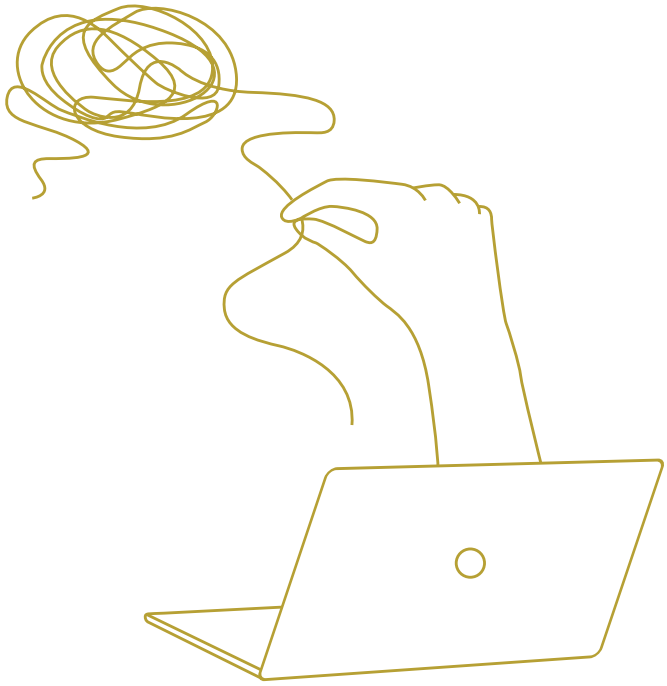
DANCHURCHAID

actalliance

ONLINE HARASSMENT AND CENSORSHIP OF WOMEN HUMAN RIGHTS DEFENDERS



February 2023

Important notice: None of the people depicted in the report have anything to do with the interviewees.

ACRONYMS

CSO	Civil Society Organisations
DCA	DanChurchAid
GBV	Gender-based Violence
HRD	Human Rights Defenders
LGBT+	Lesbian, Gay, Bisexual, Transgender and Other Identities
UN	United Nations
WHRD	Women Human Rights Defenders

CONTENT

[Photo: Mikkel Østergaard]

01. ONLINE HARASSMENT AND CENSORSHIP OF WOMEN HUMAN RIGHTS DEFENDERS

Many human rights defenders (HRDs) experience external deliberate censorship, takedowns or online harassment because of their work in digital spaces. Women encounter particularly high levels of harassment. According to a UN Human Rights Council panel, women and girls are **27 times** more likely to be harassed online than men.[1] In addition to the impacts on physical and mental health and dignity, the threat of online abuse is also leading many women to practice self-censorship or to "log off" social media, perpetuating and entrenching inequalities within the space.

Because social media has become an intrinsic part of modern life and a vital tool to organise, spread information, advocate, and educate about social issues, the harassment of HRDs in general and Women Human Rights Defenders (WHRDs) poses a serious threat to human rights and democracy.

In this report, DanChurchAid (DCA) will outline the scope of online harms faced by WHRDs and uncover areas for improvement in both regulation and tech platform operation and content moderation. It will provide recommendations of actions to be taken by tech sector platforms, states and civil society to combat online harassment and ensure a strong, inclusive democratic debate online.

The report is based on publicly available data and surveys, as well as interviews with DCA partners and staff from Palestine, Israel, Ukraine, Kenya, Nepal, Myanmar and Cambodia. To protect their safety, all interviewees will remain anonymous.

Main authors of the report are Maia Kahlke Lorentzen, Cybernauterne, and Karina Pultz, Senior Human Rights Advisor, DCA. Other contributors to the report are Christina Dahl Jensen, Team Leader and Senior Innovation Advisor, DCA, Joy Anne Icayan, Civic Space Advisor, DCA and Adaline Hui, intern, DCA. The report is based on open-source global reports and qualitative interviews with partners of DCA. ■

[Photo: Srey Vann]

02. THE SCOPE OF THE PROBLEM

Multiple surveys and studies have shown that women and girls feel the negative impacts of social media and digital communication platforms most acutely. The largest international survey to date, done by Plan International in 2020, found that 58% of young women and girls have experienced online harassment.[2] The harassment ranges from receiving threatening or hateful messages, to unsolicited sexual messages and images, having their images shared in degrading or demeaning context, being cyberstalked, defamed and blackmailed.

The data, covering over 14.000 participants from more than 22 countries, highlights the same tendencies across low- and high-income countries, and across all social media platforms included in the study.[3]

While some may think that online harassment is less harmful than physical violence, the report sheds light on some alarming consequences: One in five of the respondents had left or reduced their use of social media platforms, and 12% reported that they changed the way they express themselves on account of being harassed.

Though depressing, the findings are not surprising. Similar research done by Amnesty International has shown that women experience high levels of online harassment, and highlights the intersectionality of the issue. In a project that mapped abuse on Twitter across 150 countries, Amnesty International found that women of colour were 34% more likely to receive abusive and problematic Tweets than white women, and that black women alone were 84% more likely to be targeted by abuse.[4] According to a survey conducted by the LGBT+ advocacy group, The Trevor Project, 40% of LGBTQ youth reported being the target of online bullying. Additionally, a study by the Cyberbullying Research Center found that LGBT+ individuals are more likely to experience cyberbullying compared to their non-LGBTQ peers.

If regular women on social media experience high levels of harassment, it should come as no surprise that women who are part of the public sphere experience event greater levels of online harassment and harm due to increased exposure and excessive scrutiny.

According to a survey done by Inter-Parliamentary Union from 2016, 81.8% of women parliamentarians across five regions reported experiencing psychological violence on social media. Over a third of the respondents claimed the harassment had ***"undermined their ability to fulfil their mandates and freely express their opinions."*** [5]

According to a study published by International Women's Media Foundation in 2021, 63% of women journalists indicated they had been threatened or harassed online.[6] Over a third of the respondents reported that the harassment led them to avoid covering certain types of stories that were likely to stoke negative reactions.

Both studies point to the intersectional nature of digital violence and online harassment. Women parliamentarians and journalists belonging to ethnic or religious minority groups, lower castes, classes, or LGBT+ minorities, face increased levels of harassment targeted at their identities.

Many WHRDs who work in politics and media trying to effect change on politically charged agendas are under hard opposition from state actors and civil society groups. The office of the UN High Commissioner highlights how this threatens WHRDs specifically,

"WHRDs face all of the same risks and violations as human rights defenders generally. However, the consequences of these violations are often gender-specific for WHRDs due to the prevailing social and cultural norms in a given context. WHRDs can also face additional gender-specific threats and violence, in both public and private spheres, such as gendered verbal abuse (online and offline), sexual harassment, rape and sexual violence, which can also lead to further violations, such as stigmatisation."[7]

WHRDs often challenge societal norms and issues related to stereotypes of gender, sexuality, family, and conservative notions of what "proper" female behaviour looks like. Furthermore, issues like gender and LGBTQIA+ equality, sexual and reproductive health, sex workers' rights, and gender-based violence (GBV) are often portrayed as being somehow foreign or cultural imports, making WHRDs akin to foreign agents or traitors to their culture or nation state. The UN High Commissioner writes in their update that ***"the assertion that these defenders are somehow advocating or attempting to import 'foreign' or 'Western' values which contradict national or regional culture. State agents or representatives are often alleged to be responsible for such stigmatization."*** [8] ■


[Photo: Albert Hillert]

03. THE NATURE OF THE PROBLEM

Digital tools are vital in the strive for human rights. Messaging and communication apps are crucial for organising efforts. Social media has revolutionised the ability of marginalised and disenfranchised groups to reach larger audiences, mobilise and have their voices heard without gatekeeping by traditional media. However, these same digital tools are also weaponised by civil society and state actors to stigmatise, shame and defame WHRDs through gender stereotypes. Accusations of attempting to destabilise traditional gender roles, family values and national cohesion are also hurled at WHRDs.

The following chapter will shed light on the nature of the digital harassment and violence directed at WHRDs,

informed by a series of interviews conducted among DCA partners and staff. Based on the information collected from the interviews, the nature of the online harassment can be understood in four overarching categories:

- Sexist hate speech and threats
- Defamation and disinformation
- Privacy violations
- Censorship

Sexist hate speech and threats

As highlighted in the first chapter of this report, sexist and derogatory remarks aimed at women are abundant on social media, and no less so for the WHRDs interviewed for this report, where only one WHRD of all the interviewees did not experience gendered attacks on her personally. The comments often target their looks and are either derogatory or sexualised remarks on their body types, age, dress, hair, and makeup. Many comments also revolve around their perceived transgressions against traditional gender roles; if they are married, they are asked why they are not at home with their kids; if they are not married, they are blamed for not fulfilling their "purpose" as women. Several of the women interviewed have received rape and death threats.

One WHRD expresses it like this: "They feel threatened and irritated by young women. They try to denounce me by using my gender. They say, 'She does it just for the attention.' I have received thousands of rape and death threats."

One WHRD relays an experience from a colleague: "She started to receive photos of her home, and also photos of the school her children went to. What we see sometimes is that those kinds of attacks they try to use 'you are a woman, you should be in the kitchen, and you should not write investigative reports. You do not know what you are talking about. Think about your children.' They try to appeal to these maternal instincts."

Hate speech is a broad category that encompasses speech that directs hate or threats of violence against a group or individual, based on group characteristics like gender, gender identity, sexual orientation, race, ethnicity, religion or disabilities. While not all types of sexist remarks aimed at women fall into the hate speech category, many do.

The WHRDs interviewed for this report all experienced defaming or hateful messages, comments and discourse. All but one experienced personal comments aimed at their gender. Many did not highlight these types of experiences unless specifically asked, because sexist remarks and hateful gendered messages are largely perceived to "come with the territory". Meaning when you speak up for minority groups on rights violations, you expect to be harassed.

One WHRD described how the staff from their organisation working for LGBTQIA+ rights had their images posted on a public far-right Facebook page, in an effort to "expose" their work. The comments used the looks of the WHRDs to discredit them by calling them ugly and "exposing them", which in turn led to an onslaught of defamatory messages, derogatory comments on their looks, and violent threats. These threats come in different forms, as another interviewee explained, "Sometimes it is just somebody cursing you, and sometimes it is somebody saying, 'You deserve to die' or 'We are coming to kill you.'" Another WHRD expressed that she experienced attacks when raising issues in relation to racism, women's rights and the rights of ethnic minorities in her country.

The WHRDs interviewed often came across threats of violence that were gendered in nature. Some were direct threats of sexual violence or rape, while others directed their intentions toward family members, for instance, threats aimed at children, or messages to male family members encouraging them to assert dominance or control over the WHRDs. One of the interviewees recounts how her grandparents received harassing calls about her during a defamation campaign aimed at the organisation she worked for, in which she was also a target.

The interviewees reported the threats being made on several platforms, ranging from comments on social media, to emails and phone calls. They also noted that they could not always tell which threats were to be taken seriously. As one WHRD notes, "It is hard to know what a real threat is and what is not." Also, the fact that harassment campaigns travel across platforms make it overwhelming and difficult to protect yourself against them.

The type of hate speech and threats the interviewees experienced often intersect with other forms of oppression. WHRDs belonging to minority groups are likely to experience hate speech directed at both their gender and their minority status.

Multiple studies and projects have highlighted that hate speech and harassment plays into existing inequalities and biases reflected in the cultural context. Amnesty International's report from India showed that while one in seven tweets mentioning women politicians in India were "problematic", Muslim women received 94.1% more ethnic or religious slurs than women from other religions. Women politicians belonging to marginalised castes received 59% more caste-based abuse compared to other women.[9] One interviewee describes how the threats of rape she received were interwoven with racist ethnic stereotypes directed at Muslim men. "There is a saying in Hebrew: 'We will send you to Gaza.' That is like a way to threaten to rape."

Defamation and reputational attacks

The Office of the UN High Commissioner for Human Rights highlights defamation as a serious obstacle to the work of WHRDs, stating that attacks against WHRDs often focus on their reputation and/ or their sexuality as non-conforming with dominant stereotypes of 'appropriate' behaviour by women and men. [10]

This aligns with the experiences of the WHRDs interviewed, who all experienced smear campaigns, defamation and reputational attacks directed at both their organisations and them.

Apart from containing the types of sexism and hate speech outlined in the previous section, these defamation attacks can also include fabrications and direct falsehoods. The so-called "gendered disinformation" is a type of misinformation specifically fabricated to play into gender stereotypes and double standards. Examples range from manipulated media reports that portrays women in sexualised situations, accusations of indecency or infidelity, or rumours about trading sexual favours for influence and power. Gendered disinformation is often weaponised against women in politics, as it undermines credibility and hijacks agendas. Ample examples from the United States, Rwanda and recently Finland have shown us gendered disinformation directed at female politicians.[11]

Social media platforms enable the creation and dissemination of mis- and disinformation. This includes fake news and manipulated content but can also come in the form of "real" news stories or social media commentary that deliberately misrepresents the views of the organisation or the WHRDs. Mis- and disinformation are often the starting points for wider harassment campaigns, hate speech, threats and other forms of harassment that follow in its wake.

One interviewee told us of the harassment she and her colleagues faced due to a defamation video going viral on social media. The video falsely accused the organisation of supporting terrorists and exposed staff names and photos, causing them to receive targeted harassment and threats from people who believed the information stated in the video to be true. The talking points from the video were also picked up by mainstream media outlets, leading to further exposure and thus perpetuating and prolonging the harassment and disinformation cycle.

WHRDs told us that defamation and reputational attacks are rarely constrained to social media but are amplified and given credibility by the participation of mainstream media and public figures like politicians or political pundits. If a mainstream media outlet covers a piece of misinformation or a rumour that is going around on social media, they give it a veneer of respectability or veracity, that in turn risks increasing the harassment of the WHRDs, who are left to combat the onslaught of harassment while trying to refute the false accusations.

"I receive dozens of threats and curses every day. Whenever the media would amplify the fake news, the threats would increase. I do not have resources to handle all the harassment and disinformation."

Debunking falsehoods or correcting misrepresentations takes a lot more time and effort than creating and spreading them. The result is that organisations and WHRDs spend a disproportionate amount of their precious and limited resources on this type of self-defence, instead of on their core agenda and goals.

Many WHRDs would try at first to ignore false information in order to not lend it credibility. However, when misinformation is picked up by mainstream media outlets or propagated by politicians, there is a risk of damaging relationships with partners or funders. Hence, organisations have to eventually attempt to address such issues. But according to the interviewees, the cases are often dealt with, with limited success.

The reputational damage and the organisational fatigue resulting from disinformation and reputational attacks are often more effective at silencing WHRDs than direct threats and insults. This plays into the dynamic of shrinking civic spaces as defamation and reputational attacks creates serious barriers for the work of civil society organisations (CSOs), as well as grassroots organisers and activists.

Privacy violations

With so much of our private lives taking place on the internet, social media and digital accounts, privacy violations have become part of the harassment playbook.

The interviewed WHRDs all reported instances of attempted and successful hacks of social media accounts, organisational databases, email accounts, and voicemails. Much of the information obtained through these illegal means can be used for smear campaigns or defamatory purposes. If, for instance, a WHRD's private photos are hacked and subsequently leaked, the reputational and psychological damage is substantial.

One interviewee experienced multiple attempts to hack and access their organisational database which contains sensitive information from anonymous sources. Breaches of databases and digital accounts can also enable attackers to carry out digital vandalism, like sharing false information on

social media, erasing critical data or shutting down digital accounts.

Furthermore, this sensitive information can be used to gain access to a target's contacts and their whereabouts. One WHRD describes how someone had hacked their voicemail in order to get access to not only their phone number, but also the phone numbers of their friends and grandparents, which resulted in them getting harassed as well.

Many of the interviewed WHRDs have, to some extent, experienced a type of privacy violation known as doxing, which is the practice of disclosing private information about an individual with malicious intent. One WHRD had her address and phone number shared on platforms like Telegram and Facebook by people who were hostile to her work. She recounted, **"My phone and my computer kind of became a part of the enemy's tool. (...) My email was published. It was not a secret."**

This highlights the dilemma of doxing, where a simple (and most likely) legal act, like the sharing of a publicly available email address, can lead to harassment and have dangerous implications for the mental and physical safety of the target. The information could have been obtained through legal means, for example, from a public Facebook profile or public records, while other information is gathered through illegal means. Doxing is also not always explicitly threatening, making it more difficult to report to platforms and authorities. As doxing is not illegal in most countries, it in turn encourages harassment and intimidation of a target without legal consequences or implications by the perpetrator.

Digital platforms have lowered the cost and access to the surveillance of WHRDs, be it through following their social media, hacking digital accounts or via more far-reaching and illegal means, like installing spyware on their devices. One WHRD tells of infiltrators and informants in their online groups passing on information, making it even harder to vet for these types of threats in online groups and on social media sites.

Censorship

Traditionally, censorship has been the purview of state actors, but with the mass proliferation of social media platforms, community guidelines and terms of service have become a form of regulation of speech on par with legal regulation.[12]

Social media platforms have community guidelines to ensure the safety of users and limit the spread of harmful and illegal content. Content that is deemed to be in conflict with these guidelines can be removed outright, but many platforms also use automated suppression of content that is deemed to be harmful or offensive by algorithms. While this can effectively limit the virality and reach of problematic content, community guidelines and mechanisms to upkeep them can also be weaponised against human rights defenders and CSOs, for example, through mass reporting.

A WHRD in Ukraine told us how any content related to the war is at risk of being algorithmically repressed, meaning followers will not see the posts on their social media feeds. Others had their accounts temporarily locked or deleted for posts that mentioned Russia and depicted the Ukrainian flag, but without an explanation of why this was deemed to be in conflict with community guidelines.

Reporting and news from ongoing violent conflicts often end up algorithmically repressed or removed because they conflict with community guidelines on sensitive or offensive content. These contents can also be flagged for depicting violence or extremism. However, this also poses a serious impediment for human rights organisations trying to communicate their work and reporting on human rights violations through social media platforms.

This has long been highlighted by organisations working in Palestine, as documented by Human Rights Watch.[13] Multiple organisations, Palestinian citizens and their allies experienced a digital version of a "media blackout", wherein their content was either repressed, hidden or removed without transparency about the causes or decision-making process.

As one WHRD puts it, **"We know that Facebook has a connection with the Israeli government. They can delete what they want and they can keep things the way they want. (...) It is bigger than us, because it is a government decision, with a lot of money."**

WHRDs also often experience having either content removed or their accounts locked or shut down due to mass-reporting. This tactic, where people flag inoffensive content as offensive, can be a way to get accounts and pages of political opponents removed and deplatformed from social media. Since there are rarely official routes to appeal decisions of removed accounts from social media after they have been flagged, this can be an effective way to silence human rights defenders and organisations. ■

[Photo: Mikkel Østergaard]

04. WHO ARE THE PERPETRATORS

When analysing threats of sustained and organised digital violence and harassment, the perpetrators are usually grouped into two larger categories:

1. **State actors:** This category ranges from elected officials, ministerial departments, security agencies, state-run cyber armies[14], as well as state-paid influence agencies and state-run media.[15]

2. **Social actors:** Any organised group from civil society, ranging from political cyber activists, corporate paid influencer agencies, astroturf groups, members of mainstream media, or online troll-communities.

It can be difficult to ascertain who is behind an organised online attack. Social media and digital tools make it easy to be anonymous, assume fake identities and perpetrate harassment or disinformation campaigns at scale, obfuscating that just a few active operators are running the show.

There have been cases of viral Twitter trends organised by just a few bot-like accounts being picked up by mainstream media outlets and reported as if the contents were the opinions of real citizens.[16] Similarly, there were cases of state-run influencer agencies pretending to be active political groups during elections.[17] These incidents were exposed due to the diligent work of investigative journalists and fact-checkers. But often, it is hard to prove who is behind the online harassment that WHRDs face. Some may be orchestrated initially, and then evolve into spreading organically by unaffiliated social media users believing the misinformation or choosing to pile on.

Many of the WHRDs that DCA partners with describe that states or governments are involved in the harassment they experience. One WHRD explained how a vicious online campaign involving both social actors and the state changed their view on the society they lived in. "I thought I was running a human rights organisation in a democratic context. I do not see it that way anymore. The regime turned against me. It is not just civil society organisations or political groups that are against us. It is the whole system: state, legal system, educational system, against us. It is fuelled by online discourse, and a mainstream media outrage cycle. It sent me on a journey to rethink my politics."

Another describes how a person with malicious intent tried extracting statements that could be used to portray the organisation in a negative light. "They recorded me giving legal advice to someone that told me, 'I want to go protesting and I need your advice on how I can protest.' And I was like 'I cannot tell you how to protest, (…) you can do whatever you want. But [the organisation] can't give you support if the way that you protest is in a violent way.' And [the caller] was pushing me to say things that I do not want to say, because they wanted to record me and to publish it saying that we are violent, and we support violence. They edited the recording, and they said, 'Look how [the organisation] supports people to do graffiti and so on.'" The misleading article in question was then shared by the president of the country, resulting in further miscrediting the organisation.

The two types of actors tend to overlap. A harassment campaign may be started by a governmental agency, garnering attention from social actors, reaching a stage where both regular citizens and members of mainstream media participate. This often leads to an overwhelming amount of attention. A WHRD describes it as a more challenging aspect than the specific subject matter of the harassment, "It was not about the specific harassment. It was to be overwhelmed from the amount of hatred that you get. And it gets to you from all sides and all day long. And you turn to be kind of numb towards it. It happened very synchronised because things happen in real life usually by politicians - the people who fuelled it in real life were politicians. It seemed very organised then. But today I do not think it was just a campaign like that. I think it just got out of control. It was very, very successful."

It is not only the type of actors that overlap. Also, the harassment campaigns travel across platforms whereas the tech companies moderate in pillars and do not cooperate sufficiently to detect and respond to the harassment campaigns. ■

[Photo: Jakob Dall]

05. THE CONSEQUENCES OF DIGITAL VIOLENCE AGAINST WHRDS

For WHRDs that are subject to online harassment, it can be quite a challenge to keep up their spirits. Many describe feeling a desire to withdraw from social media, and this notion is supported by a 2020 study on patterns of cyber violence.[18] Out of the 356 surveyed women above 18 years of age, 148 of them had experienced cyberviolence within the past year. 18% of those exposed had withdrawn from social media as a result. An even larger portion of the 148 described feelings of anger (70%) and worry (35%) after being subjected to harassment.

One WHRD explains that being subjected to harassment had made her reconsider her relationship with the platforms. "My first response would be to rethink my relationship with digital media and technology and to see to what extent the platform is safe and offers really genuine freedom of expression."

This sentiment is shared by other WHRDs, who have chosen to adjust their online and social media behaviour. Many have chosen not to be active on public platforms like Facebook and Twitter, out of fear of harassment. Some are still using social media, but as one Kenyan WHRD puts it, "Our Twitter is very ruthless (...) so I see an issue that I really believe in and want to stand up for, but I end up keeping quiet because I know they are going to attack me."

The fear of harassment leads to self-censorship and limits the ways in which the WHRDs can make an impact. It has dangerous implications, essentially worsening the already bad conditions for HRDs doing work online. Harassment of other WHRDs makes an impact on all who work in the field, "Every single time a woman like me, or any other feminist or woman human rights defenders are targeted by online abuse, their confidence of course diminishes. (...) Especially woman politicians who are also HRDs in a sense."

While the impacts on the work of WHRDs can be devastating in itself, it also impacts other aspects of their life. Anger and worry can begin to dominate their lives. One WHRD describes how she started to feel paranoia, especially about her electronic devices. "The phone was seen as part of the enemy's tools. The phone or computer feels like your enemy. Email, social media, phone calls. I only used Facebook at that time."

In conflict and war settings, the lack of due diligence of content moderation can lead to severe human rights violations, such as enforced disappearances and killings. One WHRD in conflict setting says, "They (the de facto duty bearers, ed.) monitor our Facebook accounts and share them on Telegram, calling for people to share information on the activists, saying 'we need to arrest them'. And within a week, the activist would be arrested or killed." Another WHRD explained how the digital threat changed from a high level of online harassment against minority groups to more hacks against organisations stealing sensitive information during war time: "Before the war, there were some online attacks, especially doxing (...) especially towards LGBTQIA plus and women rights organisations (...) after the war we have had more focus on protecting against hacking attacks."

While this report mainly deals with WHRDs, it is important to note that these consequences apply to all women who engage in digital spaces. As one WHRD puts it, "You know, it is a part of the deal. Once you are there, that is part of the deal. It is to put yourself on the line."

In the following section, measures that can help change the conditions online for the better for both WHRDs and regular users alike will be discussed. ■

[Photo: Thomas White]

06. SOLUTIONS

Reducing the online harms that WHRDs face is no simple task. The problem is multifaceted, based on systemic inequalities and oppressive tactics predating modern digital platforms. It is further exacerbated by tech and social media platforms that were developed with scant thought for the protection of human rights.

Therefore, any solutions must also be multifaceted and will only succeed when combined with the efforts of governments, the tech sector, and civil society.

The following section contains an outline for recommendations for advocacy on legal frameworks and justice for online harms, and recommendations for practicing content moderation within a human rights framework aimed at tech platforms.

Legal frameworks for reducing digital harms

Regulators and politicians looking to mitigate online harms and harassment through a legal framework should ensure that legislation adequately covers:

- Protection of free speech and freedom of association and assembly
- Protection of the right to privacy
- Protection of freedom of access to information

Furthermore, efforts should be made to ensure the prohibition of:

- Hate speech and incitement of violence
- Libel and defamation
- Threats of violence
- Blackmail
- Unauthorized access to another person's digital devices

National legislation will in most countries cover some or all of these offenses, whether they are made online or offline.

Other types of offenses that are specific to online contexts, as many of them are only made possible in a digitalized society, will often not be covered in current legislation. These offenses could be, but are not limited to:

- **Image-based sexual abuse:** Colloquially known as "revenge porn". Multiple states have prohibited the practice of non-consensual sharing of intimate images. Some states also prohibit the sharing of non-consensual manipulated pornographic images and videos, also known as deepfakes. As these two types of harassment are leveraged against women at all levels of society and represent a threat to the victim's physical safety in many countries, a robust legal framework to prohibit these is advised.
- **Doxing:** Most countries do not have legislation criminalising the hostile sharing of publicly available information, like private address or phone numbers, without consent.

It is advisable to criminalise the sharing of private individuals' addresses and contact information without their consent. To avoid this being misused to silence critics and hinder advocacy campaigns and investigative journalism, exceptions should be made for contact information and official addresses that are of public interest, for instance, when belonging to elected officials, corporations or media.

- **Stalking:** While stalking is being criminalized under gender protection acts in multiple countries, lawmakers should be mindful to include cyberstalking and stalking by unknown perpetrators as well.
- **Online impersonation:** Social media has made it easy to impersonate organisations and individuals. In most countries, this is not illegal unless it is used for monetary gains. However, impersonation of an individual on social media can be used for harassment and disinformation tactics and can have damaging psychological impacts on the targets. Therefore, legislating against impersonation, with an exemption for satire purposes, is advisable.

Furthermore, initiatives that can increase the safety of human rights defenders online and offline are:

- **Secret addresses for at-risk individuals:** At-risk individuals, like WHRDs, often experience doxing of their private residence through publicly available information, like phone registries or land registry records. In some countries, like Denmark, it is possible to be unlisted from digital registries, so your phone number and private address is only known to the state and those you choose to share it with. An option like this could protect many at-risk individuals from doxing and privacy intrusions.
- **Legal injunctions against one or multiple harassers:** An injunction orders a harasser to seize communication by any means towards a victim. Getting injunctions against perpetrators of online harassers is in most cases extremely difficult but should be made easier in case of organised harassment.

Seeing the vast scope of harassment of women and WHRDs through digital tools, it is tempting to try and solve the complicated issues through legislative measures.

However, the WHRDs interviewed for this report pointed out that their national legislation in most cases covers and criminalises most of the harassing behaviour they experience. As one WHRD stated, "The laws are ok. Our legal text could be updated to better include digital definitions, but what we really need is better enforcement."

Another states her appreciation for the attempts at legislation but has little faith in their actual impact. "Even when the law is quite clear - your data protection act, your gender-based violence act, (...) they have very weak provisions on punishment for perpetrators - especially for online violence and online abuse - and there is a very, very long and tedious process to receive any form of justice."

While many countries would benefit from updating legislative texts to clearly include online harassment, threats, stalking and other types of online harms, the fact that these laws are rarely enforced in a fair and adequate manner remains.

Many of the interviewees relay that they do not feel that police and other authorities take reports of online harassment and digital violence seriously, and that when they do, their understanding of the methods and resources

for investigation fall short and fail to bring justice or mitigate harms.

Better legislation of online harms will not have the intended effects unless paired with education of police forces and allocation of resources to investigate and prosecute perpetrators of online harassment and violence. As with the issues of gender-based violence like domestic abuse and rape, sensitivity training should be included to avoid victim blaming and further traumatisation of victims.

For those WHRDs who are experiencing persecution from their state or violence from police departments, however, there is little hope that help combatting online harms and harassment will come from those same institutions.

Balancing free speech while safeguarding victims' rights is a delicate act. The effort to eradicate online harms can have unintended effects on already marginalised voices.

Some states with questionable human rights records have weaponised legislation around digital harms, fake news, and online harassment to enact censorship and prosecute human rights defenders. When Egypt's cybersecurity law included paragraphs on fake news, they were subsequently used to prosecute and silence critical media and HRDs.[19]

As many human rights defenders operate in a shrinking civil space and are at risk from state persecution, legislation that is nominally about combating online harms risks becoming a tool to silence their voices or disrupt vital human rights work.

Regulating the tech platforms

There is growing political momentum to regulate big tech platforms, and to create legislation aimed at social media companies. Britain's Online Safety Bill, which was debated in parliament in 2022, the European Union's Digital Services Act[20], the draft EU AI Act and Germany's NetzDG legislation, which was approved by the Bundestag in 2017, are just a few examples.

These legislative frameworks make social media companies liable for the content that their users post, impose a mandate on them to remove content that the government deems illegal within a specific timeframe, or even as with the UK Online Safety Bill, restricts encrypted services and penalises speech that is deemed to potentially cause "psychological harm" with jail time.

Older legislative frameworks, like the US Communications Decency Act, Section 230, stands to be repealed or reinterpreted.[21] The legislation created the foundation for US Big Tech's explosive growth as it protected social platforms from lawsuits over harmful user-generated content. This protection of US social media companies might be lifted or changed, forcing companies to transform their approach to moderating content. One consequence could be that the legal risk of moderating content will shift to individuals. By shifting the legal responsibility to individuals, democratic-oriented governments run the risk of stifling individual moderation, posing a threat to communication communities which traditionally are used by WHRDs in authoritarian states to voice their cause.

While regulating tech platforms is key to creating better safety for users and protecting both privacy and democratic integrity, both the British and German legislation have been criticised by human rights and internet freedom organisations for imposing state censorship on digital platforms under the guise of combating harassment and protecting users from online harms.[22] As Joe Mullin from the Electronic Frontier Foundation points out in their criticism of the UK's Online Safety Bill, ***"When governments around the world pressure websites to quickly remove content they deem 'terrorist', it results in censorship. The first victims of this type of censorship are usually human rights groups seeking to document abuses and war."*** [23] Also, the US's consideration to ban TikTok could inadvertently encourage other governments to tighten their grip over and block dominant US-based platforms, shrinking the digital civic space.[24]

The EU's Digital Services Act has more promising tenets, putting emphasis on getting service providers to do more to reduce online harms, although it also allows government agencies and providers to remove content they deem illegal or dangerous, as well as compromising the right to online anonymity in some areas.[25]

The dilemma of regulating online harms, while also protecting anonymity, free speech, and freedom of information is obvious.

For all the criticism levied at social media companies for inadequate resource allocation to content moderation efforts and for platforming harmful content, the perspective of leaving this task solely in the hands of the states with their tarnished record on upholding human rights is however also less promising.

It is clear that greater global stewardship is needed to ensure digital technology promotes human rights, inclusive sustainable development, and international stability. But the world has yet to develop adequate global frameworks to govern the digital domain.

The Santa Clara Pinciples

In 2018, a group of civil society organisations produced the first version of the Santa Clara Principles of content

moderation. The second revision of these principles was published in 2021 after an 18-month long open consultation process during the pandemic.

In the meantime, the biggest tech companies including Apple, Meta, Google, and Twitter have endorsed these principles, thus creating a strong platform for civil society organisations collectively and individually to base their content moderation advocacy on.

They consist of five foundational and three operational principles for how to ensure transparency and accountability in the moderation of user-generated content, give impacted users access to due process, and ensure fair, unbiased, proportional enforcement of community guidelines and terms of service while respecting users' rights:

"Foundational Principles are overarching and cross-cutting principles that should be taken into account by all companies, of whatever business model, age, and size, when engaging in content moderation. They set out each principle and guidance as to how to implement that principle. The Operational Principles set out more granular expectations for the largest or most mature companies with respect to specific stages and aspects of the content moderation process. Smaller, newer, and less resourced companies may also wish to use the Operational Principles for guidance and to inform future compliance."[26]

Foundational Principles

1. **Human Rights and Due Process:** Companies must integrate human rights, particularly freedom of expression, and due process in all stages of the content moderation process, including their automated processes.
2. **Understandable Rules and Policies:** The rules and policies of content moderation, including account suspension and deletion must be clear, precise, available and accessible.
3. **Cultural Competence:** Workers making decisions about user content must understand the language, culture, and political and social context of the posts they are moderating.
4. **State Involvement in Content Moderation:** It should be clear to users when state actors are involved in moderation of their content and whether such involvement was based on a legal framework.
5. **Integrity and Explainability:** Content moderation systems both human-driven and machine-automated should work reliably and effectively, with accuracy and non-discrimination, subject to assessment and auditing.

Operational Principles

1. **Numbers:** Companies must publish a comprehensive suite of quantitative information about the scope and scale of their content moderation processes.
2. **Notice:** Companies must in detail notify users whose content is deleted or whose account is suspended or deleted with the reason for the moderating action taken by the company.
3. **Appeal:** Companies must make channels available for users to appeal content moderation decisions, including human review of the appeal and possibility for users to present additional information.

The Santa Clara Principles are a set of standards established to create transparency and accountability, and even though they have been widely endorsed by the biggest tech platforms, they still need to be widely adopted.

Looking at the principles from the perspective of the WHRDs interviewed for this report, they serve as a solid framework and starting point to protect human rights defenders, however, they need to be specified and expanded in the following areas.

The lack of cultural and linguistic context for social media content moderators continuously came up in the interviews with DCA partners. Many experienced that reporting of issues fell through the cracks due to lack of prioritization of languages other than English, and that it was difficult to report threats and harassment in local languages and dialects. Also, threats and harassment can be veiled or masked through euphemisms, making it difficult for a content moderator without significant knowledge of the local context, slang, and discourse to see through and properly deal with, as is the case of content moderation in Palestine/Israel, which is moderated from Morocco. One interviewee said that a person from Morocco would never be able to understand the complexities moderating content between Palestine and Israel. The "Facebook Files" leak documented Meta's repeated content moderation failures in Ethiopia and how they have been traced to real-world violence.[27] The failures of Facebook in Ethiopia are a symptom of a deep geographic and linguistic inequality in the resources devoted to content moderation and which countries and situations social media platform companies deem relevant to focus on. In countries where the market share is low or where the market is less attractive, having qualified content moderators and staff in country is down prioritised, as in Nepal. "There are many words in Nepali language that are very vulgar. (...) people won't hesitate in using those words. When they are doing that, they hardly use the English language. (...) People tend to use those kinds of words to spread hatred. (...) They will write with English letters but still Nepali things."

Therefore, following the Santa Clara Principles includes

hiring sufficient and culturally competent staff to ensure content moderation is done with a proper understanding of language and cultural context. A Kenyan WHRD gives a perspective from first-hand experience.

"Some of the tech companies are quite understaffed, that is why it will remain a very huge challenge on the [African] continent and the global south. They have one or two individuals working for the whole of the continent, on issues to do with trust and safety. Trust and safety are the biggest data point - or entry data point for some of these tech companies. Therefore, they need to staff more!"

In some situations, during natural catastrophes, war, or conflict, special measures regarding the understanding and changing nature of online threats and harms need to be applied.

A Ukrainian WHRD explained in her interview that the online threat landscape her organisation worked in changed overnight with the Russian invasion, which resulted in a deluge of new cyberthreats.

"The issue of Digital security became very crucial and vital under the current situation in Ukraine. During the war, incident response[28] became one of our main activities, because there was a great increase in the number of cyberattacks on Ukrainian civil society."

Social media platforms like Meta and Twitter adequately responded to this threat, by setting up a special operations centre and allowing users to lock their social media profiles. This serves as an example of best practice for social media dealing with constant developing threats in conflict or war zones and is a model to be followed also in less high-profiled conflicts, to ensure user safety.[29]

Flagging of manipulated media and disinformation is also vital in conflict situations, but according to the Ukrainian WHRD, many larger social media platforms effectively silenced or censored information content about the war, indiscriminately hiding posts like "Russia", or the Ukrainian flag. Here, it is vital to work with localized fact checking initiatives to accurately flag false content, and a special operations centre assembled to monitor and moderate platforms in conflicts that have heightened risks of cyberattacks and disinformation should be adequately equipped to deal with fact-checking and manipulated media. The WHRD interviewee further suggested that the platforms reach out and collaborate more with civil society organisations to provide context.

"The local organisations can provide information about specific actors and explain why the things they do actually are a violation of the terms of the platforms."

The operational principles of **Notice** and **Appeal** are also crucial to WHRDs. Many users are still kicked off platforms with no explanation or meaningful process of appeal. The lack of a direct contact point is especially crucial for human rights defenders, who often experience deplatforming and hostile reporting.

In countries with lack of access to independent media, access to social media platforms is a crucial communications channel for many WHRDs, making them vulnerable to the censorship of social media platforms. This highlights the need for transparency around decisions on algorithmic repression of specific types of messaging and decisions to remove posts or profiles, as well as opportunities to appeal decisions.

While community guidelines are vital to the health and safety of social media users, they can also be misused for mass-reporting content creators, organisations and HRDs in bad faith, thus depriving them of their platform and restricting their free speech. Enacting the principles of notice and appeal can ensure that accounts are not restricted or removed without proper review and explanations, and that appeal options exist, making it harder to weaponise bad faith reporting and deplatforming against WHRDs.

An organisation like Access Now has long worked with Human Rights Defenders to restore access that was unjustly revoked. It's vital that major social media platforms continue to build relationships with these types of organisations and ensure better appeal options across territories.

A request that continuously came up from the WHRDs in the interviews was the lack of a direct contact to appeal to or request support from platforms, when being targeted by mass harassment and disinformation campaigns. When WHRDs expressed an easy access to Tech platforms, it was ad hoc, random and oftentimes built on personal connections. Big social media companies like Meta, Twitter, TikTok, YouTube, etc. should ensure a complaint mechanism in each territory, that can be a point of contact and dialogue for civil society and human rights defenders who are experiencing harassment, censorship, or other impediments to their work on the platforms.

Tech platforms should also prioritize collaborating with local fact-checking and research initiatives who can provide them with the necessary context to make moderation decisions.

Earlier examples of solutions and mitigations of online harms

implemented after input or pressure from HRDs, academics, citizens groups or experts include: YouTube's mass removal of extremist content from ISIS and QAnon, Meta's experiments with reporting features for sharing of illegal intimate images and Twitter's updated Terms of Service, which are the most comprehensive in outlining user protection and mitigating the spread of disinformation.[30] These are all examples of how social media platforms can vastly improve the safety and experience of their users, when working with relevant input from experts on online harms.

Foundational and operational principles of content moderation

Considering the number of users on the most popular social media platforms and the amount of content they upload continually, it has long been clear that moderating all that content, making sure it conforms to the platforms' community guidelines, and terms of service, i.e. content moderation at scale, is anything but a trivial problem. For that reason, content moderation has developed into a global, multi-million dollar industry in large part outsourced by the big tech companies to third-party service providers.

In practice, content moderation begins with reporting (or "flagging") of violating content by users of the platforms as well as by automated content recognition systems. These reports are then reviewed in quick succession by human content moderators who decide whether to allow or delete the content or whether to refer the report to a superior in cases of doubt.

In certain cases, content deletion is followed by the suspension or outright deletion of the infringing user's account. Complicated cases of reported content that are escalated to superiors can lead to changes or clarifications of content guidelines.

More and more, automated systems are being employed not only to flag content, but also to make judgment about whether to delete the content or not, without any human review.

A specific form of content moderation that often can go unnoticed is algorithmic downranking, as well as de-linking or depublication of user-generated content and account profiles, colloquially known as "shadow banning".

It is important to point out that both human content moderation workers and automated content recognition systems routinely make mistakes, resulting in large amounts of false positives and false negatives, where non-violating content gets deleted and violating content stays up.

Because the basis of content moderation is user reporting, the flagging tools can be weaponised as a form of harassment, where mass reporting of legitimate content results in its deletion, in effect amounting to censorship. Both the accidental and deliberate deletion of legitimate content affect otherwise vulnerable users disproportionally, e.g. LGBT+ individuals, women, racial and ethnic minorities, and sex workers. ■

[Photo: Jjumba Martin]

07. FINAL RECOMMENDATIONS

The findings in this report highlight the necessity of addressing the harassment, censorship, and silencing of WHRDs on social media platforms, and suggest improvements in legislation, enforcement and content moderation. In short, DCA has the following recommendations for improving content moderation online:

- **Multistakeholder collaboration is key.** Regulators, platforms and civil society need to collaborate and exchange information in order to find the best models for content moderation that detect online harassment and misinformation campaigns in due time and, at the same time, respect freedom of speech. This needs to be done at an international, national and local level.
- **Invest more resources in protecting human rights online in the global south.** Social media platforms that facilitate a public space need to invest more resources in the global south.
- **New regulation should differentiate between tech platforms,** not only in relation to size but also in relation to the products they are offering, the technology they are built on and the business models they provide.
- States need to live up to the UN Guiding Principles and hold tech platforms to account and **include a vision of online protection of human rights in their own, their foreign, trade and development policies** to combat online harassment and other human rights violations, polarisation and conflict.
- **Tech platforms need to share information and data with each other and collaborate** to protect WHRDs from online harassment campaigns.
- Tech platforms in general need to be **more approachable and more responsive.** There is a need for the platforms to create equal access for users in all countries where their service is offered.
- Tech platforms need to **practice due diligence when it comes to elections, polarisation and conflict** on the rise.

Aside from advocating for change with states and technology companies, **a larger cultural shift in attitudes towards women,** be they politicians, Human Rights Defenders or regular schoolgirls, is needed, so that the number of girls and women experiencing online harassment will be closer to 0% than the current 58%. Achieving this goal requires a joint effort by multiple stakeholders, including lawmakers and regulators in the states, activists and experts in civic spaces, journalists and producers in media, educators and experts in academia, and last but not least, the people working at the tech companies who built these platforms we have come to rely on. ■

[Photo: Kajta Kjar Levin]

08. NOTES AND SOURCES

1. *"Human Rights Council holds Panel discussion on online violence against women human rights defenders"*, Office of the United Nations High Commissioner for Human Rights, 2018.

2. As there is no single global entity that consistently compiles data on this issue, data from a variety of trusted sources has been used to inform the scope of the problem. The data quoted is by no means exhaustive, but looks to some of the more recent surveys and analysis of gendered online harm.

3. Facebook, Instagram, WhatsApp, TikTok, Twitter, and Snapchat were included in the survey. *"Abuse and harassment driving girls off Facebook, Instagram and Twitter"*, Plan International, 2020.

4. *"Crowdsourced Twitter study reveals shocking scale of online abuse against women"*, Amnesty International, 2018. https://www.amnesty.org/en/latest/press-release/2018/12/crowdsourced-twitter-study-reveals-shocking-scale-of-online-abuse-against-women/

5. *"Sexism, harassment and violence against women parliamentarians"*, Inter-Parliamentary Union, 2016. https://www.ipu.org/resources/publications/issue-briefs/2016-10/sexism-harassment-and-violence-against-women-parliamentarians

6. *"Attacks and Harassment: The Impact on Female Journalists and Their Reporting"*, International Women's Media Foundation, 2018. https://www.iwmf.org/wp-content/uploads/2018/09/Attacks-and-Harassment.pdf

7. *"Information series on sexual and reproductive health and rights: Women Human Rights Defenders"*, Office of the United Nations High Commissioner for Human Rights, 2020. https://www.ohchr.org/sites/default/files/Documents/Issues/Women/WRGS/SexualHealth/INFO_WHRD_WEB.pdf

8. Ibid

9. *"Troll Patrol India: Exposing Online Abuse Faced by Women Politicians in India"*, Amnesty International Decoders, 2020. https://decoders.amnesty.org/projects/troll-patrol-india

10. *"Situation of women human rights defenders"*, Report of the Special Rapporteur on the situation of human rights defenders, A/HRC/40/60, 2019.

11. **In the United States:** Congresswoman Katie Hill had private photos leaked as revenge by her ex-husband, forcing her to shut down her campaign: *"Former Rep. Katie Hill, who lost revenge porn lawsuit, files for bankruptcy"*, Seema Mehta, LA Times, 2022. https://www.latimes.com/politics/story/2022-07-12/california-former-rep-katie-hill-who-lost-revenge-porn-lawsuit-files-for-bankruptcy

    In Rwanda: A woman political candidate's fake nude photos went viral during her campaign: *"Fake nude photos were used to 'silence me', disqualified Rwandan candidate says"*, Stephanie Busari & Torera Idowu, CNN, 2017.

    In Finland: The Prime Minister's leaked party video went viral with accusations that she did drugs and critiques of her "indecent" dancing: *"Finnish PM says videos of her 'boisterous' partying shouldn't have been made public"*, Rob Picheta & Pierre Meilhan, CNN, 2022.

12. *"Jillian York: The global impact of content moderation"*, ARTICLE 19, 2020. https://www.article19.org/resources/the-global-impact-of-content-moderation/

13. *"Israel/Palestine: Facebook Censors Discussion of Rights Issues"*, Human Rights Watch, 2021. https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues

14. *"How Vietnam's 'influencer' army wages information warfare on Facebook"*, James Pearson, Reuters, 2021.

15. *"Inside Russia's Notorious 'Internet Research Agency' Troll Farm"*, Spyscape, 2022.
https://spyscape.com/article/inside-the-troll-factory-russias-internet-research-agency

16. *"Uncovering a Twitter Bot Army Mobilised Against Al Jazeera"*, Marc Owen Jones, Amanatech, 2017.
https://perma.cc/39MC-W75G

17. *"How Russian Facebook Ads Divided and Targeted US Voters Before the 2016 Election"*, Issie Lapowsky, Wired, 2018.
https://www.wired.com/story/russian-facebook-ads-targeted-us-voters-before-2016-election/

18. *"Cyber violence pattern and related factors: online survey of females in Egypt"*, Hassan et al., Egyptian Journal of Forensic Sciences, 2020.
https://www.researchgate.net/publication/339164810

19. *"Egypt's new cybercrime law legalizes Internet censorship"*, Reporters Without Borders, 2018.
https://rsf.org/en/egypt-s-new-cybercrime-law-legalizes-internet-censorship

20. *"Proposal For a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC"*, European Commission, 2020.
https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN

21. *"How the Supreme Court ruling on Section 230 could end Reddit as we know it"*, Tate Ryan-Mosley, MIT Technology Review, 2023
How the Supreme Court ruling on Section 230 could end Reddit as we know it | MIT Technology Review

22. *"Germany: Flawed Social Media Law"*, Human Rights Watch, 2018.
https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

23. *"The UK Online Safety Bill Attacks Free Speech and Encryption"*, Joe Mullin, Electronic Frontier Foundation, 2022.
https://www.eff.org/deeplinks/2022/08/uks-online-safety-bill-attacks-free-speech-and-encryption

"Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content", Jilian C. York et al., Electronic Frontier Foundation, 2019.
https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content

"Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act", Christoph Schmon et al., Electronic Frontier Foundation, 2022.
https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act

"Beyond the Brussels effect", Andrea Renda, Foundation for European Progressive Studies, policy Brief, March 2022 220301 beyond the brussels effect.pdf (feps-europe.eu)

24. *"A TikTok Trend You Can't Ignore: Addressing the Risks by Protecting Privacy and Bolstering Transparency"*, Allie Funk, Freedom House, 2023
A TikTok Trend You Can't Ignore: Addressing the Risks by Protecting Privacy and Bolstering Transparency | Freedom House

25. Most social media platforms looking to grow their user base are investing in moderation practices that are enforced by automation and AI rather than humans, e.g. as documented by the Wall Street Journal in an investigative series about Meta. Recent statements made by Elon Musk regarding his acquisition of Twitter also included a desire to automate content moderation.

"How Social Media Platforms' Community Standards Address Influence Operations", Jon Bateman, Natalie Thompson, Victoria Smith, Carnegie Endowment for International Peace, 2021
How Social Media Platforms' Community Standards Address Influence Operations - Carnegie Endowment for International Peace

"The Facebook Files", Newly Purnell et al., The Wall Street Journal, 2022.
https://www.wsj.com/articles/the-facebook-files-11642035385

"Documents detail plans to gut Twitter's workforce", Jeremy B. Merrill et al., The Washington Post, 2022.

26. *"The Santa Clara Principles: On Transparency and Accountability in Content Moderation"*, ARTICLE 19 et al., 2018.
https://santaclaraprinciples.org

27. *"Facebook's Content Moderation Failures in Ethiopia"*, Caroline Allen, Council on Foreign Relations, 2022
https://www.cfr.org/blog/facebooks-content-moderation-failures-ethiopia

28. *"Incident response"*, a term in cybersecurity, for identifying and investigating data breaches, intrusions, or related security events.

29. *"Facebook, Twitter highlight security steps for users in Ukraine"*, Sheila Dang, Reuters, 2022.
https://www.reuters.com/world/europe/facebook-is-letting-users-ukraine-lock-their-social-profiles-security-2022-02-24/

30. On Twitter moderation and the potential effects of Elon Musk's takeover: *"Elon Musk is wrong: Research shows content rules on Twitter help preserve free speech from bots and other manipulation"*, Filippo Menczer, The Conversation, 2022.
https://theconversation.com/elon-musk-is-wrong-research-shows-content-rules-on-twitter-help-preserve-free-speech-from-bots-and-other-manipulation-182317

On Facebook's attempts to combat "revenge porn":
"Facebook is expanding its unconventional approach to combating revenge porn", Nick Statt, The Verge, 2018.
https://www.theverge.com/2018/5/23/17386972/facebook-revenge-porn-combat-uploading-hashing-nude-photos

On YouTube's mass-removal of extremist content:
"YouTube to Remove Thousands of Videos Pushing Extreme Views", Kevin Roose & Kate Conger, The New York Times, 2019.
https://www.nytimes.com/2019/06/05/business/youtube-remove-extremist-videos.html

DanchurchAid
Meldahlsgade 3, floor 3 & 4
1613 Copenhagen V
Denmark

Telephone +45 33 15 28 00
mail@dca.dk

nødhjælp.dk
danchurchaid.org