of misogynistic content. The lack of independent access to companies' data means that large scale, multi-platform, longitudinal, independent studies that can monitor, evaluate, and compare effectiveness of platform algorithms and approaches is also a gap that needs addressing.

Deficiencies in content moderation algorithms include a limited ability to detect nuanced online violence against women journalists. But simultaneously, they can also lead to censorship of non-offensive content, due to the algorithms' inability to interpret the wider conversational and cultural context. Journalists in Eastern Europe, Asia, Latin America and the Arab States interviewed for this study reported experiencing algorithmic censorship applied to feminist content. Beyond images, the algorithms for detecting misogynistic and other abusive posts have also been found to censor legitimate conversations, due to lack of local context. One such reported example involved Facebook users being muted, banned, or warned when posting about a well-known UK landmark (Plymouth Hoe), as its name was erroneously flagged as misogynistic.

These cases highlight the crucial importance of a human-in-the-loop approach to content moderation, which is not only sensitive to local cultural and linguistic context, but also complemented by an effective appeals process - including the ability to appeal refusals to remove abusive content - and robust transparency policies. Moreover, as Twitter itself noted in a policy paper: "Content moderation is more than just leave up or take down. Regulation should allow for a range of interventions, while setting clear definitions for categories of content."

29 "Facebook Executives Shut Down Efforts to Make the Site Less Divisive. The social-media giant internally studied how it polarizes users, then largely shelved the research", https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Gap 6: Deficiencies in collaboration and multistakeholderism

There is little systemic engagement by the companies with civil society, media, governments and experts in developing policy responses to gendered online violence. Some hopeful signs of collaboration are slowly emerging, but there are still often gaps in stakeholder representation. Interviewees said, for example, that in Mexico, where company representatives meet periodically with government officials and civil society representatives to discuss trends in digital attacks and share good practices, there are only limited exchanges with independent news media and academic experts.

More needs to be done to bridge the industry divides between big tech companies and news outlets to proactively address the issue of gender-based online violence against journalists, as indicated by Sveriges Radio CEO Cilla Benkö. Reflecting on International News Safety Institute (INSI)-facilitated conversations between (mostly Western) news organisations and the platforms about addressing the online abuse of women journalists, she said: "It is good that as industries we communicate with each other in an open and constructive manner. But much more needs to be done" (Benkö, 2021). Exploring the "more" could lead to country-level facilitated workshops designed to improve understanding and trigger more effective responses.

At the level of process, Facebook has established a 'Trusted Partners' programme to provide more context to the flow of requests to flag, block or remove content and people on its platform or on WhatsApp. This programme has been operational in select countries in Asia, the Arab States, Latin America and Europe according to research for this study, and parallel research (Sinpeng et al., 2021). The programme offered journalists and human rights defenders the option to escalate complaints to the platforms via civil society organisations and provided more support to help counter the automatic decisions being made by moderation software. But there is scant public information about this programme and awareness is low, particularly in the Asia Pacific (ibid). Twitter also provides a one-on-one service via their Latin America Policy team, in partnership with ABRAJI (Associação Brasileira de Jornalismo Investigativo, also known as Brazilian Association of Investigative Journalism), ARTICLE 19 and Reporters without Borders (RSF), with a fast-track channel to prioritise complaints from these organisations. However, according to interviewees, many cases fall through the cracks, and freelance women journalists along with others who lack the contacts, are at a disadvantage.

Facebook told researchers in an email that the Digital Security Helpline of the civil society group Access Now (see Access Now, n.d.) has high priority access to their systems. However, there was limited awareness among the research participants about this service. For example, although the service has a Tunis-based contact point, according to Faouzia Ghiloufi - executive member of the National Syndicate of Tunisian Journalists and Vice Chair of IFJ Gender Council - Tunisian civil society organisations do not have direct contact with social media platforms, and the onus falls on the individual journalists to reach out and report to the social media platforms. While civil society contact points for rapid escalation in cases of gendered online violence targeting journalists could be valuable, this does not exempt the companies from having staffing capacity to cover local languages and contextual understanding.

4. The need to move fast and fix things

Most interviewees assessed that the companies' enabling role in gender-based online violence against journalists could not be properly addressed until the

business models and technical design were overhauled to prioritise safety and human rights over profit. The dominant players set up their highly profitable business models to maximise traffic and engagement, rather than to protect journalists, human rights defenders, democracy, or marginalised communities. Therefore, these models favour inflammation rather than accurate information and this principle is embedded in the design decisions of the platforms[30]. This underpins the view of Dr. Michelle Ferrier, founder of TrollBusters, that the companies are complicit in maintaining an environment of tension that serves business purposes and also reinforces patriarchal norms in digital spaces.

A core problem is reliance on attention, not least that is centred around harassing or abusive engagements, which grows market share and boosts already astronomical profits. This can be damaging to women journalists in particular, even if legal. South African investigative journalist Qaanitah Hunter linked the platforms' business models, which "incentivise and reward trolling", to their reluctance to effectively combat gendered online violence. "They have no interest in getting hate off [their platforms]," her compatriot, editor Ferial Haffajee said. ARTICLE 19's Legal Officer Paulina Gutiérrez said "the business model is a problem in itself" and that it enables attention-driven pile-ons against women journalists and many other groups at risk and those facing discrimination.

The scale of the gender-based online violence problem and its networked nature mean that it can go viral very easily and quickly, with wide-ranging impacts including on democratic deliberation and digital citizenship more broadly. These developments, along with UN-level acknowledgement of Facebook's "determining role" in human rights abuses against the Rohingya in Myanmar (Miles, 2018; McPherson, 2020), and its facilitative function in the 2021 Capitol Hill insurrection (Silverman et al., 2021; Zuckerberg, 2021; Reuters, 2021) highlight the urgency for action.

A report by civil society monitoring organisation, GLAAD, titled the "Social Media Safety Index" has assessed how companies detect abuse against LGBTQ users, and argues that "bad actors" have learned how to game AI systems (GLAAD, 2021). This is one reason for the NGO urging the "need for human moderation — as well as a corresponding need for ethical and responsible employment practices in relation to these workers". According to GLAAD, Reddit conducted a study on hate and abuse on its services and made these findings public – and recommended that Facebook do likewise. In addition, GLAAD called on Facebook to allow an "independent audit, specifically focused on its lack of transparency" (ibid.).

Guardian US' Julia Carrie Wong pointed to the need for an overhaul of Facebook's ethical and normative frameworks - led from the top:

30 One example is a recent study of the role of highly active, abusive users in shaping content recommendation on US Facebook pages https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/

> I don't necessarily need Facebook to provide me with personal on-platform protection, but my society is being damaged by what's happening on Facebook... the QAnon movement was allowed to grow so big on Facebook with Facebook support, rather than with Facebook trying to limit it... it's a lack of being willing to make moral judgments that would protect fundamental features of a liberal democracy.

However, former HuffPost UK editor-in-chief[31] Jess Brammar said that expecting the platforms to not just 'clean house', but rebuild the infrastructure was a futile wish: "I feel like it's too late to come from the platforms… We are years into this now". She called for broader social and political reform to address the problems at the root. "[I]t lets off the people who have got us into this situation... if we just look at the platforms."

5. Conclusion

While structural sexism and misogyny, populist political instigators and partisan news media contribute to gendered online violence, the platforms bear a major responsibility for enabling and facilitating the problem – and for addressing it. For women journalists to be able to work safely online, the policy gaps identified must be addressed. Business models and algorithms must be restructured and redesigned. And more effective and comprehensive tools and protocols for detection, reporting, moderation and countering of online attacks on journalists are required. Additionally, there is a strong need for transparency; for independently defined and evaluated measures of the effectiveness of abuse countermeasures. It is time to move away from the current approach of largely ineffective self-regulation.

31 At the time of the interview Brammar was editor-in-chief of HuffPost UK. In September 2021 she was appointed the BBC's executive news editor of news channels.- a move which led to further pile-ons: https://www.pressgazette.co.uk/bbc-boss-fears-jess-brammar-effect-will-affect-hiring-of-journalists-with-diversity-of-views/.

Recommendations for Action

The following research-based recommendations are proposed for consideration by social media and search companies as key vectors responding to online violence against women journalists globally.

Big tech companies could:

1. Continuously review their policies, algorithms and moderation processes, to address the evolving nature of gender-based online violence, while working closely with women journalists and civil society groups to co-design new solutions.

2. Develop more sophisticated abuse reporting systems with capacity for escalation for women journalists under attack (and their employers), recognising their particular vulnerabilities along with the implications for press freedom.

3. Implement a coordinated multi-stakeholder approach to protecting women journalists from online violence, which brings together all platforms, female journalists, civil society, news organisations, governments, and independent experts - at national and international levels.

4. Initiate platform-platform cooperation, since online violence often jumps across platforms and exploits the weaknesses of each.

5. Implement proactive countermeasures which reverse the onus on women targets having to report online violence to start with. This might involve using human moderators and artificial intelligence technology to more effectively filter out threats, abuse and harassment at the point of origin.

6. Retain data documenting attacks to aid targets wishing to access and use it for research or legal action. Such proactive steps could also link to monitoring processes to develop an 'early warning system'[32] so as to better protect women journalists at the outset, or in the midst of an attack.

7. Build shields that enable users to proactively filter abuse which could be quarantined for review and response. Such systems should also provide prioritised pathways for women journalists under attack and news organisations seeking to report online violence.

8. Provide authorised independent researchers with secure and privacy-preserving access to archives of moderated content and user appeals in

32 ICFJ and University of Sheffield computer scientists are in the process of developing such an 'early warning system' under commission from the UK Foreign, Commonwealth and Development Office (FCDO)

a standardised format, to enable transparency and independent audits of moderation decisions about threats made to women journalists.

9. Use the findings of such independent audits to adjust both human and algorithmic moderation practices, to strike a better balance between protecting freedom of expression and prohibiting abuse.

10. Implement an effective human-in-the-loop approach to content moderation coupled with a timely and effective appeals process - including effective systems to appeal against company refusals to act against online violent content and perpetrators.

11. Report transparently on how human moderators and artificial intelligence algorithms are trained to detect online abuse.

12. Define effective policies for detecting and penalising repeat offenders, to stop the same abusers assuming new online identities after action taken such as suspension or de-platforming.

13. Develop markers for abuse perpetrator accounts, similar to systems used to identify disinformation purveyors.

14. Establish clear and transparent community rules on what constitutes online violence and cease making exceptions for political actors, influencers, public figures and other high-profile users, whose high number of followers makes it easy for them to instigate abuse pile-ons.

15. Create more effective content moderation tools that provide sufficient support for all languages in which their services are offered (including vernacular or slang), and which are sensitive to contextual and cultural norms.

16. Technical solutions should be supported by human contact points who are familiar with a country's cultural, political, linguistic, and religious context and are well versed in local languages. These people should also possess press freedom, gender and journalism safety expertise, and be able to assist women journalists under attack.

17. Establish task forces and carry out proactive programmes to protect women journalists from certain abuse types, such as the dissemination of intimate images and doxxing.

18. Take effective steps against the use of bots, false accounts and sock puppet networks to prevent coordinated attacks and pile-ons that are frequently used in targeted online violence against women journalists.

19. Conduct regular human rights impact assessments as well as retrospective studies into the problem, including review of company policies and responses to gender-based online violence, and make the findings public.

20. Provide detailed transparency reports on actions taken against online violence against women journalists, broken down on a national level and including meaningful quantifiable metrics, beyond the total number of accounts removed and posts moderated. Reports need to also include appeals and their outcomes, along with data about notifications and responses to online violence reported by women journalists. They should also include statistical representation and analysis of content that stays up after being reported by journalists as abusive, offensive or threatening - not just on what is taken down.

21. Monitor the intersectional nature of attacks on women journalists who are targeted more than others because they belong to religious, racial, or ethnic minorities, Indigenous groups or identify as members of the LGBTQ community.

22. Strike a better balance between supporting freedom of expression and prohibiting online violence, and recognise that international human rights instruments and UN resolutions require that women journalists be able to work online free from threats and harassment.

23. Support independent research (i.e. with no strings attached) on campaigns of violence against women journalists, and responses to these.

A NOTE ABOUT OUR METHODOLOGIES

The survey method adopted was 'purposive sampling', with 'snowballing' techniques used to generate responses within the international field of journalism. The results, therefore, are not generalisable, although it is legitimate to extrapolate many patterns that may well have wider applicability. To avoid illegitimate or inauthentic responses and ensure data integrity, the survey was distributed digitally via the closed networks of UNESCO and ICFJ, our research partners, civil society organisations focused on media development, journalism safety and gender equality, and groups of professional journalists. The survey ran from September 24th to November 13th 2020 and it garnered 901 valid responses. The survey results were then disaggregated along gender lines, and a subset of data from 714 respondents who identified as women was isolated for analysis. In parallel, we identified 183 interviewees through the survey and institutional outreach, as well as via the networks of the research team. The interviews were conducted face-to-face (where COVID-19 restrictions allowed) and via digital channels. Most of the interviews were undertaken synchronously by the researchers identified in this report. The vast bulk of interviewees chose to be publicly identified after being offered the option to remain anonymous.

For the big data case studies on Maria Ressa and Carole Cadwalladr 2.5 million social media posts were collected over the course of five years and 13 months respectively. Relevant subsets of these collections were identified for network analysis and deeper investigation via Natural Language Processing (NLP). The results were synthesised with the long form qualitative interviews and contextualised via detailed timelines developed through desk research.

The University of Sheffield (UK) granted ethics clearance for the English language version of the survey and English language interviews. Translations of the survey into other languages were conducted by UNESCO and reviewed by ICFJ. The University of Sheffield also provided ethics clearance for quantitative data gathering and analysis associated with the big data case studies featured here.

in cooperation with