SCOPING PAPER

From visibility to vulnerability: when social media fails human rights defenders

March 2026

UN SPECIAL RAPPORTEUR ON HUMAN RIGHTS DEFENDERS

IN SOLIDARITY & HOPE

TABLE OF CONTENTS

INTRODUCTION

Since their creation, social media platforms have become indispensable tools for those who defend human rights. People turn to them to gather information, raise awareness of successes and challenges, share reports and documentation, build and organise communities and networks, and amplify marginalised voices.

In the course of her work, however, and in consultations with people defending human rights, the UN Special Rapporteur on the situation of human rights defenders ('the Special Rapporteur') has found that these same platforms significantly hinder human rights activism. Through at best faulty moderation practices, algorithmic suppression, inadequate responses to online harassment and compliance with repressive demands from States, social media platforms contribute, intentionally or otherwise, to silencing, deplatforming and endangering human rights defenders (HRDs). [1]

Given these concerns, the Special Rapporteur is publishing this scoping paper examining the impact of social media on the right to defend human rights. Its findings underscore the urgent need for companies behind social media platforms to uphold their human rights responsibilities, and for States to respect their human rights obligations and ensure that digital spaces remain safe and enabling environments for those who defend human rights.

[1] The Special Rapporteur defines a human rights defender as a person who, individually or with others, acts peacefully to promote or protect human rights in accordance with the UN Declaration on Human Rights Defenders.

METHODOLOGY

This scoping paper is based on research and consultations [2] conducted by the office of the UN Special Rapporteur on human rights defenders with civil society organisations working alongside people defending and promoting human rights across all regions.

The scope of the paper was limited by time, resources and access, and does not claim to be exhaustive. Rather, it reflects the experiences of the organisations who were able to contribute to the consultation process. It should be read as an indicative mapping of key patterns and risks, rather than a comprehensive assessment of all platform-related impacts on the right to defend human rights.

The UN Special Rapporteur on Human Rights Defenders shared her concerns regarding the policies of Meta Inc. and X Corp. and their impact on the right to defend human rights in two official communications sent to the companies on 30 December 2025, informing them that the communications and their responses would be included in this scoping paper. No response had been received as of the publication of this paper.

[2] Consultations were engaged on under the Chatham House rules.

1. INTERNATIONAL LAW & STANDARDS

Adopted by consensus at the UN General Assembly in December 1998, the UN Declaration on Human Rights Defenders [3] ('the HRD Declaration') declares in article 1 that "Everyone has the right, individually and in association with others, to promote and to strive for the protection and realization of human rights and fundamental freedoms at the national and international levels." This is reinforced in article 12.1, which declares that "Everyone has the right, individually and in association with others, to participate in peaceful activities against violations of human rights and fundamental freedoms", while article 6 holds that everyone has the right to share information, views and knowledge on all human rights, and to draw public attention to those matters.

Article 6 of the HRD Declaration builds on article 19 of the International Covenant on Civil and Political Rights, which protects the right to freedom of opinion and the right to freedom of expression, defined as the "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice". Freedom of expression may be restricted pursuant to article 19(3), which requires that restrictions must be clear and precise ("provided by law"), and are necessary to achieve "respect for the rights or reputations of others", "the protection of national security or of public order (ordre public)", or of "public health and morals".

States are the primary duty bearers under international human rights law, and must protect against human rights violations within their territory or jurisdiction by third parties, including companies. Under the United Nations Guiding Principles on Business and Human Rights, however, companies themselves have the responsibility to: a) avoid causing or contributing to adverse human rights impacts; and (b) to prevent or mitigate such impacts directly linked to their operations, products or services from their business relationships. To meet their responsibility to respect human rights, enterprises should have a) a policy commitment to meet their responsibility to respect human rights; b) a human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights; and c) processes to enable the remediation of any adverse human rights impacts they cause or contribute to. The responsibility to respect human rights constitutes a global standard of conduct applicable to all businesses and exists irrespective of the capacity or willingness of States to meet their own human rights obligations.

---

[3] A/RES/53/144, available in multiple languages at: https://www.ohchr.org/en/special-procedures/sr-human-rights-defenders/declaration-human-rights-defenders-different-languages

2. CONTENT MODERATION FAILURES & COMPROMISED VISIBILITY

Human rights defenders across all regions report a range of content moderation failures on social media platforms, with severe impacts on the right to defend human rights. These include the removal of legitimate content, shadowbanning, arbitrary takedowns, inconsistent enforcement of community guidelines, and an overall lack of transparency on internal processes.

2.1 REMOVAL OF LEGITIMATE HUMAN RIGHTS CONTENT

In information received in preparation of this paper, several organisations reported incidents in which legitimate human rights content was removed for allegedly violating platform policies.

One organisation recorded that posts related to human rights in **Palestine** have been taken down by Meta-owned platforms under policies relating to "Dangerous Individuals and Organisations" or "Violent Content," often without clear explanation or recourse. These included documentation of extrajudicial killings, arbitrary arrests and peaceful protests, and have particularly impacted human rights defenders and journalists, some of whom received no notification or explanation for the censorship. HRDs also reported inconsistencies: for example, posts criticizing hateful or extremist content were taken down while the original hateful or extremist content they were critiquing remained online [4]. A digital rights organisation also consistently documented the censorship and removal of content relating to human rights in Palestine by social media platforms, including Facebook, X and Instagram [5]. Similarly, another organisation reported takedowns of content documenting war casualties in **Syria**, including posts that could have contributed to international accountability efforts. [6] The reasons generally provided by platforms for the removal of such content include glorifying dangerous individuals or organisations, violent or graphic content. Photos of victims, especially children, have been flagged as child exploitation.

[4] Submission (A) received by the Special Rapporteur
[5] Submission (B) received by the Special Rapporteur
[6] Submission (C) received by the Special Rapporteur

Similar issues are seen in other regions and on other platforms. On TikTok, content related to human rights abuses in **Xinjiang** and **Tibet** has been repeatedly flagged, removed or restricted in recent years for allegedly going against community guidelines. [7][8] A Belarusian human rights organisation reported that their publications about victims of political repression in **Belarus** were sometimes flagged as inappropriate or political content on Facebook and Instagram, with appeals yielding varied outcomes, with little explanation [9]. In **Cambodia**, journalists have seen the suspension of their Facebook accounts after publishing content on land disputes, corruption, and deforestation, and during the 2023 national elections several independent media accounts were blocked, severely restricting access to independent information. [10] One regional human rights organisation reports that one of their paid campaigns on ASEAN and human rights was flagged as inappropriate or sensitive and taken down from social media platforms in the **Philippines** and **Cambodia**, without clear justification and despite running smoothly in other Asian countries. [11]

2.2 SHADOWBANNING & VISIBILITY SUPPRESSION

Shadowbanning, where users' posts become invisible to others without notice, is also widely reported by human rights defenders.

Several organisations have found that Palestinian human rights defenders' accounts and content related to human rights in Palestine had been notably shadowbanned by Meta's platforms, Facebook and Instagram, with these practices increasing since the start of the genocide being perpetrated by Israel [12], but also reported before it. [13] This typically means that the information shared by these HRDs, whether to document human rights violations on the ground, raise international awareness or fundraise to support victims, will be seen by less people, preventing their work from being as effective. Affected users receive no notification or justification. Such cases can be difficult to understand and verify, as Meta provides little transparency about algorithmic enforcement [14][15].

---

[7] The Guardian, "Revealed: how TikTok censors videos that do not please Beijing", September 2019 https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing
[8] Radio Free Asia, "TikTok deletes videos related to Uyghur human rights violations", November 2024 https://www.rfa.org/english/uyghur/2024/11/05/uyghur-tiktok-censors-abroad/
[9] Submission (D) received by the Special Rapporteur
[10] Submission (E) received by the Special Rapporteur
[11] Submission (F) received by the Special Rapporteur
[12] Amnesty International, "Israel/OPT: Post-ceasefire: Israel's genocide in the occupied Gaza strip continues", 27 November 2025 https://www.amnesty.org/en/documents/mde15/0527/2025/en/
[13] SMEX, "Israel's Fierce Attack on Palestinian Narrative and Content", June 2022 https://smex.org/israels-fierce-attack-on-palestinian-narrative-and-content/
[14] Submission (A) received by the Special Rapporteur
[15] Submission (B) received by the Special Rapporteur

Organisations working on freedom of expression have found that shadowbanning has disproportionately affected feminist women human rights defenders in **Lebanon**, who report having their posts, particularly in Arabic, removed while hateful content targeting women remains. [16][17] WHRDs advocating for sexual and reproductive rights are also affected. [18]

In an example documented by one of the submitting organisations, a **Chilean** collective of migrant journalists reported that posts covering an anti-racism event on Instagram were not appearing on timelines. Despite repeated requests for support, no clear explanation was provided for why the posts were hidden, and visibility was only restored after prolonged exchanges. The lack of transparency left the collective uncertain about the reasons behind the restriction and discouraged them from posting further. [19]

Shadowbanning has also been used to the detriment of those involved in LGBTIQ+ rights advocacy. In 2024, policy changes [20] at Meta algorithmically limited the reach of "political content" on Instagram and Threads, without clarifying exactly what that meant and how it would be applied. In effect, this change dramatically reduced the reach of credible, informative posts from several LGBTIQ+ rights Instagram accounts [21], including those of **US** organisations Human Rights Campaign and GLAAD. This development compounded Meta's history of shadowbanning LGBTIQ+ content, often by incorrectly flagging it as sexually explicit. [22]

2.3 ALGORITHMIC & AUTOMATED MODERATION FAILURES

Many of the aforementioned moderation failures seem to stem from the use of automated or AI-driven moderation systems. According to one non-governmental organisation, such technologies are often insufficiently advanced to be able to reliably distinguish between

---

[16] Submission (B) received by the Special Rapporteur
[17] SMEX, "Research Report: Confronting Structural Silencing: Challenges and Resistance Among Digital Feminist Activists in Lebanon", April 2025 https://smex.org/research-report-confronting-structural-silencing-challenges-and-resistance-among-digital-feminist-activists-in-lebanon/
[18] Submission (C) received by the Special Rapporteur
[19] Submission (A) received by the Special Rapporteur
[20] Instagram, "Continuing our Approach to Political Content on Instagram and Threads", February 2024 https://about.instagram.com/blog/announcements/continuing-our-approach-to-political-content-on-instagram-and-threads
[21] Accountable Tech, "Meta's Political Content Limit Causes Steep Drop in Reach for Accounts", July 2024 https://accountabletech.org/research/metas-political-content-limit-causes-steep-drop-in-reach-for-accounts/
[22] Human Rights Campaign, "Meta's New Policies: How They Endanger LGBTQ+ Communities and Our Tips for Staying Safe Online", January 2025 https://www.hrc.org/news/metas-new-policies-how-they-endanger-lgbtq-communities-and-our-tips-for-staying-safe-online

lawful and unlawful content, leading to the removal of vast amounts of legitimate expression, including by human rights defenders. [23]

Another organisation cited an example where automated filters misclassified legitimate posts documenting repression during the 2024 **Venezuela** election as 'incitement to violence', leading to them being blocked. [24] They also reported that a reproductive rights collective in **Brazil** had their account permanently suspended for sharing information on Misoprostol, a medicine for ulcers that had long been used by Brazilian women to induce abortions. [25]

2.4 ACCOUNT SUSPENSIONS & MASS REPORTING

People and civil society organisations defending human rights also face account suspensions and restrictions triggered by coordinated mass reporting campaigns carried out by trolls, state-linked actors, right-wing and anti-rights groups.

In **Central America**, one regional organisation noted the use of bots and coordinated abuse campaigns to generate false reports, leading to takedowns and account loss for people defending human rights. These incidents often occur without warning or rationale, and appeal processes are opaque or inaccessible, particularly for users without direct access to the platforms' Trusted Partner channels. [26] In **Bangladesh**, feminist, queer and secular human rights defenders experienced similar issues after coordinated mass reporting campaigns resulted in automatic suspensions of their accounts on Facebook and Instagram, without any meaningful recourse. [27] In **Iran**, one organisation reported that Iranian women's rights activists' Instagram accounts were disabled or shadow-banned after mass reporting attributed to Iranian authorities. [28] In **Belgium**, an Instagram account dedicated to anti-racism education and administered by a young WHRD has been deactivated seven times between 2023 and 2025 due to mass reporting from anti-rights groups, leading her to ultimately turn the account private and stop posting on it. [29] In **Malaysia**, one organisation recounts that the Facebook and Instagram accounts of a women's rights organisation and two of its staff members were suspended with no explanation following a wave of online harassment. Although their accounts were later reinstated, the incident affected their communications work and the organisation had to undergo an internal review on its digital strategy. [30]

---

[23] Submission (G) received by the Special Rapporteur
[24] Submission (A) received by the Special Rapporteur
[25] Submission (A) received by the Special Rapporteur
[26] Submission (H) received by the Special Rapporteur
[27] Submission (A) received by the Special Rapporteur
[28] Submission (B) received by the Special Rapporteur
[29] https://www.instagram.com/p/DHvS29aomok/
[30] Submission (G) received by the Special Rapporteur

3. COMPLIANCE WITH REPRESSIVE GOVERNMENT REQUESTS

As people using social media platforms for the promotion and protection of human rights face interference from governments seeking to silence dissent and control narratives [31], some companies frequently comply with government censorship requests that conflict with international human rights law or facilitate state surveillance.

3.1 COMPLIANCE WITH CENSORSHIP REQUESTS

In information received by the Special Rapporteur in preparation of this paper, several organisations highlighted instances where Governments exerted pressure on social media platforms to align their content moderation practices with specific political viewpoints or to directly censor human rights defenders. Such pressure can take the form of threats of throttling (i.e. limiting the bandwidth made available to the platform), advertisement bans or legal sanctions such as fines.

In **Türkiye**, where social media platforms have become the primary source of news and a vital tool for independent journalism, the Government has in recent years tightened its grip on social media platforms through a series of amendments to the Internet Law. One organisation reported very high rates of compliance with requests for content removal from the Turkish Government across major social media platforms: 95% for TikTok, 85% for X and 79% for Instagram in the second half of 2024) [32]. Around the time of the arrest of opposition leader Ekrem İmamoğlu in March 2025, the Government ordered the blocking of hundreds of accounts, including of human rights defenders, journalists, student groups, civil society organisations and women's rights groups. While X complied with many of these blocking orders because of the threat of throttling, Meta stated that it took no action [33] and reportedly faced a substantial fine [34]. The submitting organisation

[31] See A/HRC/38/35, paragraph 13, A/80/341, paras. 12 and 76.
[32] Submission (G) received by the Special Rapporteur
[33] Meta, "Case Studies - Content alleged to violate Turkish Law", March 2025 https://transparency.meta.com/reports/content-restrictions/case-studies
[34] Associated Press, "Meta is fined by Turkey after refusing to restrict content on Facebook and Instagram" April 2025 https://apnews.com/article/meta-facebook-instagram-turkey-erdogan-59f16571e884fd6f0f9597a72771131f

further reported that Meta [35], TikTok [36], YouTube [37] and X [38] were all pressured to appoint in-country representatives under threat of escalating legal repercussions, despite warnings that these would give the Turkish government greater control over their platforms. [39]

In **India**, *Bolta Hindustan*, an independent media platform covering social issues followed by 300,000 subscribers, was banned from YouTube in April 2024 after a notification from the Ministry of Information and Broadcasting. Google's legal team did not initially provide the media platform with any reason for the ban. [40] Press freedom advocates denounced the ban, and the account was later reinstated by YouTube. [41] Similarly, Twitter (now X) has been accused of censoring Kashmiri voices at the Indian Government's request. In 2019, the Committee to Protect Journalists reported that Twitter prevented India-based users from accessing thousands of tweets sharing information about the situation in **Jammu and Kashmir** during media blackouts and Internet shutdowns, following legal notices sent by the Indian government. [42]

In **Pakistan**, livestreams of Baloch protests on Facebook and YouTube have faced slower Internet speed or were removed without explanation. [43]

3.2 ENABLING GOVERNMENT SURVEILLANCE

Another concern is the lack of transparency and control over personal data shared by human rights defenders on their social media pages. In cases documented by one submitting organisation, user data belonging to HRDs had reportedly been provided to State authorities without their knowledge or consent. In **Guatemala** for example, Meta shared the user information of a human rights defender, not only related to the account requested by authorities, but from all their social media accounts. The defender was never

[35] Facebook, "An Update on Facebook in Turkey", January 2021
https://about.fb.com/news/2021/01/an-update-on-facebook-in-turkey/
[36] TikTok, "An update on TikTok in Turkey", January 2021
https://newsroom.tiktok.com/an-update-on-tiktok-in-turkey?lang=en
[37] YouTube, "An update on YouTube in Turkey", December 2020
https://blog.youtube/news-and-events/update-youtube-turkey/
[38] Balkan Insight, "X Appoints Representative in Turkey, Bowing to Govt Pressure", May 2024
https://balkaninsight.com/2024/05/23/x-appoints-representative-in-turkey-bowing-to-govt-pressure/
[39] Submission (G) received by the Special Rapporteur
[40] Submission (F) received by the Special Rapporteur
[41] Medianama, "YouTube Reinstates Blocked and Demonetised Channels", May 2024
https://www.medianama.com/2024/05/223-youtube-reinstates-blocked-demonetised-channels/
[42] Committee to Protect Journalists, "India uses opaque legal process to suppress Kashmiri journalism, commentary on Twitter", October 2018
https://cpj.org/2019/10/india-opaque-legal-process-suppress-kashmir-twitter/
[43] Submission (F) received by the Special Rapporteur

officially notified about this and only learnt about it after a trusted source with the knowledge of the facts informed them. [44] This raises serious concerns, particularly in contexts where Governments are known to target dissenting voices. The processes by which platforms assess and respond to requests from national authorities remain opaque. As a result, people defending human rights may be subjected to surveillance, intimidation or reprisals without being aware that their private data has been accessed or disclosed, and without access to effective remedy.

4. DIGITAL ATTACKS & PLATFORMS' INADEQUATE RESPONSE

Social media platforms have reproduced the intersecting patterns of harassment and abuse that many human rights defenders already face in their work, in a wide range of digital forms. Online abuse has become both more prevalent and more coordinated, with human rights defenders consistently targeted by harassment, smears and disinformation campaigns that spread across social media platforms and often include the use of automated "bots" [45]. These tactics, sometimes supported by artificial intelligence tools, are becoming increasingly prevalent. In the face of this, platforms have failed to provide adequate mechanisms to respond or protect users affected by such attacks.

As an example of the extent of the problem, Global Witness noted in a recent report [46] that 92% of land and environmental human rights defenders they surveyed had experienced online abuse from a range of actors including individuals, bots and coordinated campaigns. Globally, HRDs received more abuse on Facebook than on any other social media platform, followed by X (Twitter) and Instagram, according to the study. [47]

---

[44] Submission (A) received by the Special Rapporteur
[45] Resolution 58/23 adopted by the Human Rights Council on 4 April 2025 https://docs.un.org/en/A/HRC/RES/58/23
UN Women, "Tipping point: The chilling escalation of online violence against women in the public sphere", 2025, https://www.unwomen.org/en/digital-library/publications/2025/12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai
Front Line Defenders, "Global Analysis 2024/2025", https://www.frontlinedefenders.org/sites/default/files/1609_fld_ga24-5_output.pdf
[46] Global Witness, "Toxic platforms, broken planet", 16 July 2025 https://globalwitness.org/en/campaigns/digital-threats/toxic-platforms-broken-planet/
[47] Global Witness, Ibid

4.1 ONLINE GENDER-BASED & IDENTITY-BASED ABUSE

People who already face oppression in one form or another typically face greater risks and harsher abuses when using social media platforms to defend and promote human rights. In all regions of the world, HRDs, especially women and LGBTIQ+ defenders, are facing disinformation, smear campaigns, visual defamation, deepfakes, blackmail, threats, harassment, doxxing and impersonation, which are often ignored or inadequately addressed by platforms. Such attacks are aimed at delegitimising, intimidating, discrediting and ridiculing the defenders and their work.

One organisation in Central America found that digital attacks accounted for about a quarter of all attacks against WHRDs the organisation documented between 2020 and June 2025, a number that has risen continuously each year. In **El Salvador**, digital attacks accounted for more than half of all attacks against WHRDs. Nearly 90% of these attacks were carried out on social media. [48]

Another organisation reported that **Syrian** WHRD Hiba Ezzideen Al-Hajji was subjected to a defamation campaign on Facebook targeting her, her family and the organisation she is a part of because of a post advocating against forced marriages. She received numerous death threats through a flood of posts on her own account, as well as on her organisation's page, with calls for physical violence against her. [49][50] A few years prior, she had already been smeared and threatened via Telegram. [51]

In **Bangladesh**, one organisation highlighted that the political transition of 2024 marked a turning point, where online harassment evolved from isolated incidents into coordinated campaigns, with a sharp rise in doxing, visual defamation, meme-based hate speech, mass reporting of HRDs' accounts on Facebook and Instagram, and the circulation of targeted misinformation. Such attacks often carried gendered and religiously coded abuse, especially targeting feminists, secular human rights defenders, and LGBTQI+ human rights defenders. [52]

Several human rights organisations reported that the growing use of artificial intelligence (AI) to generate deepfakes and manipulated images had increased the vulnerability of HRDs, in particular WHRDs, online. [53][54][55]

---

[48] Submission (H) received by the Special Rapporteur
[49] Submission (A) received by the Special Rapporteur
[50] See AL SYR 3/2025
[51] See AL OTH 129/2023, to which Special Procedures mandate holders received no response from Telegram
[52] Submission (A) received by the Special Rapporteur
[53] Submission (H) received by the Special Rapporteur
[54] Submission (J) received by the Special Rapporteur
[55] Submission (A) received by the Special Rapporteur

This has been the case in **Venezuela**, where WHRD Martha Lía Grajales was the target of a defamation campaign based on an AI-generated video promoting the false narrative that her human rights work was a cover for plotting against the government. The video was widely shared on social media and WhatsApp, with the involvement of digital militias dedicated to carry out and amplify such attacks. Once disseminated, such content is hard to contest or remove. [56]

In all the aforementioned examples, platforms have reportedly been slow or ineffective in responding to these harms.

Multiple submissions highlighted the role of social media platforms in fuelling and amplifying harmful content based on their algorithms, which prioritise sensational, polarising and violent content and thus increases the reach of harassers and abusers. [57] [58][59] One organisation warned that the profit-driven algorithms currently in place on mainstream social media platforms directly contribute to digital violence against WHRDs in **Mesoamerica**, which are part of a continuous cycle of violence that often translates into the physical realm. [60] According to the aforementioned survey by Global Witness, 84% of environmental defenders across Africa, Latin America and Asia who reported being targeted due to activism believe the online harm either "directly" or "somewhat" contributed to the offline harm. [61]

4.2 LACK OF EFFECTIVE COMPLAINTS MECHANISMS

Across all submissions, organisations stressed the lack of responsive, transparent, or human-driven complaint systems when human rights defenders try to report abuse. The Global Witness survey found that only 12% of the human rights defenders who reported their abuse and harassment to social media platforms were satisfied with the response that they received from them. [62] According to one of the submitting organisations, platforms such as TikTok and Telegram do not respond at all to requests for help from specialised organisations in the **Central America** region. Moreover, many processes require fluency in English and are managed by bots or automated replies. [63] In **Bangladesh**, one organisation found that over 90% of abuse reports made to Facebook in relation to online harassment of human rights defenders were either ignored or resulted in retaliatory actions against the complainant instead of the perpetrator, including

[56] Submission (A) received by the Special Rapporteur
[57] Submission (I) received by the Special Rapporteur
[58] Submission (H) received by the Special Rapporteur
[59] Submission (C) received by the Special Rapporteur
[60] Submissions (B), (F), (H) and (J) received by the Special Rapporteur
[61] Global Witness, "Toxic platforms, broken planet", 16 July 2025
[62] Ibid
[63] Submission (H) received by the Special Rapporteur

suspension of accounts. [64] One **Liberian** WHRD highlighted how the lack of response of social media platforms to incidents of online attacks against WHRDs promoting women's political participation in the country had in some cases led to physical attacks against them. [65] One organisation, which dedicates time and resources to running a digital safety helpdesk for HRDs, reported that most of the tickets it tracks are eventually resolved. However, the platforms' response time of up to 14 days, and the occasional lack of response or negative resolution, contribute to further endangering HRDs. Furthermore, it is important to note that most HRDs do not have access to such sustained support and resources, resulting in comparatively less positive resolutions in their dealings with social media platforms. Concerns were also raised regarding cases of account impersonation or harassment in chats, which are common issues for WHRDs in the **Middle East and North Africa** and can lead to severe harm. Yet, the organisation reports that platforms routinely fail to solve such issues, simply asking for users to fill out a form with no tangible result. [66]

5. UNEVEN TREATMENT OF USERS & CONTENT

Several organisations identified disparities in how social media platforms treat human rights defenders and the content they post, notably based on their language and region. [67]

5.1 LANGUAGE DISPARITIES

One major issue is the lack of adequate language capacity within content moderation teams and tools. This is especially evident in the moderation of languages other than English, French or Spanish, where harmful or hateful content often goes unrecognised and unaddressed.

One organisation's research on online harassment in **Bangladesh** has demonstrated how platforms like Facebook fail to detect and act on abusive content in Bangla. Despite repeated reports, abusive accounts often remain active, and hate terms such as

[64] Submission (A) received by the Special Rapporteur
[65] Submission (K) received by the Special Rapporteur
[66] Submission (C) received by the Special Rapporteur
[67] Submissions (A), (B), (C), (F) and (I) received by the Special Rapporteur

"Shahbagi" [68], used in a derogatory way against HRDs and activists, are not flagged or removed. [69] At the same time, over-moderation in other languages creates a different but equally harmful dynamic. Several organisations noted that Arabic-language posts, especially those related to **Palestine**, **Syria** and broader regional rights issues, are more aggressively moderated than posts in other languages. [70]

Overall, content in languages other than English, French or Spanish tends to be disproportionately removed or flagged, as social media platforms lack local language moderation capacity. [71] One human rights organisations in **Cambodia** raised similar concerns, noting that Khmer-language takedowns lacked explanations, and that no support infrastructure existed for HRDs who only speak Khmer. [72] Another organisation reported how Urdu was a "blind spot" for platforms when it comes to hate speech and threats against human rights defenders in **Pakistan**. [73] In **Iran**, Instagram's process of moderating Persian-language content has also exposed deficiencies in the application of the platform's content moderation policies in non-Western languages. [74]

5.2 REGIONAL DISPARITIES

Several organisations also reported significant regional disparities in how human rights defenders and their content are treated on social media platforms. [75] According to research by one submitting organisation, although the rate at which HRDs report harms against them to platforms is fairly even between different global regions, African defenders were less likely to receive a response to their complaint than their European counterparts (50% vs. 72% respectively). [76] Another organisation identified similar gaps in how platforms address harassment against HRDs across different regions. [77]

---

[68] Term that originated during the 2013 Shahbag protests, used as derogatory slang to refer to left-leaning, liberal or secular individuals, often implying islamophobia or allegiance to media trials and mob justice.
[69] Submission (A) received by the Special Rapporteur
[70] Submissions (A), (B) and (C) received by the Special Rapporteur
[71] Submissions (A), (B), (C) and (E) received by the Special Rapporteur
[72] Submission (E) received by the Special Rapporteur
[73] Submission (B) received by the Special Rapporteur
[74] Article 19, "Iran: Meta must overhaul Persian-language content moderation on Instagram", June 2022 https://www.article19.org/resources/iran-meta-persian-language-content-moderation-instagram/
[75] Submissions (A), (B), (F) and (I) received by the Special Rapporteur
[76] Submission (I) received by the Special Rapporteur
[77] Submission (A) received by the Special Rapporteur

6. DIFFICULTIES ENGAGING WITH SOCIAL MEDIA PLATFORMS

6.1 DISAPPOINTING OUTCOMES FOR HUMAN RIGHTS DEFENDERS

In the face of the aforementioned restrictions, rights violations and biases, human rights defenders themselves have very limited access to remedy.

One organisation's research shows that in the absence of dedicated civil society support to accompany HRDs in their complaints to social media platforms, the effectiveness of their responses remains severely limited. Platforms like Meta have community standards that often fail to recognise many types of online abuse as going against their policies, especially when the harmful content is framed as part of public debate or freedom of expression. As a result, many reports of harassment are ignored or rejected, leaving human rights defenders with no meaningful recourse. Even when platforms do take action, the response is often slow and does not match the seriousness of the abuse. [78]

Another organisation identified that the main obstacles to WHRDs' access to remedy in cases of digital attacks in **Mesoamerica** were the lack of adequate mechanisms, difficulties in accessing them, especially for those who do not speak English, long waiting periods to obtain a response, the lack of follow-up, and the fact that responses were often automated and managed by bots, making it nearly impossible to obtain a human response. [79] As a result, some defenders choose not to report the abuse they face. 65% of WHRDs who reported digital attacks to this organisation did not report them to the platforms themselves, mostly because of the inefficiency of their mechanisms. [80] Another organisation's research on online harassment in **Bangladesh** showed a similar pattern, with HRDs interviewed recounting that Facebook's reporting mechanisms were often convoluted, opaque, and discouraging, leading many to become disillusioned and stop pursuing formal complaints. [81]

In certain cases involving direct threats or unauthorised use of defenders' images, platforms have occasionally responded by removing posts or taking down offending profiles. However, such interventions are rare and inconsistent. Similarly, positive resolutions are rare when it comes to appeals related to blocks and suspensions.

---

[78] Submission (A) received by the Special Rapporteur
[79] Submission (H) received by the Special Rapporteur
[80] Submission (H) received by the Special Rapporteur
[81] Submission (A) received by the Special Rapporteur

The opacity of the complaint and appeal processes, the lack of transparency as well as the absence of human interlocutors contribute to human rights defenders rarely receiving the answers and remedy they seek. This leads many defenders to seek support from third party civil society organisations to try and obtain results.

6.2 CIVIL SOCIETY AS INTERMEDIARIES: A FLAWED & UNSUSTAINABLE MODEL

Some well-established human rights organisations are maintaining direct channels of communication with social media platforms in order to flag cases and extend support to human rights defenders. While these channels have allowed for comparatively more satisfactory resolutions and speedier processes, civil society organisations often find them unreliable or inconsistent. [82] One organisation reports that engagement with social media platforms on behalf of defenders is often unproductive, with no meaningful systemic change. [83] Similarly, another stated that results were never guaranteed. When the organisation tried to secure the reinstatement of a **Kazakh** HRD's account by establishing a direct line of communication between the HRD and Meta, it received no follow-up despite repeated attempts. [84]

Even initiatives such as Meta's Trusted Partner Program, created to foster collaboration between the platform and civil society organisations, have yielded inconsistent results. One organisation reported that, in instances where HRDs' content was suppressed or flagged, it would attempt follow-up through these official support channels. While responses from Meta's Trusted Partner Channel were more frequent than from X or Telegram, they often consisted of automated replies lacking substantive justification. Even when explanations were provided, they rarely acknowledged or understood the broader context of the situation, and escalation did not guarantee resolution. [85]

Overall, the engagement between civil society organisations and platforms seems to be marked by inconsistency, lack of transparency, and growing disengagement on the part of the companies. Despite this, NGOs are striving to fill the gaps left by platforms and increasingly playing intermediary roles, documenting harms, engaging with platforms, and advocating for users. [86] However, this is not a sustainable model, as it places the burden of protection and support on civil society, while the platforms' lack of accessible and effective mechanisms for HRDs at risk persists. [87]

---

[82] Submissions (A), (B), (C), (F) and (I) received by the Special Rapporteur
[83] Submission (I) received by the Special Rapporteur
[84] Submission (B) received by the Special Rapporteur
[85] Submission (A) received by the Special Rapporteur
[86] Submissions (A), (B) and (C) received by the Special Rapporteur
[87] Submission (A) received by the Special Rapporteur

6.3 PLATFORMS' ERODING COMMITMENTS TO HUMAN RIGHTS

In recent years, in what can only be seen as an indication of their eroding commitment to human rights, mainstream social media companies have started to gradually pull back from collaboration with human rights defenders and NGOs.

Despite a long history of working directly with social media platforms, both informally and as an official member of various trusted partner programmes, one organisation stated that engagement had grown less and less productive, with response times increasing significantly in 2025. Harmful content was sometimes left up for 10 to 15 days after reporting, even in cases of serious threats against WHRDs in the **Middle East and North Africa**. It was also reported that social media platforms were far less responsive than they used to be to public pressure. [88] Other civil society organisations concurred that platform engagement on human rights has substantially declined over the past three years, with diminishing interest in civil society input. [89]

These changes happened in parallel to social media companies resorting to mass layoffs of key personnel and teams devoted to trust, user safety, human rights and countering misinformation, as well as the dismantling of advisory bodies and programmes centred on these issues. Affected platforms include X [90], Meta [91], YouTube [92] and TikTok [93]. Such structural changes have drastically weakened any avenues of dialogue human rights defenders and civil society organisations previously had, and have impacted the platforms' ability to address individual cases and structural issues [94].

Simultaneously, many mainstream social media platforms rolled back crucial policies that were contributing to the protection of users' rights. Meta, X and YouTube have all

---

[88] Submission (C) received by the Special Rapporteur
[89] Submissions (I) and (J) received by the Special Rapporteur
[90] Amnesty International, "Global: Twitter's decision to disband safety council threatens wellbeing of users", December 2022 https://www.amnesty.org/en/latest/news/2022/12/global-twitters-decision-to-disband-safety-council-threatens-wellbeing-of-users/;
Electronic Frontier Foundation, "Monetization, Not Human Rights or Vulnerable Communities, Matter Most at Twitter Under Musk", November 2022 https://www.eff.org/deeplinks/2022/11/twitters-monetizable-users-not-human-rights-matter-most-under-musks-rein-leaving;
Submission (F) received by the Special Rapporteur
[91] Free Press, "Big Tech Backslide: How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections", December 2023 https://www.freepress.net/big-tech-backslide-how-social-media-rollbacks-endanger-democracy-ahead-2024-elections; Submission (F) received by the Special Rapporteur
[92] Free Press, ibid
[93] Business and Human Rights Centre, "Germany: TikTok content moderators strike over mass layoffs & plans to replace them with AI", August 2025 https://www.business-humanrights.org/en/latest-news/deutschland-content-moderatorinnen-protestieren-gegen-stellenabbau-geplante-ersetzung-durch-ki-algorithmus/
[94] Submission (C) received by the Special Rapporteur

all weakened privacy protections for users to allow for the training of their AI tools. Meta has scaled back on third-party factchecking in favour of "community-based" or automated moderation, which raises concerns regarding the spread of misinformation. X and YouTube have also rolled back election-misinformation policies. [95]

While Meta has since 2022 published annual human rights reports detailing how they identify and mitigate human rights risks on their platforms, its latest issue [96] does not deal with the aforementioned challenges in any substantive way. Other platforms, such as X, TikTok or YouTube, do not issue dedicated, standalone reports on human rights.

7. CONCLUSION: IMPACT ON DEFENDERS' ONLINE PRESENCE & ORGANISATIONAL STRATEGY

The cumulative impact of the forms of digital repression mentioned above is a shrinking digital civic space, with human rights defenders being gradually silenced online.

When confronted with visibility restrictions, harassment, or surveillance via social media, and with no effective or adequate platform support to remedy such violations, a high number of human rights defenders report feelings of fear and anxiety for themselves and their communities, as made clear in submissions received for this paper [97] and in the Special Rapporteur's interactions with HRDs. As a result, they can choose to reduce their visibility, avoid sensitive topics, or leave platforms entirely.

Women human rights defenders in **El Salvador** have pulled back from public discourse due to sustained digital harassment [98]. In **Pakistan**, a 2023 study by Media Matters for Democracy found that 8 out of 10 women journalists were self-censoring online due to the fear of digital and physical violence. [99] In **Cambodia**, one organisation found that one-third of journalists now avoid posting content critical of the authorities on social media. [100]

---

[95] Free Press, ibid
[96] Meta, "2024 Human Rights Report", December 2025, https://humanrights.fb.com/wp-content/uploads/2025/12/2024-Meta-Human-Rights-Report.pdf
[97] Submissions (A), (H), (I) and (J) received by the Special Rapporteur
[98] Submission (H) received by the Special Rapporteur
[99] Submission (J) received by the Special Rapporteur
[100] Submission (E) received by the Special Rapporteur

Elsewhere in **Asia**, self-censorship on politically sensitive themes remains prevalent [101]. Nearly half of the land and environmental defenders surveyed by Global Witness reported that the abuse they suffered online had led to a loss of productivity and diminished capacity to campaign on these issues. [102]

The overall impact is a diminished digital presence and reduced public engagement by human rights defenders, in particular those belonging to marginalised communities and those working on sensitive issues. This lays the groundwork for an online ecosystem that reinforces existing power dynamics, leaving communities that are already overexposed to harms at risk of being completely erased from digital spaces.

Some human rights defenders and groups are exploring alternative, more autonomous platforms, where safety and privacy are prioritised over reach. Others have adjusted their strategy when it comes to social media presence, deciding to limit their digital footprints, use pseudonyms or coded language and avoid real-time posting, inter alia [103]. However, mainstream social media platforms remain essential for most human rights defenders. Newer platforms still have very limited reach and far less active users, and the current digital ecosystem remains dominated by a handful of large, profit-driven social media platforms. This leads human rights defenders to be forced to balance visibility with safety. Several organisations told us about the difficulty for defenders to fully disengage without risking sacrificing the impact of their work, with the lack of safe alternatives leading to continued exposure to risks. [104]

[101] Submission (F) received by the Special Rapporteur
[102] Global Witness, ibid.
[103] Submissions (A), (C), (F) and (I) received by the Special Rapporteur
[104] Submissions (A), (B) and (C) received by the Special Rapporteur

RECOMMENDATIONS

TO COMPANIES OPERATING SOCIAL MEDIA PLATFORMS:

- Ensure **content moderation and curation policies** are strictly aligned with international human rights law and do not impinge upon the right to defend human rights;
- Refrain from developing and using **AI** in content moderation and removal practices without third-party, human fact-checking and verification;
- Improve **transparency** on content moderation decisions and appeals;
- Enhance **privacy and security tools**, including stronger encryption, anonymous reporting, and safeguards against doxxing and surveillance;
- Review the **effectiveness of reporting mechanisms** for online abuse, and invest resources in trusted, accessible, **human-led support** mechanisms to provide assistance to human rights defenders in need of it, staffed by teams trained in relevant languages, cultures and political contexts;
- Ensure faster response times, especially to **trusted partners and urgent cases**;
- Invest in **gender-sensitive** moderation tools and protocols and incorporate specific gender-based harassment and abuse sections in community guidelines;
- Implement external **audits of algorithms** to curb the amplification of disinformation and hateful or abusive content, and commit to inclusive design that does not marginalise human rights content;
- Resist compliance with **state censorship requests** that conflict with international human rights norms;
- Publicly log and **publish state takedown requests** and platform responses;
- Reverse recent **rollbacks in hate speech, misinformation and privacy policies** and Trust & Safety and Human Rights staffing;
- Establish and expand **communication channels with human rights defenders and civil society**, especially in regions where no structured mechanisms exist, and ensure they have a say in the design of policies and support systems that work for them.

TO STATES:

- Ensure that social media companies respect freedom of expression and **uphold human rights standards**;
- Consider **regulatory solutions** to address the power of dominant platforms;
- Recognise digital violence as part of the broader continuum of attacks against HRDs;
- Investigate and prosecute online violence against HRDs;
- Adopt clear **legal and policy frameworks** that hold social media platforms accountable for their role in enabling online harassment, especially when such actions threaten the safety of HRDs;
- Refrain from coercing social media platforms to implement censorship or surveillance.

HRC-SR-DEFENDERS@UN.ORG
WWW.SRDEFENDERS.ORG

UN SPECIAL RAPPORTEUR ON HUMAN RIGHTS DEFENDERS

IN SOLIDARITY & HOPE