ARTICLE 19

Online harassment and abuse against women journalists and major social media platforms

2020

First published by ARTICLE 19, 2020

ARTICLE 19
Free Word Centre
60 Farringdon Road
London EC1R 3GA
UK
www.article19.org

ARTICLE 19 works for a world where all people everywhere can freely express themselves and actively engage in public life without fear of discrimination. We do this by working on two interlocking freedoms, which set the foundation for all our work. The Freedom to Speak concerns everyone's right to express and disseminate opinions, ideas and information through any means, as well as to disagree from, and question power-holders. The Freedom to Know concerns the right to demand and receive information by power-holders for transparency good governance and sustainable development. When either of these freedoms comes under threat, by the failure of power-holders to adequately protect them, ARTICLE 19 speaks with one voice, through courts of law, through global and regional organisations, and through civil society wherever we are present.

Contents

Executive summary

Women journalists face a distinct set of risks in carrying out their work. Although digital technologies have created new opportunities for women journalists and activists to communicate and organise, they have also reproduced patterns of harassment that women journalists face in their work.

Online harassment and abuse of women journalists have become more prominent and more coordinated in recent years and can occur on the basis of their reporting, or purely on the basis of being women operating in the public sphere. The objective of these types of attacks is to silence, stigmatise and intimidate women journalists.

ARTICLE 19 believes that social media companies have a role to play in both enabling women journalists' right to freedom of expression and also addressing gender-based harassment and abuse against them on their platforms. Concerns have been raised regarding how the three major social media platforms – Facebook, Twitter and Youtube (Google) – have defined prohibitions of offending content in their rules, policies and community guidelines as well as how they are enforcing these rules in practice.

ARTICLE 19 has previously offered a number of recommendations on the steps companies should take in order to demonstrate their commitment to the protection of freedom of expression. These recommendations were largely grounded in the UN Guiding Principles on Business and Human Rights, a non-binding set of international principles that companies ought to follow in order to guide their decision-making and policy development insofar as they may have an impact on the human rights.

ARTICLE 19 recognises that the major social media platforms have created policies to address harassment and abuse and that these policies are updated regularly and could be applied in cases of gender-based harassment and abuse against women journalists. However, the terms are often broad and vague, causing confusion, but also leaving platforms the flexibility to use these policies to their own needs. ARTICLE 19 notes that there is often a lack of consistent enforcement of the rules despite all three platforms providing reporting mechanisms.

This briefing looks closer at how these recommendations have been applied in practice to address gender-based harassment and abuse against women journalists. It also examines both the positive and negative aspects of the three social media platforms regulations on gender-based harassment and abuse. Finally, it offers a list of recommendations for the companies. In particular, social media companies should:

- Develop dedicated sections on gender-based harassment and abuse in their policies and community guidelines that are easily accessible and available in local languages.
- Increase transparency regarding the methods and internal processes for the elaboration of policies and community guidelines, their use of algorithms and on the complaints mechanism.
- Undertake human rights and gender discrimination impact assessments.
- Improve their internal redress mechanisms, respecting due process safeguards.
- Notify their decisions to affected parties and provide sufficiently detailed reasons for the actions they take against particular content or accounts.
- Consider further partnering with women journalists and civil society groups to develop practical strategy of research-focused and community-lead solutions on gender-based harassment and abuse.
- Consider joining or improve their engagement in multi-stakeholder regulatory bodies such as social media councils, that would allow better public oversight of their practices, including in the area of gender-based harassment and abuse.

Introduction

Women journalists face a distinct set of risks in carrying out their work. Entrenched discrimination means women journalists are at a heightened risk of complex abusive behaviour, that is committed, abetted or aggravated, in part or fully, by the use of information and communication technologies, such as mobile phones, the Internet, social media companies, and email.[1] This ranges from direct or indirect threats of physical or sexual violence, offensive messages, targeted harassment (often in the form of 'pile-ons', with multiple attackers, to privacy violations (such as "doxing, stalking, and non-consensual dissemination of intimate sexual images").[1]

Online harassment and abuse of women journalists have become more visible and more coordinated in recent years and can occur on the basis of their reporting, or purely on the basis of being women operating in the public sphere. The objective of these types of attacks is to silence, stigmatise and intimidate women journalists. This has a detrimental impact not only on freedom of the media and the inclusion of women's perspectives in public debate, but also on equality and women's equal right to freedom of expression.

Reports and studies show that in many instances, it can force women journalists to abandon certain types of coverage and their journalistic activities, diminishing their engagement in public discourse and limiting their contribution to newsgathering and reporting on politics, sports, economics, corruption, criminality, and other issues historically performed by men in the news sector.[2]

Concerns about the risks that journalists in many countries around the world face, is not exclusively targeted at women; men and gender-diverse journalists are also facing online and offline attacks, various forms of harassment and intimidation.[3] However, women journalists are exposed to additional threats,[4] caused by deep-rooted discriminatory practises, systemic inequality and restrictive gender stereotypes. This problem intensifies when prejudice and intolerance intersect with other forms of discrimination, including but not limited to racism, nationality, ethnicity, religion, and sexual identity and orientation.[5]

Since online harassment and abuse are predominantly occurring on social media, major social media companies have been at the centre of the discussions about the need to develop tailored and effective solutions to address the problem of gender-based harassment and abuse on their platforms. It is without doubt that policies and practices of major social media companies should be reviewed from the perspective of freedom of expression standards but also how they fare in addressing online harassment and abuse against women journalists from a gender and equality perspective.

[1] It should be noted that each of these problematic conducts may be defined differently in domestic legislation or in recommendations of regional and international human rights bodies; Other institutions, such as social media companies and academics, have also produced their own lexicon to conceptualise this phenomenon. While there is no universally agreed terminology to capture this phenomenon and its different forms, in this report, ARTICLE 19 employed the term "online harassment and abuse" as a generic term to capture the type of such conduct...

In 2018, ARTICLE 19 examined the extent to which major social media companies' community standards/guidelines, in particular those of Facebook, Twitter and YouTube, comply with international standards on human rights. In *Sidestepping Rights: Regulating Speech by Contract,* ARTICLE 19 offered a number of recommendations as to the steps companies should take in order to demonstrate their commitment to the protection of freedom of expression. These recommendations were largely grounded in the UN Guiding Principles on Business and Human Rights, a non-binding set of international principles that companies ought to follow in order to guide their decision-making and policy development insofar as they may have an impact on the human rights of their users. In 2019, ARTICLE 19 further launched our *Missing Voices* campaign in order to spur companies into action and encourage them to step up their transparency and due process efforts with a view to better protect their users' rights. Meanwhile, ARTICLE 19 has been developing a potential model of multi-stakeholder regulation – Social Media Councils - that would allow for better public oversight of the major social media companies, whilst avoiding the pitfalls of hard regulation, notably handing over control of users' speech to the State.

In this briefing, ARTICLE 19 builds on this work and examines how three dominant social media companies – Facebook, Twitter and Youtube – have responded to calls to address various forms of gender-based harassment and abuse in their community guidelines and practices. However, ARTICLE 19 is very mindful of the wider context of the problem of online gender-based harassment and abuse and the role of other players in the wider Internet ecosystem, such as private messaging services. The briefing first examines what are the responsibilities of the major/ dominant social media companies under human rights standards and how they implement their responsibilities in their community guidelines and practices. The briefing highlights the positive and negative aspects of these tools and approaches, and provides recommendations for improvement.

Human rights responsibilities of the private sector

International standards

Over the past decade, international and regional human rights bodies and special procedures have developed a body of recommendations on the responsibilities of social media companies to respect human rights. These include in particular the following recommendations:

- **The Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework** (the Guiding Principles) provide a starting point for articulating the role of the private sector in protecting human rights on the Internet.[6] The Guiding Principles recognise the responsibility of business enterprises to respect human rights, independent of State obligations or the implementation of those obligations. In particular, they recommend that companies should:[7]
    - Make a public statement of their commitment to respect human rights, endorsed by senior or executive-level management;
    - Conduct due diligence and human rights impact assessments in order to identify, prevent and mitigate against any potential negative human rights impacts of their operations;
    - Incorporate human rights safeguards by design in order to mitigate adverse impacts, and build leverage and act collectively in order to strengthen their power vis-à-vis government authorities;
    - Track and communicate performance, risks and government demands; and
    - Make remedies available where adverse human rights impacts are created.

- In his May 2011 report to the Human Rights Council, the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (UN Special Rapporteur on FoE) highlighted that – while States are the duty-bearers for human rights – Internet intermediaries also have a responsibility to respect human rights and referenced the Guiding Principles in this regard.[8] He also noted the usefulness of multi-stakeholder initiatives, such as the Global Network Initiative (GNI), which encourage companies to undertake human rights impact assessments of their decisions as well as to produce transparency reports when confronted with situations that may undermine the rights to freedom of expression and privacy.[9] He further recommended that intermediaries should only implement restrictions to these rights after judicial intervention; be transparent in respect of the restrictive measures they undertake; provide, if possible, forewarning to users before implementing restrictive measures; and provide effective remedies for affected users.[10] The UN Special Rapporteur also encouraged corporations to establish clear and unambiguous terms of service in line with international human rights norms and principles; and, to continuously review the impact of their services on the freedom of expression of their users, as well as on the potential pitfalls of their misuse.[11]
- In his April 2018 Report to the Human Rights Council, the UN Special Rapporteur on FoE called on companies to recognise that the authoritative global standard for ensuring freedom of expression on their platforms should be human rights law, not the varying laws of States or their own private interests, and that they should re-evaluate their content standards accordingly.[12] He also made it clear that companies should embark on radically different approaches to transparency at all stages of their operations, from rule-making to implementation and development of "case law" framing the interpretation of private rules.[13] Finally, he recommended that companies should open themselves up to public accountability, suggesting that this could take the shape of Social Media Councils.[14]
- In her 2013 Report, the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights (IACHR Special Rapporteur on FoE), also noted the relevance of the Guiding Principles[15] and further recommended, *inter alia,* that private actors establish and implement service conditions that are transparent, clear, accessible, and consistent with international human rights standards and principles; and ensure that restrictions derived from the application of the terms of service do not unlawfully or disproportionately restrict the right to freedom of expression.[16] She also encouraged companies to publish transparency reports about government requests for user data or content removal;[17] challenge requests for content removal or requests for user data that may violate the law or internationally recognised human rights;[18] notify individuals affected by any measure restricting their freedom of expression and provide them with non-judicial remedies;[19] and take proactive protective measures to develop good business practices consistent with respect for human rights.[20]
- In the 2016 report on Standards for a Free, Open and Inclusive Internet,[21] the IACHR Special Rapporteur on FoE recommended that, *inter alia,* companies make a formal and high-level commitment to respect human rights, and back this commitment up with concrete internal measures and systems; seek to ensure that any restriction based on companies' Terms of Service do not unlawfully or disproportionately restrict freedom of expression; and put in place effective systems of monitoring, impact assessments, and accessible, effective complaints mechanisms.[22] He also highlighted the need for companies' policies, operating procedures and practices to be transparent.[23]
- At European level, the Committee of Ministers of the Council of Europe, in its Recommendation on the protection of human rights with regard to social networking services, recommended that social media companies should respect human rights and the rule of law, including procedural safeguards.[24] Moreover, in its March 2018 Recommendation o**n the roles and responsibilities of internet intermediaries**, the Committee of Ministers adopted detailed recommendations on the responsibilities of internet intermediaries to protect the rights to freedom of expression and privacy and to respect the rule of law.[25] It recommended that companies should be transparent about their use of automated data processing techniques, including the operation of algorithms.

Civil society's recommendations

Additionally, recommendations that social media companies should respect international human rights standards have been made by a number of civil society initiatives:

- The **Manila Principles on Intermediary Liability**[26] make clear that companies' content restriction practices must comply with the tests of necessity and proportionality under human rights law;[27] and that intermediaries should provide users with complaints mechanisms to review decisions to restrict content made on the basis of their content restriction policies.[28]
- The **Santa Clara Principles on Transparency and Accountability in Content moderation** provide recommendations of practices that social media companies should undertake to provide transparency over their content moderation on their platforms. They call for companies to, at a minimum, disclose information about how many posts are removed, notify users about content removal, and give users meaningful opportunities to appeal take downs and have content restored.[29]
- Similarly, the **Ranking Digital Rights** project has undertaken a ranking of the major Internet companies by reference to their compliance with digital rights indicators. These include the following freedom of expression benchmarks: (i) availability of Terms of Service; (ii) terms of service, notice and record of changes; (iii) reasons for content restriction; (iv) reasons for account or service restriction; (v) notify users of restriction; (vi) process for responding to third-party requests; (vii) data about government requests; (viii) data about private requests; (ix) data about Terms of Service enforcement; (x) network management (telecommunication companies); (xi) identity policy (internet companies).[30]
- Finally, the **Dynamic Coalition on Platform Responsibility** is currently seeking to develop standard Terms and Conditions in line with international human rights standards.[31]

Addressing gender-based harassment and abuse by major social media companies

As noted earlier, online harassment and abuse against women journalists can take many forms, such as sending direct or indirect threats of physical or sexual violence, sending offensive messages, targeted harassment (often in the form of 'pile-ons', with multiple perpetrators), to privacy violations (such as "doxing, stalking, non-consensual dissemination of intimate sexual images").*

At present there is an ongoing and intense debate about the actions taken by three major/dominant social media companies – Facebook, Twitter and Youtube (Google) – to address gender-based harassment and abuse faced by women journalists on their platforms. The concerns have been raised about both how they define prohibitions of offending content in their rules, policies and community guidelines and how they enforce these rules in practice. In this section, ARTICLE 19 separately examines both aspects.

Community guidelines, policies and rules on online harassment and abuse

Overall, over the years, Facebook, Twitter and Youtube have had to take steps to address the serious criticisms directed at them by various human rights groups, journalists, lawmakers and institutions.[32] More generally, social media companies have been roundly condemned for profiting from polarising content because it drives greater engagement from users and it therefore generates greater advertising revenue.[33] Based on available information, it also appears that all three companies have responded by adopting a range of measures, which have become more sophisticated over time. These include the following measures.

All three companies state that various forms of harassment and abuse are not tolerated on their platforms and have dedicated sections in their rules, policies and community guidelines that can be applied to the offensive content in question. Each policy section describes its rationale and scope, as well as providing specific examples of the content and conducts that violate their policies and community guidelines. These policies apply to all users, they are not geared towards women journalists or women in particular but are applicable to various forms of harassment and abuse that they experience on these platforms.

* While there is no universally agreed terminology to capture this phenomenon and its different forms, in this briefing, ARTICLE 19 has employed the term "online harassment and abuse" as generic term to capture the type of offending conduct.

Various forms of online harassment and abuse are intrinsically broad concepts and States usually struggle to define many of them with sufficient precision in respective domestic legislation.[34] Definitions of such offending content in rules, policies and community guidelines of social media companies may sometimes reflect existing offences, such as threats of violence. These prohibitions in community guidelines comply with permissible restrictions on the right to freedom of expression, under international human rights law. However, in other cases, definitions go beyond that to include offensive and distasteful comments.

Further, given the variety of offending content that can be captured under the 'gender-based harassment and abuse', prohibitions can be found in more than one section of the community guidelines or rules. For example:

- On Facebook, some forms of gender-based harassment and abuse against women journalists can be removed under the section on "Violence and Criminal Behaviour" rules or under "Safety", "Objectionable Content" (such as "Hate speech" or "Violent and Graphic Content" or "Sexually Explicit" content) or Privacy rules. To a certain extent, these sections might overlap, e.g. 'credible violence' with 'hate speech.'
- On Twitter, various forms of gender-based harassment and abuse can fall under various sections of its Safety and Security rules (e.g. Privacy, Sensitive Content or Security).
- YouTube policies that can be applied to gender-based harassment and abuse include "Threats," "Hateful content," "Harassment and Cyberbullying," or "Privacy."

Assessment

Positive aspects

- Detailed policies, that include examples of offending content, is positive. In particular, it is useful when the rules and community guidelines list factors that companies take into account in assessing certain issues, e.g. the credibility of threats of violence or protection of private data.
- From the freedom of expression perspective, it is also positive when community guidelines distinguish between public figures and private individuals. For example, in their "Bullying and Harassment" section,[35] Facebook states that it distinguishes between ordinary users and public figures in order to enable discussion, including critical commentary of people who have a large public audience. For "public figures," it states that it removes attacks that are "severe" and "certain attacks where the public figure is directly tagged in the post or comment." On the other hand, it is not clear to what extent it might consider some women journalists, especially those with a large number of followers, as public figures or whether any exceptions are applied if they are targeted for journalistic activities. Similarly, Twitter includes public interest exceptions to its removal policy, if the content "contributes to understanding or discussion of a matter of public concern."[36]

- It is positive that companies regularly update and clarify their policies and community guidelines. For instance, Facebook regularly updates its policies on a wide range of content[37] and has taken steps to make those changes more apparent.[38] This is part of a wider stated commitment by Facebook in its Terms of Service to "better explain how we combat abuse and investigate suspicious activity."[39] On substance, these changes tend to widen the scope of speech, which is prohibited on the platform. For example, in June 2019, YouTube announced that it was changing its community guidelines to ban videos promoting the superiority of any group as a justification for discrimination against others based on *inter alia* gender and sexual orientation.[40] In practice, the purpose of this change was to crack down on videos promoting white supremacist and related ideologies, which had been outwardly tolerated up until then.[41]
- Another positive aspect is creating resources for users with examples, such as "test your knowledge" material on YouTube to help creators understand what content is permitted on the platform and some basic examples on how YouTube views the content in question.[42]

Negative aspects

- As noted earlier, different forms of gender-based harassment and abuse can fall under different categories in the community guidelines and policies. Although understandable, this could create some confusion and impede reporting from users, particularly in circumstances where the companies do not explain the distinction between overlapping concepts.
- Although policies and community guidelines are generally drafted in relatively plain language, they are also drafted in broad terms giving companies flexibility to interpret them according to their own needs. This results in inconsistent and sometimes apparently biased outcomes. In the absence of more concrete examples and explanations being given on how the guidelines are applied, it is difficult to know what content actually gets removed from these platforms. This is particularly the issue for gender-based harassment and abuse that women journalists face on these platforms.

Enforcement of community guidelines, policies and rules

While social media companies continue to remove vast amounts of content,[43] they have also developed more sophisticated responses to deal with 'problematic' content. In particular, as noted by the UN Special Rapporteur on FoE, they can "restrict its virality, label its origin, suspend the relevant user, suspend the organisation sponsoring the content, develop ratings to highlight a person's use of prohibited content, temporarily restrict content while a team is conducting a review, preclude users from monetising their content, create friction in the sharing of content, affix warnings and labels to content, provide individuals with greater capacity to block other users, minimise the amplification of the content, interfere with bots and coordinated online mob behaviour, adopt geolocated restrictions and even promote counter-messaging".[44]

Concerns about the lack of a consistent enforcement of gender-based harassment and abuse of women journalists focus on three aspects - reporting or identification mechanism, appeals and complaints of harassment and abuse, and remedies.

Assessment

Positive aspects

- All three companies provide various reporting mechanisms that are available to women journalists experiencing gender-based harassment. Their policies explain how and where to initiate a report or give users an option to specify which policy/community guidelines have been violated. For example, Facebook allows to report particular accounts, pages, posts and so on,[45] and has communication tools that allow users to request other users to take content down (social reporting).[46] YouTube has a dedicated webpage outlining enforcement and reporting options;[47] users can report videos,[48] abusive users[49], legal complaints[50], privacy violations,[51] with other additional reporting tools being available to capture the whole range of content that users may find problematic.[52] In other words, different report forms are available depending on the type of complaint at issue. Twitter has a dedicated page on reporting offending content, which includes the specific types of content and actions to be reported, as well as a form to add the required information to report the various elements of the incident.[53]
- All three companies continue to invest in the upgrade of their systems, including those that help them detect more offending content on their platforms, including in a wider range of languages.[54] They also use a "trusted flagger" system to fast-track reports of violations of their policies and community standards.[55]
- Additionally, the three companies have also been developing new features giving more control to users over what they see. They have also implemented more user-centred approaches to let users manage their privacy settings, among others. These include, for example, Twitter's 'hide replies' feature[56] or on Facebook a feature to hide the content on newsfeed or blocking users, or "unfriend" them.[57] All three companies have put in place some appeals mechanisms for the users in cases of removals on the content on the basis of their community standards.[58] In addition to content removal, the companies have also adopted penalties against repeated offenders, e.g. by applying community guidelines strikes.[59] Beyond redress in individual cases, appeals can also enable the identification of systemic problems. For instance, a volume of complaints or appeals on a given topic may point to deficiencies in the algorithms used to identify 'problematic' content.
- In addition to enhancing their policies to respond against problematic situations for their users, all three companies have undertaken different approaches and measures to support women journalists on their platforms. This has included initiatives to promote their stories and content, hosting events and trainings to explain their policies and community standards, as well as their security and moderation tools created to improve the perception of security women can face online. For example, @twitterwomen promotes content related to women's inequality, empowerment and non-discrimination, including the videos of #Herstory campaign, which Twitter develops alongside UN bodies, civil society and other stakeholders. Facebook's Journalism Project partnered with trusted reporters to report and escalate cases of journalists being harassed on their platform.[60] Twitter also ran regional campaigns to promote women's rights and counter narratives for discrimination, responding to constant calls to consider context in their responses.[61]
- In the case of Facebook, it has gone a step further with the creation of an external body, the Facebook Oversight Board.[62] As such, this new governance structure could be well-placed to identify and help Facebook address systemic issues, including online harassment and abuse of women journalists on the platform.

Negative aspects

- Although enforcement mechanisms are laid out, they are not always easy to find, and significant shortfalls in enforcement remain. Many studies document that when women journalists report gender-based harassment and abuse to the companies through these mechanisms, the companies' responses are often lacking or are inconsistent with stated objectives of the policies.
- Overall, there is a lack of transparency over the actions taken in response to gender-based harassment and abuse. Transparency reports do not capture detailed information about reports, appeals and actions taken under applicable policies on harassment, abusive behaviour, bullying and violence, or private information or privacy violations. Transparency reports do not cover all the categories of the respective policies. Some of them remain broad or provide insufficient information to examine the scale and intensification of the problem.
- There is also a lack of transparency in relation to the use of algorithms in order to detect some forms of online harassment and abuse which means that the companies are more likely to be prone to gender bias. It is also unclear how algorithms can be trained to take into account various free speech concerns – such as context (e.g. political, social, cultural), if at all.
- It is also not clear how the companies apply the exceptions to the content of women journalists. For instance, to what extent are some reports of harassment and abuse not acted upon for the reasons of considering this content as acceptable criticisms (Facebook) or "hyperbolic speech" (Twitter), and to what extent this accounts for inconsistencies of removals.
- Relatively little information is available about the way in which these companies use machine learning for the purposes of content flagging. In particular, it is unclear what criteria are used to flag particular pieces of content. This is especially concerning in relation to gender-based harassment and abuse given that automated systems are notoriously bad at understanding context. Equally, these three companies provide limited meaningful information about their 'trusted flagger' systems and the extent to which content flagged through these mechanisms is subject to adequate review.

Recommendations

In light of the foregoing, ARTICLE 19 suggests that the three major social media companies should improve their policies and practices in response to gender-based harassment and abuse on their platforms. In particular, we make the following recommendations:

- Social media companies should voluntarily accept and apply all core international human rights and women's rights instruments with a view of contributing to universal human rights protection and elimination of discrimination of women on their platforms.
- In order to respond to gender-based harassment and abuse, the three social media companies should consider developing dedicated sections in their policies and community guidelines. This could include consolidated overview of what parts of policies could be applied to the content and organised in a way that could be easily found in one place.
- The policies and guidelines should be easily accessible and available in local languages.
- Social media companies should further develop and provide case studies or more detailed examples of the way in which they apply their policies to gender-based harassment and abuse and how the existing policies are applied in practice.
- Social media companies should publish information about the methods and internal processes for the elaboration of policies and community guidelines, and to what extent they apply gender mainstreaming to the policy development. The policy development should continue to include consultations with a broad range of women's rights and gender experts and civil society.
- Social media companies should undertake a human rights and gender discrimination impact assessment that identifies, prevents and mitigates any negative impact of their operations on the rights to freedom of expression, privacy, participation and non-discrimination of women and women journalists.
- Social media companies should ensure that their appeals process complies with the Manila Principles on Intermediary Liability and Santa Clara Principles, particularly as regards to notice, the giving of reasons of their decisions, and appeals processes;
- Transparency should be a basic requirement that pervades everything that social media companies do, including in area of gender-based harassment and abuse. In particular, social media companies should:
    - Be more transparent about their use of algorithms to detect various types of gender-based harassment and abuse, provide more essential information, and explain to the public how their algorithms are used to present, rank, promote or demote content. They should also publish information about the way in which their algorithms operate to detect allegedly harmful content, including gender-based harassment and abuse under their community standards. In particular, this should include information about rates of false negatives/false positives and indicators, if any, to assess content that is likely to become viral, e.g. by reference to exposure to a wider audience.
    - Publish detailed information consistent with the Santa Clara Principles about gender-based harassment and abuse on their platforms. It is particularly important not to limit statistical information to removal of content, but also include data about the number of appeals processed and their outcome. Any transparency reporting requirements should aim to provide far more qualitative analysis of content moderation decisions.
    - Publish detailed information about "trusted flagger" schemes, including the roster of trusted flaggers, how and the criteria under which they have been selected and any 'privileges' attached to such status.
- Social media companies should improve their internal redress mechanisms, respecting due process safeguards. These should also be able to address any refusal to remove content, such as gender-based harassment and abuse, that is arguably in breach of the companies' community standards.
- Social media companies should notify their decisions to affected parties and give sufficiently detailed reasons for the actions they take against particular content or accounts. They should also provide clear information about any internal complaints mechanisms.
- Social media companies should give greater access to datasets, to independent researchers, whether academics, journalists or otherwise, in order for them to verify that the companies' systems and algorithms are operating as the company says it does.
- Social media companies should consider further partnering with women journalists and civil society groups to develop practical strategy of research-focused and community-lead solutions on gender-based harassment and abuse. They should support journalism initiatives promoting gender inclusion and gender mainstreaming programmes. They should also further develop, and strengthen and amplify, their efforts to counter-narratives against gender inequality and positive measures to promote gender diversity on their platforms.
- Social media companies should consider joining multi-stakeholder regulatory bodies such as social media councils, that would allow better public oversight of their practices, including in the area of gender-based harassment and abuse. In particular, the Facebook Oversight Board should consider gender harassment and abuse cases in their case work and provide the appropriate guidance on how better and more effectively address this issue, in line with international freedom of expression standards.

About ARTICLE 19

ARTICLE 19: Global Campaign for Free Expression (ARTICLE 19), is an independent human rights organisation that works around the world to protect and promote the rights to freedom of expression and information. It takes its name and mandate from Article 19 of the Universal Declaration of Human Rights which guarantees the right to freedom of expression.

ARTICLE 19 has produced a number of standard-setting documents and policy briefs based on international and comparative law and best practice on issues concerning the right to freedom of expression. Increasingly, ARTICLE 19 is also examining the role of international internet technical standard-setting bodies and internet governance bodies in protecting and promoting freedom of expression.

If you would like to discuss this brief further, or if you have a matter you would like to bring to the attention of ARTICLE 19, you can contact us by e-mail at info@article19.org.

Endnotes

1 *C.f.* APC, Technology-related violence against women – a briefing paper, 2015. Council of Europe (CoE), CoE Factsheet Hate Speech, 2017; European Commission, What is gender-based violence?, 2018; European Union Agency for Fundamental Rights (FRA), Violence against women: an EU-wide survey, 2014; Human Rights Council (HRC), Report of the Special Rapporteur on violence against women (Special Rapporteur on VAW), its causes and consequences on online violence against women and girls from a human rights perspective, July 2018; OSCE Representative on Freedom of the Media (OSCE RFoM), New Challenges to Freedom of Expression: Countering Online Abuse of Women Journalists, 2016; OSCE RFoM, Legal Responses to Online Harassment and Abuse of Journalists: Perspectives from Finland, France and Ireland, 2019.

2 See, e.g. UNGA, The safety of journalists and the issue of impunity, A/72/290, 4 August 2017, para 9; IACHR, Annual Report 2016, Report of the Office of the Special Rapporteur for Freedom of Expression, Chapter IV, OEA/Ser.L/V/II.Doc. 22/17, 15 March 2017; Committee to Protect Journalists, The threats follow us home: Survey details risks for women journalists in U.S., Canada, 4 September 2019; Amnesty International, Toxic Twitter, Triggers of Violence Against Women on Twitter, 2018; Trollbusters & IWMF, Attacks and Harassment, The impact on Female Journalists and Their Reporting, 2018; AMWIK & ARTICLE 19, Women's Journalists Digital Security, 2016; DRF, Female Journalists in New Media, Experiences, Challenges and Gender Approach, March 2019; Fundacion Karisma, Misogyny in the Internet, Colombia, 2016, p. 4- 5; IMS, The safety of female journalists: Breaking the cycle of silence and violence, IMS- booking series-2019, September 2019; Reporters Without Borders, Online Harassment of Journalists, Attack of the Trolls, August 2018.

3 UNESCO, Intensified attacks, New Defenses, Development in the Fight to Protect Journalists and End Impunity, 2019.

4 Special Rapporteur for Freedom of Expression, Female journalists and Freedom of Expression, Inter-American Commission for Human Rights, CIDH/RELE/INF.20/18, 31 October 2018, para 3.

5 *Ibid.*

6 Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect and Remedy' Framework, developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, 7 April 2008, A/HRC/8/5A/HRC/17/31. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

7 *Ibid.*, Principle 15.

8 The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 45.

9 *Ibid.* para 46.

10 *Ibid.*, paras 47 and 76.

11 *Ibid.*, paras 48 and 77.

12 *Op. cit.,* para 70

13 *Ibid.,* para 71

14 *Ibid.,* para 72.

15 OAS Special Rapporteur on FOE, Freedom of Expression and the Internet, 2013. The report noted that "the adoption of voluntary measures by intermediaries that restrict the freedom of expression of the users of their services - for example, by moderating user-generated content - can only be considered legitimate when those restrictions do not arbitrarily hinder or impede a person's opportunity for expression on the Internet;" paras 110-116.

16 *Ibid.*, paras 111-112.

17 *Ibid.*, para 113.

18 *Ibid.*, para 114.

19 *Ibid.*, para 115.

20 *Ibid.*, para 116.

21 OAS Special Rapporteur on FOE, Standards for a Free, Open and Inclusive Internet, 2016, paras 95-101.

22 *Ibid.*, para 98.

23 *Ibid.*, para 99.

24 Committee of Ministers of Council of Europe, Recommendation CM/Rec (2012)4 of the Committee of Ministers to Member States on the protection of human rights with regard to social networking services, *adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies.* These recommendations were further echoed in the Committee of Ministers Guide to Human Rights for Internet users, which states "your Internet service provider and your provider of online content and services have corporate responsibilities to respect your human rights and provide mechanisms to respond to your claims. You should be aware, however, that online

service providers, such as social networks, may restrict certain types of content and behaviour due to their content policies. You should be informed of possible restrictions so that you are able to take an informed decision as to whether to use the service or not. This includes specific information on what the online service provider considers as illegal or inappropriate content and behaviour when using the service and how it is dealt with by the provider;" Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, p. 4.

25 Recommendation CM/Rec (2018) 2 of the Committee of Ministers to member states on the roles and responsibilities of internet intermediaries, *adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies.*

26 The Manila Principles on Intermediary Liability, March 2015. The Principles have been endorsed by over 50 organisations and over a 100 individual signatories.

27 *Ibid.,* Principle IV.

28 *Ibid.*, Principle V c).

29 The Santa Clara Principles on Transparency and Accountability in Content Moderation, February 2018.

30 Ranking Digital Rights, Corporate Accountability Index, 2015 Research Indicators.

31 Dynamic Coalition on Platform Responsibility is a multi-stakeholder group fostering a cooperative analysis of online platforms' responsibility to respect human rights, while putting forward solutions to protect platform-users' rights.

32 *Ibid.*

33 See, e.g. The Independent, Social media companies 'actively' serve up extremist material to users to maximize profits, MPs say, 24 April 2019.

34 See for instance the Nova Scotia Cyber Safety Act 2013, which sought define bullying. The Act was struck down by the Supreme Court of Nova Scotia: see *Crouch v. Snell,* 2015 NSSC 340, p. 47 ff.

35 Facebook Community Guidelines, Bullying and Harassment.

36 https://help.twitter.com/en/rules-and-policies/public-interest.

37 See, e.g. Facebook Product Policy Forum, 15 November 2018.

38 See for instance Facebook's Recent Updates page, which allows users to see recent amendments made to its community standards.

39 See Facebook Terms Update.

40 See YouTube, Our ongoing work tackle hate, 5 June 2019

41 See The Verge, YouTube just banned supremacist content, and thousands of channels are about to be removed, 05 June 2019.

42 See Youtube, Harassment and Cyberbullying Policy.

43 See for instance, European Commission, EU Code of Conduct on countering illegal hate speech, Fourth monitoring round, February 2019.

44 See Special Rapporteur on FOE, A/74/486, para. 51. See also YouTube, *op. cit.* See also Facebook Rules for Monetisation; Facebook's initiatives are available from here.

45 See, How to Report Things on Facebook.

46 See, What is social reporting on Facebook?

47 YouTube, Reporting and Enforcement.

48 *Ibid.*

49 See YouTube, Reporting and Enforcement.

50 *Ibid.* More information is available at Removing Content from YouTube.

51 YouTube Privacy Complaint Process.

52 YouTube, Other reporting options.

53 See Twitter, Someone on Twitter is engaging on abusive or harassing behavior?

54 See YouTube, *op. cit.* See also Instagram's investment in AI to detect 'bullying content': Wired, Twitter and Instagram unveil new ways to combat hate, again, 07 November 2019.

55 See for instance European Commission Code of Conduct on Countering Illegal Hate Speech, Factsheet, February 2019.

56 The Verge, Twitter rolls out 'hide replies' to let you tame toxic discussions, 19 September 2019.

57 Facebook, How do I report inappropriate or abusive things on Facebook (example: nudity, hate speech, threats)?

58 For the shortfalls of these appeals mechanisms, see ARTICLE 19's Sidestepping Rights policy, and our Missing Voices campaign.

59 YouTube, Community Guidelines Strike basics.

60 Faceook Journalism Project.

61 In Mexico and Colombia #fuerzaenmivoz campaign was created to strengthening women's voices, disseminating Twitter policies and security tools and promoting the initiatives of civil society addressing online harassment against women in the platform. In 2019, Twitter supported a campaign with a group of organisations in Mexico working on online harassment against women #internetesnuestra. In India #Positionofstrength aimed at helping to bridge the gender equality gap online. Twitter Latin American section is working with the OAS on a Digital Security Guide that will be useful for women journalists.

62 For more details about the Facebook Oversight Board, see here. For ARTICLE 19's concerns about it, see here.

article19.org