Violence Against Women and Girls (VAWG) Code of Practice

Professor Clare McGlynn

Professor Lorna Woods

Contents

Introduction

Why has this guidance been created?

This guidance has been created due to the high prevalence of Violence Against Women and Girls (VAWG) perpetrated in the digital sphere. This includes technology-facilitated abuse (activities carried out with the use of technology and communication equipment, including hardware and software) enabling abusers to stalk, harass, surveil, and control victims. It is prepared with regard to the Online Safety Bill and the obligations placed on regulated providers, as set out in the Bill, to prevent harm against adult and child users. This is a 'living' document that will continue to evolve as the Online Safety BIll progresses through Parliament. It has been prepared by Carnegie UK, The End Violence Against Women Coaltion, Glitch, NSPCC, Refuge, 5Rights and academics Lorna Woods and Clare McGlynn.

Who is this guidance for?

This Code of Practice provides detailed guidance for all tech companies to help them understand and respond to the breadth of online violence against women and girls. This guidance is targeted at specific adverse human rights impacts arising from specific technology product-types. The basic principle is simple. The UNHRC B-Tech project[1] makes clear that this includes:

> *"A company identifying whether and how the design, development, promotion, deployment and use of its products and services could lead to adverse human rights impacts"*
>
> *"Beyond product design, business processes – and in the context of social media community standards and moderation standards – should also be included."*[2]

All online services, and in fact the whole of society, have an obligation to tackle gender-based violence and can achieve more in collaboration.

What is Violence Against Women and Girls?

The term Violence Against Women and Girls (VAWG) is used to mean acts of violence or abuse that are targetted at, and disproportionately affect, women and girls.

The Council of Europe Convention on preventing and combating violence against women and domestic violence (also known as the "Istanbul Convention"), to which the UK is a signatory (but still has yet to ratify a decade later), defines VAWG as:

> *"a violation of human rights and a form of discrimination against women and shall mean all acts of gender-based violence that result in, or are likely to result in, physical, sexual, psychological or economic harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life" (Article 3)*

1 OHCHR, B-Tech: "Identifying Human Rights Risks Related to End-Use"; (2020) https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf

2 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (A/74/486), 19 October 2019, para 92, available: https://www.undocs.org/A/74/486 [Accessed 22 July 2021] .

VAWG is rooted in gender inequality and men's sense of entitlement. This is the same as abuse experienced in the offline world. This gender inequality intersects with multiple, overlapping structural inequalities. These inequalities include race, ethnicity, religion, sexual orientation, gender identity, disability and age and other characteristics that shape women and girls' experiences of VAWG online.[3] For example, Black and minoritised women face disproportionate threats and experiences of online VAWG due to misogynistic racism.[4] Any consideration of online VAWG must have an intersectional analysis of how this abuse is being perpetrated and experienced at its core.

Non-binary people, as well as transgender people, gender non-conforming people, and people with diverse sexualities are targeted for forms of gender-based violence based on their sexuality, gender identity and/or gender expression. This targeting is also based in gender inequality and men's sense of entitlement. While holding the focus on VAWG, the principles outlined here apply across all forms of gender-based violence and include all groups who may be affected.

For the purposes of this Code, online VAWG should be understood as part of a continuum of violence against women and girls which is not a solely "virtual" phenomenon separated from violence "in real life".[5] It does not exist in a vacuum, but both stems from, and sustains, multiple forms of offline violence. It is often difficult to distinguish the consequences of actions that are initiated in digital environments from offline realities, and vice versa.[6] For example, research by Women's Aid found that 85% of women who experienced online abuse from a partner or ex-partner said that it was part of the pattern of abuse they also experienced offline.[7] Almost 1 in 5 women (17%) who experienced domestic abuse on social media felt afraid of being attacked or subject to physical violence because of this.[8]

What are the wider impacts of VAWG?

Online VAWG is part and parcel of the structural and systemic inequality of women and girls, as reported by the UN Special Rapporteur on VAWG:

> *"Women and girls across the world have increasingly voiced their concern at harmful, sexist, misogynistic and violent content and behaviour online. It is therefore important to acknowledge that the Internet is being used in a broader environment of widespread and systemic structural discrimination and gender-based violence against women and girls."*

This reinforces the fact that the impact of VAWG is wider than the individual instances; it also creates societal and cultural harm, with significant consequences for everyone. Preventing and combating VAWG is key to tackling the online exclusion of women and girls, and to supporting access to safe and inclusive digital spaces. The right of women and girls to safely access, navigate and enjoy online spaces, and to engage and express themselves free from fear of abuse, must be a guiding principle of regulated providers' attempts to prevent and respond to online VAWG.

This Code of Practice is also rooted in international standards and obligations for the prevention of VAWG. Article 5, paragraph 2, of the Istanbul Convention requires States Parties to take the necessary legislative and other measures to exercise due diligence to prevent, investigate, punish and provide reparation for acts of violence covered by the scope of this convention that are perpetrated by non-state actors.

---

3 Glitch-and-EVAW-The-Ripple-Effect-Online-abuse-during-COVID.pdf (endviolenceagainstwomen.org.uk)

4 https://www.amnesty.org.uk/online-violence-women-mps

5 See further, EU Advisory Committee on Equal Opportunities for Women and Men (2020), *Opinion on Combatting Online Violence Against Women*, 1 April 2020.

6 https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2020/Brief-Online-and-ICT-facilitated-violence-against-women-and-girls-during-COVID-19-en.pdf

7 https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/onlinesafety/

8 Unsocial Spaces, Refuge Report, October 2021

The Council of Europe group GREVIO (Group of Experts on Action Against Violence Against Women and Domestic Violence) has stated[9] that they consider "this obligation to cover all expressions of violence against women, including digital expressions and violence perpetrated with the help of or through technology." Similarly, the United Nations has emphasised that the Convention on the Rights of the Child calls on States parties to take legislative and administrative measures to protect children from violence in the digital environment... Such risks include... gender-based violence'. Further, States parties are called on to 'take proactive measures to prevent discrimination on the basis of sex'.[10]

What forms of online abuse does Violence Against Women and Girls Include?

VAWG encompasses a wide range of acts. In a systems-based approach listing specific types of content and/or practices is not always helpful, and there cannot be an exhaustive list as new forms of harm, and new terms for it arise regularly. It is stressed that the below list is illustrative only.[11]

- cyberharassment, including cyberbullying, online sexual harassment, unsolicited receipt of sexually explicit material, mobbing and dead naming;
- cyberstalking;
- ICT-related violations of privacy, including the accessing, recording, sharing, creation and manipulation of private data or images, specifically, including image-based sexual abuse non-consensual creation or distribution of private sexual images, doxxing and identity theft;
- recording and sharing images of rapes or other forms of sexual assault;
- remote control or surveillance, including by means of spy applications on mobile devices;
- threats, including direct threats and threats of and calls to violence, such as rape threats, extortion, sextortion, blackmail directed at the victim, their children or at relatives or other persons who support the victim and who are indirectly affected;
- sexist hate speech, including posting and sharing content, inciting to violence or hatred against women or LGBTIQ people on the grounds of their gender identity, gender expression or sex characteristics;
- inducements to inflict violence on oneself, such as suicide or anorexia and psychic injury;
- computer damage to files, programmes, devices, attacks on websites and other digital communication channels;
- unlawful access to mobile phones, email, instant messaging messages or social media accounts;
- breach of the restrictions on communication imposed by means of judicial orders;
- the use of technological means for trafficking in human beings, including for sexually exploiting women and girls

9 GREVIO, General Recommendation No. 1 on the digital dimension of violence against women, 20 October 2021 https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147

10 UN Convention on the Rights of the Child, General Comment 25: https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation

11 https://www.europarl.europa.eu/doceo/document/TA-9-2021-0489_EN.pdf

Online VAWG should also be understood as including:

- material that can be consumed as child sexual abuse material, or may further a sexual interest in children;
- material that can be used by perpetrators to signpost others to child sexual abuse content, through 'breadcrumbs', images, or hyperlinks;
- material designed or shared to trigger past victims and/or traumatise new ones;
- intentional actions which cause harm to individual women and girls, or women and girls as a collective; and
- extreme pornography which includes explicit and realistic depictions of rape, non-consensual penetration, bestiality, necrophilia and life-threatening injuries.

The prevalence of pornographic content online which depicts violence against women, coercive or non-consensual sexual practices, and/or the humiliation, degradation and exploitation of women and girls serves to normalise and minimise violence against women and girls. It is also a significant contributor to a broader societal culture which serves to reinforce the inequality of women and girls and provides a conducive context for online and offline abuse to thrive.

What is needed to prevent and tackle VAWG?

Preventing and tackling VAWG requires investment and prioritisation. VAWG cannot be tackled by online services in isolation. Online services must engage with VAWG specialists to further their understanding of the risks and harms that exist, and to work together to tackle these types of harms. Wherever possible, online services must support civil society efforts to tackle online VAWG and cooperate with endeavours to create online environments which uphold the rights of women and girls to participate without fear of violence.

Measures to prevent online VAWG must take account that the right to privacy and anonymity will be a central concern for many survivors of VAWG. This guidance cautions against safety measures which predominantly rely on the surveillance of users and content takedown, as these actions can disproportionately affect marginalised groups, including Black and minoritised survivors and LGBT+ survivors. Decision-makers should ensure action is proportionate and justified, by way of reference to human rights and equalities legislation.

Organisations and groups with expertise around VAWG must therefore be provided with opportunities to collaborate in developing safety by design approaches, identifying, defining, and responding to harm. This should incorporate organisations which represent minoritised survivors, migrant, Deaf and Disabled and LGBT+ survivors.

The training of moderators and employees, including executives, must be driven by expertise from the VAWG sector, rooted in experience and evidence, and this expertise must be remunerated adequately. The diversity – or the lack of it – of teams within tech companies must also be considered as potential contributing factors to violence against women and girls on platforms.[12]

12 https://www.safetytechnetwork.org.uk/diversity-inclusion-and-fairness-in-safety-tech-expert-panel-provides-key-insights/

Section 1) Responsibility, risk assessment, mitigation and remediation

(1) Regulated services should have a specific policy commitment to prevent and take action to combat VAWG arising on their service. This commitment should be endorsed by the UK leadership of the organisation and a board member, or person reporting into the board, appointed to be accountable for delivering it. The policy should be informed by specialist VAWG expertise. It should clearly set out the values of the regulated service.

(2) (a) Regulated services should carry out a suitable and sufficient assessment as to the risk of VAWG-related harm, taking into account international human rights standards, obligations and best practice. Risk assessments must take into account and mitigate potential harms arising from intersecting inequalities. This means the particular risks of harm to people with more than one or overlapping characteristics that typically experience discrimination and oppression, arising from the operation of the service or any elements of it. The risk assessment should be accompanied by a mitigation plan that addresses the issues raised in this Code.

(b) The risk assessment should not solely consider individual risks to individual users but also consider broader social and cultural harm, such as the ways in which all women are affected by the threat of violence and harm even if they have not directly experienced it themselves.

(c) The risk assessment should be carried out before any new service or any new feature is made available. It should include consideration of how different types of content are shared and practices carried out on the platform, and by whom.

(4) Service providers should identify suitable metrics to assess the appropriateness and success of the mitigation plan overall, and in relation to each set of risks and use them to assess effectiveness of the mitigation plan regularly (at least annually) and revise the mitigation plan accordingly.

(5) The risk assessment should be reviewed by the service provider on an ongoing basis or, if there is reason to suspect that it is no longer adequate or complete; or there has been a significant change in the matters to which it relates. Where as a result of any such review changes to a mitigation plan are required the service provider should make them.

(6) Risk assessments and mitigation plans should be recorded, retained for a period of no less than three years and published on the service provider's website in an accessible manner.

(7) All measures taken in the following guidelines, including the metrics at (4), should feed back into the risk assessment as it evolves.

Commentary

Corporate responsibility

Active leadership in companies is necessary to effectively combat VAWG that exists on or is facilitated by their service.

The United Nations Guiding Principles on Business and Human Rights (UNGP) specifies that companies should have

> *"a human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights."*[13]

As the Organisation for Economic Co-operation and Development (OECD) guidance notes:

> *"Due diligence is risk-based. The measures that an enterprise takes to conduct due diligence should be commensurate to the severity and likelihood of the adverse impact. When the likelihood and severity of an adverse impact is high, then due diligence should be more extensive."*[14]

An effective strategic approach to tackle VAWG is likely to include –

- A clear policy statement
- Strategies for corporate oversight
- Clear and effective systems and processes responsible for addressing harms against women and girls
- A company governance structure with allocated roles and responsibilities for discharge of functions of the code that includes Board accountability
- A mapping exercise to identify roles and departments relevant to discharge of functions of the code
- Clear responsibility for delivery of the risk mitigation plan as well as risk assessment process with reports required to individual(s) accountable for VAWG within the company.
- Meaningful engagement with specialist VAWG expertise

Undertaking risk assessments for VAWG

Online services should take a systemic approach to the identification and mitigation of reasonably foreseeable impacts on women and girls resulting from the design and operation of their services. Risk assessments should be founded in an intersectional understanding of how harms are directed at, and felt by women and girls with multiple and overlapping characteristics. The overall aim of a risk assessment should not be to guard against liability, but to try and prevent VAWG taking place on regulated providers' platforms in the first place, and thus reduce the amount of harm experienced by women and girls online.

VAWG risk assessments should be broader than individualised harm in recognition of the fact VAWG is not only harmful when it poses risks to individuals but that there is a wider social harm that exists from women and girls' awareness of the threats and harm they may face online. Risk assessments should take a 'rights based approach' and consider how women and girls' right to use the internet free form harm could be curtailed by the service.

13 UNGP 15 OHCHR https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

14 OECD, Responsible Business Conduct: Guidelines for Multinational Enterprises p17

Companies conducting a risk assessment for VAWG for the first time should also evaluate their existing risk management practices and processes, practices in relation to human rights impact assessments, and data protection/privacy impact assessments, to consider any gaps or tensions in those approaches and ensure that there is appropriate governance[15]. Particular attention should be paid to reliance on techniques driven by machine learning and artificial intelligence, and the well-known questions around the design and deployment of Machine Learning and AI.[16]

The complexity of the risk assessment will vary according to the size of the business, its business model, its values (including those found in its Community Standards or Terms of Service[17]) and the profile of its users (e.g. a service with a significant number of children as users).

Companies should centre the need for equalities and rights in a risk assessment process so that intersectional inequalities are not overlooked or minimised.

The OECD Guidance on due diligence for responsible business conduct[18] provides a good framework for risk assessment for VAWG, as for many other issues, as does ISO 310000.[19]

VAWG risk assessment process

The VAWG risk assessment process should be based on data and research, and an assumption that VAWG may be perpetrated on its service and is a responsibility of the provider. This will involve gathering data in a systematic manner[20] about what is happening on the service. For example, data on the nature of user complaints and how they are dealt with (recognising that complaints are not the only, or even an accurate, measure of what is happening on the service). Data may also include the results of any testing on the product to understand the nature of the problem, as well as its scale, context and triggers. The VAWG risk assessment should understand not just the fact that women and girls face particular harm (in general and in relation to particular technologies such as nudification apps), but that they face a greater likelihood of encountering harmful content and that the perpetrators of harm may not be evenly spread across the platform.

1) Regulated services should consider sharing best practice on risk assessments with other technology companies.

2) Considerations must be made for girls' age and gender as well as other possible protected characteristics that may be identifiable through 'know your user' processes that are likely to affect the way they are targeted for and impacted by VAWG

15 For guidance on human rights-friendly governance procedures, generic to any company type see the UNGPs Interpretative Guide and for technology companies the OHCHR B-Tech project.

16 Council of Europe 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on 8 April 2020) https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154

17 The set of rules about expected behaviour on a platform or service, usually against which the platform enforces sanctions

18 'OECD (2018), OECD Due Diligence Guidance for Responsible Business Conduct'

19 ISO 31000: 2018 *Risk Management – Guidelines*; see also ISO Guide 73, *Risk Management – Vocabulary*; see also The International Finance Corporation, "Guide to Human Rights Impact Assessment and Management" (2010), available here: https://www.unglobalcompact.org/library/25

20 See Danish Institute for Human Rights in collaboration with the Human Rights Centre at University of Essex 'Guidance on Human Rights Impact Assessment of Digital Activities' (2020), available: https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities

3) The risk assessment must have a specific framework for an equalities analysis of risk against women and girls with overlapping protected characteristics and who experience intersectional inequalities. This must be a core part of VAWG risk assessment, not an add-on.

4) Where a risk assessment identifies a new or non-designated risk to women and girls, the online service would be expected to notify Ofcom of the kind of content identified and the prevalence of the content.

5) Regulated services must risk assess how content is shared on the platform, including but not limited to public news feeds private news feeds, private messages and private groups.

A separate risk assessment is required for services which are accessed by children, which should take a gendered approach, in line with this guidance.

Bringing victims' experience into the risk assessment process

Understanding the lived experience of those at risk of harm from VAWG is important if an effective risk assessment is to be delivered and appropriate mitigating actions implemented. Service providers should engage with VAWG specialists and civil society in tackling VAWG harms on platforms. Too often companies have claimed to rely on internal VAWG expertise but this does not translate into wider practices across the services and is often not transparent. External VAWG organisations, including services led "by and for" minoritised and marginalised women, can provide the greatest source of insight and evidence on high level emerging issues. Internal VAWG expertise may be best placed to translate high level VAWG principles and issues into technical solutions. VAWG panels and governance forums may be appropriate to give VAWG issues legitimacy and ensure that internal VAWG champions are involved in decisions.

Risk mitigation – general

Risk mitigation is essential to stop the harms outlined in the risk assessments. A risk mitigation plan should be drawn up to address identified risks, including built in risk-mitigation strategies that support safety by design practices (see section 2).

The plan should take into account different forms of VAWG and respond appropriately and proportionately. Appropriate and sufficiently granular metrics should be identified to assess the effectiveness and success of the mitigation plan. Regulated services should work with VAWG experts to create mitigation plans and to identify these metrics. The plans and metrics should be agreed with Ofcom who will then assess the risk mitigation effectiveness using those metrics. Where the risk assessment shows user detriment and/or a broader detriment to women and girls, regulated services should halt the rollout of the product/technology or implement appropriate and effective risk mitigation, for example through better safety by design. Success measures for risk mitigation, in addition to harm prevention and reduction, should include ensuring that marginalised groups are not disproportionately adversely affected by plans. Companies should engage VAWG experts for detailed discussions on thresholds for risk assessments, with input appropriately remunerated.

VAWG, and a rights-based systems approach to mitigation

A systems approach

> *"recognises that the platforms, as well as being in a gatekeeper role, are not neutral as to how people discover and create content. Choices made by the platforms about how they design their services affect the content seen (e.g. default to autoplay, curated playlists, data voids and algorithmic promotion) and even produced (e.g. through financial incentives for content creators, or the feedback loop created through metrification; emojis create a new shorthand for communication)."*[21]

A systems and processes approach to protecting women and girls from gender-based harm must consider the different ways in which harm is caused by regulated services, and address these through design features and product changes. Measures to address such harm can be broadly grouped as:

- action to prevent users from discovering or being exposed to material that may cause them harm;
- action to reduce the escalation and amplification of content online that can contribute to, and compound, harm experienced online;
- Actions to reconcile the connections between online and offline experience of VAWG (both direct and indirect)

Measures, where possible, must focus on acting preventatively rather than remedially. Ultimately, these systems-based interventions can help to balance responses to removal with a proactive and risk-responsive user design, allowing for protection from gender-based harm that does not rely solely on content removal alone.

Section 2) Safety by Design

(1) Regulated providers must implement appropriate "safety by design" technical and organisational measures, including but not limited to those detailed in these Guidelines. The intended outcome is to

(a) minimise the risk of those harms arising from VAWG content and practices

(b) mitigate the impact of those that have arisen,

(c) enhance women and girls' freedom online

taking into account the nature, scope, context and purposes of the online platform services and the risks of harm arising from the use of the service.

21 https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/11/25105121/UN-Hate-Speech-draft-v.05a-1.pdf

(2) Companies must ensure and be able to demonstrate their systems are safe by design, including addressing the following concerns:

(a) Taking an appropriate and proportionate approach to the principle of knowing your client [KYC] to address VAWG harms spread by those using multiple, false, or anonymous identities.

(b) Ensuring that young users' settings are set to safety by default.

(c) Ensuring algorithms used on the service do not cause foreseeable harm through promoting hateful content, for example by rewarding misogynistic influencers with greater reach, causing harm both by increasing reach and engagement with a content item.

(d) That speed of transmission has been considered, for example methods to reduce the velocity at which intimate images can be non-consensually shared and therefore the risk of cross-platform contamination.

(e) Actors cannot take advantage of new or emerging tools to cause harms to women and girls. For instance –

- deep fake or audio-visual manipulation materials.
- nudification technology.
- bots and bot networks.
- content embedded from other platforms and synthetic features such as gifs, emojis, hashtag.
- other new technology

(f) Consideration of the circumstances in which targeted advertising may be used and oversight over the characteristics by which audiences are segmented.

(g) Account security systems which enable survivors of abuse, who are hacked and locked out, to recover their accounts.

(h) Systems for cross-platform co-operation to ensure knowledge about forms of offending that may present a foreseeable risk of harm in relation to attacks of those with protected characteristics.

(i) Use of tools including, but not limited to, prompts which clarify or suggest an individual's intended search.

(j) Policies concerning advertising sales in respect of promoting harmful content or for malicious intent in respect of those with protected characteristics.

Commentary

General principles relating to safety by design

Safety in the context of this Code is the reduction of VAWG-related harms and violations of human rights.

Tech companies should acknowledge and respond to the fact that their products can facilitate, and even encourage, harm. 'Safety' must be understood as a context that enables all women and girls to exercise their freedom of expression online and freedom of access to platforms without fear of VAWG. It is an approach that recognises that women and girls already remove themselves from online spaces and refrain from expressing their views. It should also acknowledge that women and girls also

currently have to exercise a degree of "safety work"[22] that inhibits and curtails their experiences and free expression, and so should not place the burden of doing further "safety work" on women and girls. Similar to the relationship between privacy by design and privacy enhancing technologies (PETs), there is a link between safety by design and the emerging field of 'safety tech'.[23] Safety by design requires that harm considerations be built in, not bolted on and should itself be rights respecting (including the right to privacy). Where safety tech is supplied by third parties supply chain considerations (see Guideline 8) apply.

DCMS in their safety by design guidance outline four principles:

- Users are not left to manage their own safety;
- Online services must consider all types of users;
- Users are empowered to make safer decisions; and
- Online services are designed to keep children safe.

In addition, the following principles should be followed:

- maximum safety within the platform should be the default (even if users can choose to change these settings);
- safety is to be embedded into the design of the product;
- allowing as much functionality as possible (avoiding unnecessary trade-offs);
- for the full life-cycle of the service;
- to include transparency and to expect user-centric and rights-respecting choices.

Safety by design and VAWG

Currently, online services' broad lack of understanding of the specific issues that impact women and girls online has resulted in the creation of online environments and platforms where gender-specific harm has become ingrained in the user experience. Regulated services must engage appropriate systems and processes that not only work to prevent women and girls experiencing VAWG on their platforms, with particular recognition of intersecting inequalities and their impacts, but also work to increase women and girls' sense of safety and freedom. Regulated services must use an understanding of the behaviours of perpetrators of harm, and the ways in which they target women and girls when building and tailoring preventative measures.

Measures must include both design features that can work to prevent harm as well as product and service decisions that mitigate risk, such as account security systems which enable survivors of abuse who are hacked and locked out to recover their accounts, or systems which make evidence gathering difficult in an investigation (for example, self-destructing content).

22 Vera-Gray, F. and Kelly, L. (2020) Contested gendered space: public sexual harassment and women's safety work. International Journal of Comparative and Applied Criminal Justice https://www.tandfonline.com/doi/full/10.1080/01924036.2020.1732435

23 United Kingdom Government, "Safety tech providers deliver products and services that enable safer online experiences for citizens" https://www.gov.uk/government/publications/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech. See also an attempt to align global trends in safety tech:
Connie Moon Sehat, "Advancing Digital Safety: A Framework to Align Global Action", World Economic Forum, 29 June 2021. Available here: https://www.weforum.org/whitepapers/advancing-digital-safety-a-framework-to-align-global-action

Preventative measures must consider the role of algorithmic product decisions, the use of technology to monitor and remove harm from services, and the role of friction built into the user experience to both protect from and prevent VAWG and content and/or practices that facilitate harm.

Understanding user groups on the platform, and any overlapping characteristics, should be used for the primary purpose of protecting users. It must not be used as a way of increasing screen time or revenue to the detriment of user wellbeing.

Section 3) Access to the online service, terms of service and content creation

Terms of Service

(1) Regulated services should have in place Terms of Service which are clear and accessible by all likely users; this includes being age-appropriate and accessible for those with disabilities and different access needs. The terms of service should include how the service responds to VAWG, including actions taken to prevent VAWG, and be visible to would-be users before they sign up to the service. Community standards should also be visible and should, where relevant, cover the content of advertising.

(2) Regulated services should undertake regular, systemic reviews of their Terms of Service and Community Guidelines to ensure that they remain up to date, effective, and proportionate.

(3) To ensure Terms of Service and Community Guidelines are effective, regulated services need to review how they are operating and how they are enforcing them.

Account Creation

(4) Regulated services should ensure, and be able to demonstrate, that their sign-up processes have taken an appropriate and proportionate approach to the principle of "knowing your client" (KYC), both in relation to users and advertisers.

Content creation

(5) Regulated services should risk assess the tools for the creation of content – this includes but is not limited to bots (including chatbots), bot networks, deepfake or audiovisual manipulation materials, content embedded from other platforms and synthetic features such as gifs, emojis, hashtags.

Commentary

Guideline 3 concerns one of the basic building blocks of safety by design: a sign-up process and tools to create content, as well as Terms of Service.

Terms of Service

Terms of Service constitute the contract between the service provider and the user. They are important in communicating the service provider's values. As such, they may include community standards (though sometimes Terms of Service and Community Standards are used interchangeably) or acceptable use policies, understood as the content and behaviour rules the provider will enforce. The Community Standards should make clear the service provider's position on VAWG.

This is not the same as saying, however, that platforms must actively seek out criminal content, or monitor generally[24]. Such general monitoring has adverse impacts for all users' freedom of expression and privacy and would be very difficult, if not impossible, to justify. There is a need to ensure that the Terms of Service are not rendered meaningless and that there is some mechanism that is proportionate and appropriate to ensure that they provide a realistic expectation for the user of the types of content and behaviour that they will and will not encounter on the service.

Terms of service should be easily visible before a user signs up to the service, be easy to understand (by the age groups using the service) and be available in languages used by the service's users. This is important as part of transparency, but also to hold service providers and users to account.

Terms of Service and Community Guidelines should be kept under review, and revised where appropriate taking into account not just changes in external context but also learning from risk assessments, metrics on effectiveness of mitigation plans and complaints and moderation processes as well as any codes and or guidance from OFCOM.

For regulated services to effectively address the risk of VAWG, their terms of service must explicitly state what activity and material they determine constitutes VAWG and how they will deal with it. Most importantly, services must then enforce these principles and ensure the Terms of Service are effective and operational.

Terms of service must reflect the harms that occur to women and girls, ensuring systems and processes are continually informed by victims' perspectives and safeguarding best practice. This information might, for example, come from the internal expertise within the company or third-sector partners who provide advisory input. The provider must also explain how terms are developed, enforced, and reviewed, and the role of victims' groups and civil society in developing them.

The Terms of Service must explain the steps that regulated services will take if the terms of service are broken by users and be enforced by the online service. Evidence must be kept on individual cases, in line with GDPR requirements, regardless of the final decision. Within the service itself providers must ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms, and that these tools are kept updated.

See also requirements in Guideline 5 regarding moderation and Guideline 7 regarding transparency.

24 Note there is a difference between monitoring (for example, via an upload filter) which looks for specific content (such as on the basis of hashes or watermarks) and that which searches communications generally.

Account creation

There has been much concern about anonymous accounts and their role in online abuse and VAWG.[25] Guideline 3 refers to KYC processes but does not require regulated services to ban anonymous accounts. Indeed, it should be recognised that anonymity is an important and valued tool for the protection of women and girls, particularly those from marginalised communities, and survivors of VAWG such as rape and domestic abuse – as well as for whistleblowers and dissenting voices.

Rather, the Guideline expects the regulated provider to recognise the risk of people abusing anonymity to direct online violence towards women and girls and take steps to mitigate that risk, whether in terms of account verification, or through other interventions, (for example, enhanced user self-protection tools or reporting mechanisms).

This could be particularly relevant for services where there is user generated and user uploaded pornography, and the high risk of image based sexual abuse being perpetrated as part of that. To mitigate this potential harm, services should require user verification before uploads and require users to confirm they have consent from everyone depicted in the content to upload. This should be accompanied with messaging that informs them it is a criminal offence to upload material without the consent of those depicted, including content in violation of copyright and that the platform will take action against users for doing this.

Service providers should assess the risk of harm arising through VAWG from fake identities (for example those used for catfishing[26] or sock puppet accounts[27]), whether multiple accounts per person are permitted (and in what circumstances) and whether bots should have accounts[28] and then take proportionate steps to address these risks. Service providers should consider whether those who have been banned (for a period) from the service should be prevented from circumventing that ban for the purpose of causing harm. This should not interfere with users' right to appeal bans through the appropriate channels.

Service providers should also seek to understand any risks created by networks of accounts (for example coordination and amplification of posts). The concern is the way such networks increase not just the spread but also the speed of dissemination of abuse including across different platforms For example, misogynistic abuse and the "incel" movement. In this context, service providers could seek to understand who are the direct and indirect instigators and beneficiaries of such speech, as well as seeking to understand who is operationalising those messages and how (bots, sock puppets networks and false identities etc). Some individuals or small groups of individuals might be significant nodes in networks of misogynistic or VAWG-related behaviour that are amplified within the service[29]. Companies should have a transparent process for managing such individuals, carrying out the necessary balancing of human rights.

---

25 UK Parliament debate: Online Anonymity and Anonymous Abuse Volume 691: debated on Wednesday 24 March 2021 Available at https://hansard.parliament.uk/commons/2021-03-24/debates/378D3CBD-E4C6-4138-ABA6-2783D130B23C/OnlineAnonymityAndAnonymousAbuse

26 Where a person creates a fake identity to take advantage of another user

27 An online identity used for deception, often for the purpose of talking about or to themselves while pretending to be another person; the term is now used more broadly to include those manipulating public opinion, to circumvent restrictions, such as viewing a social media account that they are blocked from, suspension or an outright ban from a website. They are different from pseudonyms.

28 Julia Hass, Freedom of the Media and Artificial Intelligence, OSCE 16 November 2020, p. 4, available: https://www.osce.org/files/f/documents/4/5/472488.pdf [accessed 26 July 2021].

29 Renee DiResta et al., New Knowledge, The Tactics & Tropes of the Internet Research Agency 42 (2019); Brian Fishman, *Crossroads: Counter-Terrorism and the Internet*, (2019) 2 Tex. Nat'l Sec. Rev. 82, 86–87. https://www.counterhate.com/disinformationdozen [Accessed 22 July 2021]

In an interconnected world, service providers might factor into their risk assessments whether and how the individuals could spread VAWG on other services.

Content creation – service design that might increase VAWG

Each service is designed to allow and incentivise a user to create content in a different way. How content creation is designed can affect the risks of VAWG being created and disseminated. Features such as metrics or financial incentives based on popularity should be considered in relation to the motivation(s) of the creator. Outrage and content that plays on the biases of users (including sexism and misogyny) seemingly drive engagement (as clickbait headlines show), and there is a risk of cycles of ever increasingly outrageous content to drive likes and upvotes,[30] which can cause psychological harm. In some cases, the creation of highly harmful content can produce engagement and profit for the social media platform.[31] As a result of these financial incentives some 'content creators' might choose to create harmful content in pursuit of engagement and profit, but others will for ideological and recruitment purposes such as the incel communities. Addressing some of the concerns around content curation and recommender tools (see Guideline 4) may help, but services providers should seek to understand if there are features of the platform that might be exploited.[32]

The operation of these social platforms has led to the emergence of highly popular new communications media such as hashtags, emojis, photo-filters, voice notes etc. Service providers have often adopted these and encouraged their use in content creation to the extent that they become a major feature of some services. Platforms should be attentive to the fact that these methods can be abused to target women and girls, for example with misogynistic or abusive content[33], and can be hard to moderate using methods designed for text-based moderation. Service providers should include such media in their risk assessment and mitigation plans. For example, in relation to Instagram, the Center for Countering Digital Hate research[34] shows that 1 in 7 voice notes within their participants' data is abusive, and yet [it is not possible to] report them.

Disrupting bad actors

Regulated services must tackle upstream harm by disrupting bad actors who contribute to VAWG by acting in a harmful manner, through the production and distribution of harmful content, contact or conduct on their services. This may include considerations such as: barriers to producing new harmful content; action to prevent the facilitation of harmful contact such as grooming and online exploitation; mechanisms to prevent bad actors directing material at certain users, with the intention to cause harm.

---

30 W. J. Brady et al 'How Social Learnings Amplifies Moral Outrage Expression in Online Social Networks' (2021) (paper under review, available: https://psyarxiv.com/gf7t5/); Soroush Vosoughi, Deb Roy and Sinan Aral, "The Spread of true and false news online" (2018) 6380 Science 1146-51, DOI: 10.1126/science.aap9559

31 "Despite promises to keep users safe, we show how Big Tech itself makes up to $1 billion a year in advertising and other revenues from this industry, which threatens the effectiveness of a future Coronavirus vaccine.": Centre for Countering Digital Hate, https://www.counterhate.com/anti-vaxx-industry

32 DRFLab, "#InfluenceForSale: Venezuela's Twitter Propaganda Mill", *Medium* 4 February 2019, available: https://medium.com/dfrlab/influenceforsale-venezuelas-twitter-propaganda-mill-cd20ee4b33d8 [accessed 21 July 2021].

33 'AI's coming home: How Artificial Intelligence Can Help Tackle Racist Emoji in Football' Hannah Kirk Oxford Internet Institute Blog 16 July 2021 https://www.oii.ox.ac.uk/blog/ais-coming-home-how-artificial-intelligence-can-help-tackle-racist-emoji-in-football/ [Accessed 22 July 2021]

34 Center for Countering Digital Hate, Hidden Hate: How Instagram fails to act on 9 in 10 reports of misogyny in DMs, 6 April 2022 https://www.counterhate.com/_files/ugd/f4d9b9_6309420782df4942aad0ba240e190e4f.pdf

As part of their risk assessment, providers must assess how new content is produced on the platform (either created or modified). This includes but is not limited to:

- deepfake or audio-visual manipulation materials;
- nudification technology;
- use of bots (including chatbots and bot networks); and
- content embedded from other platforms and synthetic features such as gifs, emojis, hashtags that contribute to cross platform risks.

Age-appropriate barriers must be introduced to stop harmful contact with minors. Accessible and transparent user mechanisms must be in place for adult users to also implement such features that protect them from exposure to harm. This could include:

- features to prevent the direct messaging of accounts that do not follow a user;
- messages from unknown contacts reviewed by moderators; and
- control features around who can search for a profile, what content is visible for example features which filter harmful content and words appearing, and how personal content can be shared or re-distributed online.

Regulated services must consider how to stop harmful content that may originate on other websites and is moved to different platforms. For example, this could include consideration of how to prevent image-based sexual abuse, such as all forms of taking, making and sharing nude or sexual images without consent, including threats to share and altered images. Also how this abuse may be facilitated and hosted on platforms as well as the extent to which closed groups with large numbers of members are facilitating or enabling this type of abuse.[35]

A service may decide to introduce barriers to stop people sending unsolicited nude pictures without consent. This could include blurring the picture or stopping the message from being sent and warning the intended recipient.

These issues are a starting point, as are the points under sections 4 and 5. It is recommended that more work is undertaken to understand how features can cause problems with a view to potentially expanding this list.[36]

35 https://amp.theguardian.com/world/2022/jan/06/i-have-moments-of-shame-i-cant-control-the-lives-ruined-by-explicit-collector-culture

36 A model could perhaps be the survey work undertaken by the OECD on the approach to terrorist and violent extremist content: Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services OECD August 2020 No.296, available https://www.oecd-ilibrary.org/science-and-technology/current-approaches-to-terrorist-and-violent-extremist-content-among-the-global-top-50-online-content-sharing-services_68058b95-en

Section 4) Discovery and navigation

(1) Regulated services should review their recommender systems, especially their automated systems, so that they do not cause foreseeable harm, including VAWG, through –

  (a) promoting VAWG content;

  (b) suggesting groups or other users to follow that endorse or positively view VAWG or misogyny; and

  (c) rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item.

(2) Consideration must be given, in line with child-related duties, as to how to protect children to a greater degree.

(3) Platforms must consider how easily, quickly, and widely VAWG content may be disseminated by means of the service and respond appropriately.

(4) Regulated services should consider the impact of autoplay functions, especially in the context of content curated or recommended by the provider. Where the service provider seeks to take control of content input away from the person through autocomplete or autoplay (see below). The provider should consider how this might affect a person's right to receive or impart ideas.

(5) Regulated services should consider the need for explainability or interpretability, accountability and auditability in designing AI and machine learning systems, particularly with regard to the representation of women and girls, especially those from minority groups, in their data sets.

(6) A platform provider should consider the speed and ease of transmission, for example methods to reduce the velocity of forwarding and therefore cross-platform contamination.

(7) A platform provider should consider the way in which AI, machine learning systems and/or human moderators will distinguish between hateful and harmful content, reclaimed terms used by particular groups, and that of 'counter speech', minimising the risk of blocking or limiting legitimate use of terms within certain online communities and counter speech.

(8) A platform provider should be responsible for ensuring that algorithms not suggest material that is in contravention of the site's own Terms and Conditions.

Commentary

Design choices of online services, particularly recommender algorithms, determine the content which is being pushed onto users. Content that can cause serious harm, such as pornography, is often stumbled upon by children, rather than sought out. A contributor to this can be algorithm-based

recommendations.[37] There have been concerns that the effect of the recommender algorithms, especially in conjunction with auto play can prioritise extreme content, and therefore has a role in spreading online VAWG.[38] When considering the weighting of factors to promote content, care should be taken to ensure that there are no side-effects for example from heavily weighting user engagement (which says nothing about whether content is good or bad, just that it elicits a strong response). Regulated services must consider whether harm can be averted by designing their systems, including algorithmically driven newsfeeds, in ways which protect user groups from gender-specific harm. For instance, age assurance technology can be used to prevent children being exposed to harmful and age-inappropriate content online, and an investigation by BBC's Panorama found evidence of algorithm-driven misogyny.[39]

It is common in social networks to use software to select, rank and present or recommend items of content to users and to suggest text while typing.[40] Often this software contains machine learning or 'artificial intelligence'. Machine learning derives its capability from processing large data sets to inform its actions. In addition to problems around the representativeness of data sets, the people who write the machine learning software may be unaware of or unfamiliar with discrimination against women and girls, particularly those from minoritised groups, which compounds the risk of intersectional harm. For example, how AI has been programmed to be racist against dark-skinned Black women.[41]

Despite the necessary focus on Terms of Service and importance of their efficacy and enforceability, it is also must be recognised that, as stated by the UN rapporteur report:

> *"The setting of rules by social media platforms through community guidelines and moderation by algorithms is not objective. It reflects the biases and worldviews of the rule-setters, who tend to be typically from the specific sociocultural context of Silicon Valley: racially monochromatic and economically elite. The gender bias evident in content moderation reinforces the argument for companies to base their content moderation on international human rights standards."*[42]

Tools built on AI and machine learning may well run into problems common to such systems with regard to bias and lack of transparency and may contribute to the reinforcing of negative stereotypes. These tools may also have the side-effect of suppressing counter-speech.

Many such systems are often described as 'black box' in that their internal workings are not readily visible. The problems that arise from the use of AI and machine learning are not inevitable (or at least not all); the decision-making processes around their development and deployment must be scrutinised.[43] However, even 'black box' systems have outputs, which can be tested. At the statistical scale at which many social networks operate issues of bias should be discernible. Testing (see Guideline 7) should take into account how the tool is likely to be used.

---

37 In 2020 the BBFC found that 62% of 11–13-year-olds who reported having seen pornography described their viewing as mostly unintentional. https://www.revealingreality.co.uk/wp-content/uploads/2020/01/BBFC-Young-people-and-pornography-Final-report-2401.pdf

38 E. Hussein et al, 'Measuring misinformation in video search platforms: An audit study on YouTube' (2020) Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), Article 48. doi 10.1145/3392854; S. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).

39 https://www.bbc.co.uk/news/uk-58924168

40 T. Gillespie *Custodians of the Internet* (New Haven/London: Yale University Press, 2018), p. 7

41 https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

42 United Nations General Assembly, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, 30 July 2021, A/76/258

43 Committee of Experts on Internet Intermediaries, *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, (MSI-NET) Council of Europe study DGI(2017)12, p.8

Service providers should consider how to ensure that their recommender features are auditable[44] including considering and documenting the questions of what was considered when setting up the features and what the operation of the features show.[45] In this, providers should pay particular regard to special guidance on algorithmic accountability and auditing.

Autoplay and pop ups operate to push content at users without those users having chosen to engage with content, affecting a person's freedom to choose the content with which to engage. There has been concern that this, combined with the operation of the recommender machine selecting the content to be pushed, has resulted in the prioritisation of abusive and hateful content, including VAWG, (amongst other types of illegal and/or unwanted content). If autoplays are to be included, providers should consider whether other options have fewer adverse impacts; for example, autoplay only operating with user-selected playlists.

Autocompletes are a particular subset of the use of automated discovery tools and they operate to define a user's text entry or search term and thus the material that comes to that user's attention. Some autocomplete functions suggest misogynistic, racist or abusive searches[46], potentially contributing to the promotion of that message as well as being harmful to those in the relevant group. Providers should consider the adverse impacts of the use of such tools, as well as the nature and extent of any compensatory moderation or removal policies in this context. Reporting features for problematic autocompletes should be clearly visible and easy to use. Where problems arise, providers should verify that the issue is solved. See further Guideline 5 on complaints.

Some of these problems can be avoided if service providers are clear about their values as suggested in Guideline 1 and ensure that their recommendation and curation features embody those values.

Speed of transmission

Many providers aim to ensure communication is as frictionless as possible, which means that people can share content even without opening it and therefore not considering the content (and similar points may be made about 'like' buttons and similar features). These features support the virality of certain sorts of content. This is potentially problematic given the bias towards content expressing discriminatory or abusive content. Regulated services should therefore consider the constitutive role of these features in the spread of VAWG-related content.

Design choices and product functions of online services can facilitate the escalation and amplification of content that may be seen by millions of other users in a short space of time. This could be through user-to-user reshares or via algorithmic amplification. Platforms have a duty of care to stop content being amplified to users, where the escalation of such content shared online may cause victim re-traumatisation, or where the amplification of such content causes specific harm to vulnerable users, either because content was directed to them in a manner which may cause harm, or because their own content was amplified, which increased or generated harm to the user.

---

44 The issues of explainability have been discussed following the GDPR's inclusion of a right to an explanation; See eg Margot E Kaminski 'The Right to Explanation, Explained' (2019) 34 *Berkeley Technology Law Journal* 189, DOI: https://doi.org/10.15779/Z38TD9N83H. Some consider interpretability a better approach: Cynthia Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (2019) *Nature Machine Intelligence* 206-15.

45 See eg Joschua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson and Harlan Yu, 'Accountable Algorithms' (2017) 165 *University of Pennsylvania Law Review* 633, available: https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=9570&context=penn_law_review [accessed 26 July 2021].

46 'Hidden hate: What Google searches tell us about antisemitism today' Stephens-Davidowitz, Seth Published by Antisemitism Policy Trust and Community Security Trust 2019 available at https://archive.jpr.org.uk/object-uk508 [Accessed 26 July 2021]

Regulated services must put systems and processes in place to prioritise action to combat VAWG that has the potential to spread, for instance, intimate self-generated images of a child which have been shared online**.** They should also monitor for harmful content that is amplified. For instance, they must consider reviewing:

- content that becomes exponentially prevalent;
- content that is being reshared by multiple users; and
- content that is being reshared on mass by a particular user segment as identified by know your user.

Section 5) User Response, User Tools

Settings and Tools

(1) Regulated services must empower users by providing tools which, in addition to content and behaviour reporting tools, allow users to improve control of their online interactions and to improve their safety. These could include –

(a) controls over recommendation tools, so a user can choose for example to reject personalisation. Examples include –

– user-set filters (over words or topics)

– tools to limit who can get in touch/follow a user, or to see a user's posts.

– tools to allow users to block or mute users, or categories of user (for example anonymous accounts);

(b) tools for adapting privacy settings and setting privacy options as default for young and vulnerable users;

(c) controls for the user over who can and cannot redistribute their content or username/ identity in real time;

(d) the ease of use of these tools and their prominence such that users are aware they exist;

– including ease of use for children and those with accessibility issues

(e) specific tools in place for users under 18. This could include –

– Tools to stop children from receiving unsolicited messages from adults

– Measures which are targeted at the adults doing this

– Notifications to make an adult messaging a child aware of the policies of the service in relation to communication with children

– Notifications to ask a child if they know who is messaging them and to explain what children can do if they are confused or made to feel uncomfortable by it

Reporting Mechanisms

(2) Users must be able to effectively report content that is illegal or harmful to regulated services through clear and transparent flagging mechanisms. Regulated services are obligated to have effective and easy to use reporting functions and must use them to triage content for both human and automated moderation.

(3) Service providers should have reporting processes that are fit for purpose for reporting VAWG content and wider harms, that are clear, visible and accessible and age-appropriate in design. Thought should be given to reporting avenues for non-users such as teachers or family friends and support services, who are able to report without the victim needing to engage further with the harm.

(4) Service providers should have in place clear, transparent, fair, consistent and effective processes to review and respond to content reported as VAWG content. Users must be given the ability to submit third-party content to the companies' intelligence systems in relation to specific cases of content violation.

(5) Reporting processes should set out clear time frames and should inform the user directly of any decision made. Reporting processes should include a specific point of contact that is provided to users so users are able to follow up on decisions made.

Commentary

Online violence against women and girls restricts women and girls' access to online spaces, whereby consistent failures by companies to act on and respond to reports of harm and abusive content online has left users, who have found that tools and mechanisms meant to address harm are largely inadequate, disempowered.

User tools and controls designed to address risks faced by women and girls online, as well as specialist moderator training and victim support for gendered experiences of online harm (see below), are necessary for a holistic approach to online harms. This will allow users to have agency over their experience of online platforms and services.

Empowering users to engage with online services must form part of a cyclical process, whereby reviewing harms reported by users feeds into ongoing work to review emerging harm patterns and shape moderation, risk assessment and mitigation processes.

Transparency around the journey of a report and the actions taken by service providers will empower women and girls to report harms. This empowerment must not be seen as a substitute for adequate action and effective harm prevention measures. It is not the responsibility of users to avoid harm and service providers must not see this as such.

Women and girls need access to reporting systems that are designed for ease of use and are tailored to reflect the needs of gender based online harm. Options for user response and reporting must reflect the current landscape as well as the future development of harms against women and girls, providing remediation and reporting functions that can easily capture user experience and context. Additionally, consideration must be given to how the harms that impact women and girls are intersectional, and where there is overlap with other protected characteristics.

User tools

As part of their risk management, social media companies should provide tools for users that can be used if systemic risk mitigation fails.

Controls give users, or parents/carers of users, control over the type of content which they are exposed to. This helps users protect themselves and can provide users with agency over decisions about the shape of their online experience. Regulated services must empower users by having a suite of controls which can easily be accessed, with clear and transparent information about the impact of different design features and why they may or may not choose to implement them. However, this must not be a substitute for effective harm mitigation.

The platform makes design choices about whether to provide these tools and how easy they are to find and use (including providing instructions and examples in multiple languages). Given the tendency of users not to change the original settings, providers should have maximum safety settings within the platform as default (even if users can then change these settings.

Muting and blocking tools might give rise to concerns about the rights of the speaker and 'filter bubbles' The right to freedom of expression limits the ability of states to intervene in communication between willing speaker and willing listener but does not give a speaker the right to force someone to listen to that speaker. Nonetheless, the right to receive presumably also implies the right not to receive, though like the expressive right, it is not unlimited.

Complaints processes

Complaints processes provide vital early warning of VAWG problems on a service, as well as a mechanism to deal with a problem in an individual case.

The adequacy of complaints processes should be part of the risk assessment. The provider should also ensure that the design of complaints mechanisms is user-centric: that is, visible, easy to use and age and language appropriate. Complaints processes should not just be limited to complaints about individual items of content. They should allow for complaints about a series or pattern of communications as well as to features of the services itself (for example, the way the recommender algorithm works, or other 'dark patterns'[47] and nudges, or tools for creation). The regulator must regularly assess whether such processes are fit for purpose. Regulated services must work to identify trends and developments in user reporting and incorporate this in any transparency reporting obligations to the regulator.

47 A term to describe an interface to trick a user into making purchases online, signing into fake accounts etc

Good practice in responding to VAWG content that is flagged to an online service might include the following:

- all platforms must acknowledge reports within 24 hours. Reports must be actioned within a specific time frame set and published by the provider in their Terms of Service and in response to a report made (this may vary dependent on harm reported);
- data should be gathered on response times to ensure these commitments are met;
- companies should track where multiple reports are made by an individual as this may indicate increased risk of harm;
- victims must be able to provide the username of the perpetrator, rather than reporting individual pieces of content;
- reporting avenues should be provided for non-users to flag harmful content;
- users should have access to clear flagging processes that identify whether their issues are VAWG related as well intersecting with other types of abuse such as racist, homophobic abuse. This is in addition to more specific flagging categories to triage and escalate risk;
- consideration must be given to the accessibility of flagging and reporting for younger users who may not be conscious of VAWG dynamics impacting their case;
- regulated services must use the intelligence from the report or flag to prioritise its human and automated content moderation;
- in the case where content, which has had a determination by automated technology, is continuing to be flagged or reported, it must be assessed by a human moderator;
- there must be an appropriate number of VAWG-trained human moderators, taking into account the scale of any VAWG problem on the service;
- human moderators must be supported in a holistic manner which recognises the psychological impact of the work;
- harmful content or actions which have been flagged as having gendered nature must be expedited and considered by moderators with VAWG and child protection expertise;
- regulated services must explain the outcome of a report or flag in clear and simple language, outline a user's right to appeal and explain the steps a user must take if they do not agree with the determination; and
- recommender algorithms must consider content that has been recently flagged or reported and limit its spread until the content has been reviewed.

This Guideline should be considered in line with Guideline 6 on Moderation.

Section 6) Moderation

(1) Regulated services must have in place sufficient numbers of moderators, proportionate to the online service size and growth, and to the risk of harm, who are able to review VAWG content. This may include moderators who work exclusively on VAWG issues.

(2) Regulated services must put in place appropriate, updated education and training on VAWG for all staff and subcontractors involved in the content production and distribution chain. This includes senior executives, designers, developers, engineers, customer support and moderators, designed in consultation with independent VAWG experts. The moderators must be appropriately trained, supported and safeguarded.

(3) Regulated services must consider assigning moderators to specific types of VAWG content to ensure the correct moderators, trained in their specialist subjects and on related language and cultural context considerations are able to review the content in a consistent fashion.

(4) Regulated services must have in place processes to ensure that where machine learning and artificial intelligence tools are used, they operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable. For instance, technology to remove sexualised pictures must not remove photos of breast feeding. A platform provider should consider the way in which AI and machine learning systems and/or human moderators will distinguish between hateful and harmful content, reclaimed terms used by particular groups, and that of 'counter speech', minimising the risk of blocking or limiting legitimate use of terms within certain online communities and counter speech.

(5) Users must be informed of the use of such automated tools. Machine learning and artificial intelligence tools cannot wholly replace human review and oversight.

(6) If the VAWG content involves a person protected by UK law, regulated services must review the content taking into account the terms of service and UK law.

(7) Regulated services must have clear timeframes for action against flagged content, in line with the good practice outlined in the previous section. Awareness begins at the time flagged content, by means of email, in-platform notification, or any other method of communication, is received.

(8) Regulated services must act, proportionate to risk, on content which is not deemed to be illegal but is considered to break their Terms of Service, Community Guidelines, or is considered a new form of VAWG, as soon as it is identified. Acceptable actions on a piece of content which violates a provider's Terms of Service can include –

(a) removal of content;

(b) labeling as inaccurate/misleading/contrary to the rules;

(c) demonetise content;

(d) suppress content in recommender tools;

(e) termination of account;

(f) suspension of account;

(g) geo-blocking of content;

(h) geo-blocking of account;

(i) issuing a strike, if a strike system is in place;

(j) instituting delay in posting content or otherwise adding friction to the communication process;

(k) limiting number of posts over a given time period; and

(l) adding friction to mechanisms by which content may be shared.

(9) Regulated services must have systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users must be kept up to date with the progress of their reports and receive clear explanations of decisions taken.

(10) Provide holistic support for moderators who are exposed to harmful content in recognition of psychological impacts of what they are exposed to (examples may include mental health support or clinical supervision).

(11) Online services must consider putting in place an appropriate trusted flagger programme that maintains independence from the online service and from governments. The programme must include UK based non–government organisations and other experts, including the specialist VAWG sector, who will be vetted, to inform on policy development and report on new trends in harmful and illegal content. It is recommended service providers have a Trusted Flagger Policy that includes –

(a) trusted flaggers are not used as a sole provider of flagging content;

(b) trusted flaggers are appropriately compensated and incentivised for work provided to companies to ensure their compliance while not compromising their independence and impartiality;

(c) regular meetings held (with members of the trusted flagger programmes) to review content decisions and discuss any concerns;

(d) provision of support for trusted flaggers who are exposed to harmful content, as per the support provided to the companies' own moderators, whether directly employed or working for out-sourced companies;

(e) a specific Trusted Flagger reporting email address;

(f) a specific trusted flagger escalation route if no / unsatisfactory response received;

(g) clear criteria for what can be reported and what cannot;

(h) clear limited and reasonable expectation for additional information on escalation;

(i) commitment to an expectation on response times of 24 hours. Responses should include details of action taken or reasons for rejections and should include links to policies or Community Standards as relevant;

(j) willingness to reopen a case and review if additional information comes to light; and

(k) adoption of automatic suspension of content reported via Trusted Flagger route pending review.

(12) Where online services use civil society organisations for significant undertakings, they must consider remunerating them for their time and expertise.

Dispute resolution

(13) Regulated services have an obligation to instigate dispute resolution functions which allows users to raise a complaint against decisions made by the platform.

(14) Regulated services are obligated to put in place a right of appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content. All users must be given a right to appeal any measures taken against them. Users must be able to present information to advocate their position.

(15) Regulated services must acknowledge an appeal request within 24 hours of receipt. If more time is needed to assess the content the user must be informed.

(16) Regulated services must have appeals systems which must take no longer than seven days to assess appeals, except in exceptional circumstances. Exceptional circumstances could include a major disaster, or an event or incident of the same magnitude.

(17) Regulated services must explain the outcome of a dispute in clear and simple language.

(18) Complaints related to VAWG must be reviewed by a professional trained in VAWG issues for example by a VAWG specialist service.

(19) Dispute resolution procedures must be fair, transparent, and easy to use. They must not discriminate between users, introduce bias, or be applied inconsistently
Regulated services must remain conscious that children may not be able to access dispute resolution procedures and offer alternative mechanisms for children to raise issues.

Commentary

It is the service provider's duty to moderate content consistently in line with its terms of services. Moderators must be appropriately trained in understanding and tackling VAWG and providing support and signposting to victims. It may be appropriate for specialist VAWG teams of moderators to be established for larger companies or companies with gender specific problems. Staff tackling VAWG must be appropriately safeguarded and supported.

In many social networks, moderation takes place against the community standards/terms of services. Many VAWG incidents on social networks reveal deliberately or accidentally deficient community standards[48]. In risk management, service providers should look hard at the adequacy of their terms to prevent VAWG and update them in consultation with the specialist VAWG sector and other organisations that support victims of gender-based violence experienced online.

48 Ref to emoji issues w Instagram and footballers

Risk assessment of moderation

As part of their risk assessment (Guideline 1), companies should assess what form of moderation is appropriate, whether it is in-house, whether it relies on external volunteers (including trusted flaggers), or whether some form of automation should be used.[49]

Use of AI must be carefully assessed, bearing in mind established challenges regarding AI systems generally relating to accuracy and racial and gendered bias, but also those specific to content moderation. Particular attention should be paid to the question of whether such systems can adequately incorporate an intersectional and contextual analysis of content. As with the role of AI in content amplification, platforms should consider the possibility of allowing users not to be subject to AI moderation. In any event, they should inform users if and how AI is used in easy-to-understand terms, with in-built human oversight and review.

Impact of moderation

The service provider should ensure that the moderation response adopted is proportionate to the harm/intensity of the VAWG-related content and that they provide a clear, reasoned decision that explains this linkage. The decision should bear in mind the importance of freedom of expression to democracy, but also the potentially silencing effect of VAWG and the impact that has on democracy. Refuge cites that 38% of survivors said they felt unsafe or less confident online as a result of the abuse from a partner/former partner on social media. In their workshops Glitch have found that despite the vast majority of participants (96%) of Glitch's workshops stating that post-workshop, they feel that they now have the skills to be safer and more resilient online, 69% of participants have told Glitch that they will continue to censor themselves online due to anxiety or fear of how others will respond.

Take down considerations

Regulated services' approach to take down should also be part of the risk assessment. Search providers should consider whether take down (or account suspension/removal) is a proportionate response: this will depend on a range of factors including the severity of the harm caused and the track record of the user posting content. Where content has been posted by bots, this is a factor that should be taken into account.

In terms of take down times, a provider should consider what is appropriate, and even what is the right measure of responsive take down times. For example, whether it depends on the time elapsed, or the number of impressions, or a combination. This might depend on the nature of the content and the types of person harmed.

Note also that it is possible that providers may choose to remove content without waiting for a complaint within the Terms of Service.

There is consideration to be made regarding take downs and the potential for restrictions on free speech to fall disproportionately on marginalised communities whose expression may already be targeted and policed online, and for whom being able to engage in and participate in online spaces is particularly important.

---

49 Robyn Caplan has categorised types of moderation: R Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches", 14 November 2018, Data and Society Report, available: https://datasociety.net/library/content-or-context-moderation/.

Action on serial offenders

For egregious breaks of the Terms of Service, regulated services must consider investigating their other activities to determine whether that person is a serial offender. This could include assessing whether they are part of problematic groups which may exacerbate VAWG (ie "collector" groups or coordinated efforts to abuse women online). Some issues may need to be prioritised and it will be appropriate for the online service to consider harmful VAWG issues which may arise, for instance, image-based sexual abuse of a girl.

Developing Trends

Regulated services must have processes in place for moderators to raise new trends in VAWG as an issue. The service must consider whether new processes or tools are required to tackle the new type of VAWG. This information should feed back into risk assessments and harm mitigation plans.

The Online Safety Bill introduces an obligation on services to report emerging harms to the regulator. There must also be lines of reporting for specialist sector experts to be able to raise concerns around new and emerging types of VAWG.

Section 7) Transparency

(1) Online services must publish transparency reports in line with Ofcom's guidelines. These must be easy to access and understand. Online services must be prepared to answer questions on the findings.

(2) On request, online service must provide individuals with easy to digest data that the online services hold on them.

(3) Online services must uphold individuals' right to be forgotten and rights under GDPR.

(4) Online services must respond to requests for information by any Government or Ofcom appointed user advocate in the required time.

(5) Online services must proactively share information with third sector organisations where it is relevant for the organisation to safeguard the citizens that they represent on a regular basis such as quarterly meetings. For instance, this could include sharing intelligence on –

- (a) Themes and categories relating to VAWG moderation and user reporting.
- (b) Scale and dimensions of online risk experienced by women and girls.
- (c) Data relating to emerging risks and new trends in online harm perpetration.
- (d) Effectiveness of risk mitigation tools and protective measures in place relating to VAWG online.

(6) Online services must maintain effective channels of collaboration and communication with civil society organisations with expertise in these areas.

(7) Online services must consider whether decisions on gender-based harms would benefit from consultation with civil society including VAWG specialist services. For instance, when risk assessing new technology.

Commentary

Transparency reporting and information release must contain three main elements:

- collaboration and information sharing with relevant regulators;
- collaboration and information sharing with relevant civil society bodies that support the prevention and mitigation of VAWG; and
- public data sharing in line with transparency guidelines that is accessible and easily digestible for all service users.

Clear transparency allows civil society and the public to monitor online services progress in tackling gender-based harms and hold online services to account.

There is a public benefit to transparency concerning online safety. Transparency enables society to monitor the progress of the sector. It also builds confidence in the industry.

Online services are strongly encouraged to collaborate with experts on VAWG topics and achieve better outcomes for their users.

Online services that effectively collaborate with other platforms will be able to consider gender-based harms in the round and tackle issues before they appear on a platform.

It is recommended that a UKCIS working group on VAWG is established which could bring regulated services, the regulator, VAWG sector and government together and be used as a means of sharing reports and data.

Section 8) Victim support and remediation

(1) Regulated services must take steps to ensure that users who have been victims of VAWG, or are exposed to harmful content, are directed to, and are able to access, adequate support. Support can include –

(a) information from the online service about actions that can be taken to report, protect from and prevent harm, as well as guidance on the regulated services broader actions to address VAWG, such as outlined in this code of practice;

(b) signposting and access to websites or helplines dealing with the type of online harm experienced by the user or witnessed by others who may be affected by the content, even if not the designated target; For instance, the National Domestic Abuse Helpline, in England run by Refuge, Tech Safety website,[51] Childline, The Revenge Porn Helpline[52] and Report Harmful Content service[53] and StopNCII.[54] Best practice would follow a "polluter pays" model where financial contributions are made to specialist VAWG organisations providing support;

(c) reports from children or related to child safety be expedited; and

(d) information from, and contact details for, services providing victim support or mental health support after being exposed to hateful, violent, and harmful materials.

(2) Regulated services should conduct reviews of the signposting and support material they are providing, and maintain their own lines of communication with support services, to share information and ensure there is capacity for referrals.

3) Providers must have transparent processes that highlight the journey of a user report, and opportunities for regular updates and communication about action taken. If action is not taken to remove content, there must be a clear explanation as to why, together with signposting to relevant services who can help and support and potentially appeal a decision (such as the Revenge Porn Helpline or Report Harmful Content Service).

Commentary

Regulated services have a duty of care to protect their users from harm. Where they have failed in that duty, they must provide support to the victims that is trauma-informed, and victim centred.

Platforms should help protect victims by proactively suggesting and advertising tools, such as NSPCC's Report Remove and SWGfL's 'stop NCII (non-consensual intimate imagery)' helpline.[54]

50 https://www.nationaldahelpline.org.uk/ and https://refugetechsafety.org/
51 https://revengepornhelpline.org.uk/
52 https://reportharmfulcontent.com/
53 https://stopncii.org/
54 https://stopncii.org/

Remediation and support provided to users must be victim centred and all appropriate efforts must be made to avoid re-traumatisation or prolonged harm during the reporting process. Regulated services should consult victims and victim-representative groups in a respectful and sensitive manner to design remedies for people who have been harmed by VAWG. Victim support can be of help in mitigating the harm suffered by victims of VAWG, as part of the "remedy ecosystem," and go some way towards providing rehabilitation[55] [56]. The provider can facilitate users finding this support, as not all such organisations are known about or visible. Victim support is not a substitute for nor an alternate to stopping VAWG at source (as set out in Guideline 1).

The 'polluter pays' principle, endorsed by the OECD for almost 50 years suggests that the companies enabling these harms to society should pay to help rectify the damage. By ring-fencing at least 10% of the Digital Services Tax annually for ending online abuse against women and girls, this would help fund civil society organisations to carry out their vital work to support those affected by VAWG.

Section 9) Safety Testing

(1) As part of their risk assessment, mitigation processes and safety by design, regulated services should carry out or arrange for the operation of such testing and examination of their systems as may be necessary to carry out due diligence in reducing or removing content that facilitates VAWG. This approach should account for respect for the human dignity of people involved or affected by those tests, as well as ethical considerations relating to experiments involving human participants.

(2) Testing should specifically include (but is not limited to) recommendation and curation functions and automated curation and moderation systems.

Commentary

Safety testing should be at the heart of due diligence and risk assessment.

Testing is particularly relevant in an approach that focuses on very complex software systems. For over 150 years, scientific testing[57] of company processes has been intrinsic to protecting people from harm. For workers, customers and people who might be harmed by, but are not involved in the company or its products. External testing standards work best when they are transparent and, for the most hazardous services, are carried out by independent people. In some industries there are multinational agreements on

55 BTech, Access to remedy and the technology sector: basic concepts and principles, available: https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf

56 Access to remedy and the technology sector: 'a remedy ecosystem approach' OHCHR 2020 https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf

57 Peter W.I. Bartrip, "The State and the Steam-Boiler in Nineteenth Century Britain", International Review of Social History, Volume 25, Issue 1, April 1980, pp. 77-105

testing procedures and standards to protect the public. Regulated services carry out extensive testing of product features to meet their commercial goals[58] but testing for safety seems less exhaustive.

Providers should carry out testing and examination of their new services, and new features on that service or any new service of feature is made available (see Guideline 2 on Safety by Design), that enables them at a minimum to –

- understand whether the measures they have put in place are working to prevent, or appropriately mitigate, VAWG; and
- detect whether new forms VAWG have appeared.

The service provider should also test for/measure whether the measures in place to protect against VAWG, taking how users experience multiple and overlapping forms of discrimination into account, have unduly restricted other rights.

Confidence in the service provider would be enhanced if it published the results of such testing in a timely manner as well as allowing external review.

Testing should not be carried out solely against a standard, but also involve exploratory, qualitative investigation to assess exactly how a new feature could be used to facilitate VAWG at each stage of the four stage model set out in Guideline One. Such testing will work best if it involves people who have lived experience of VAWG, using an ethical and informed framework.

It is appropriate for VAWG experts to be involved in the testing of new systems and for users to be aware that they are being tested on.

58 'Facebook engineers and data scientists posted the results of a series of experiments called "P(Bad for the World)." The company had surveyed users about whether certain posts they had seen were "good for the world" or "bad for the world." They found that high-reach posts — posts seen by many users — were more likely to be considered "bad for the world," a finding that some employees said alarmed them.' Kevin Roose, Mike Isaac and Sheera Frenkel, 'Facebook Struggles to Balance Civility and Growth', New York Times, 24 November 2020. Available here: https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

Section 10) Supply Chain Issues

(1) Regulated services which outsource any part of their business, including moderation of content, applications, GIFs, images, or any other content or tools, including safety tech, should ensure the vendor adheres to the social media provider's Terms of Service and Community Standards, and where necessary take action to enforce those standards, and that they have employee and mental health protection policies in place that adhere at least to the same standard.

(2) Processes should be in place for users to report content or tools provided by a vendor which is illegal or violates the service provider's Terms of Service or Community Standards and Guidelines.

(3) Regulated services should ensure adequate information is available to the vendors on their Terms of Service and Community Guidelines to pre-empt any violations.

Commentary

This guideline is important in ensuring that regulated services do not seek to avoid responsibility for VAWG through outsourcing or ignoring the human rights and harms risks arising from it. Risk assessments therefore should include an assessment arising from business relationships. It is not enough for service providers to just draw supplier's attention to the Terms of Service and Community Standards, but should also include provision in any contractual documents, and ensure that those provisions are enforced where necessary.

It is increasingly common for regulated services to contract out parts of their business function and the impact could be felt by people who are customers or people who work for the supplier such as moderators with poor working conditions. Concerns are that providers discharging duties in removing what may 'amount to' priority offences in the Online Safety Bill, as is the current wording, this will likely require a combination of more sophisticated AI (with all existing issues of bias and unfair outcomes) and human content moderators including through outsourcing to avoid heavy penalties where there will be expansion of hidden and exploitative working conditions. [59]

This trend could continue as application programming interfaces allow componentization of a software service, sometimes in response to regulatory pressure.[60] Large social networks will be able to apply leverage to ensure that subcontractors or suppliers follow best practice, as well as international human rights norms. Smaller social networks might have to consider whether contracting out some components is worth the human rights risk for their customers.

---

59 https://www.foxglove.org.uk/2022/02/16/foxglove-supports-facebook-content-moderator-sacked-kenya/

60 "Evidence suggests that large data holdings are at the heart of the potential for some platform markets to be dominated by single players and for that dominance to be entrenched in a way that lessens the potential for competition for the market. In these circumstances, if other solutions would not work, data openness, could be the necessary tool to create the potential for new companies to enter the market and challenge an otherwise entrenched business.' HM Treasury UK Government, 'Unlocking digital competition: Report of the Digital Competition Expert Panel' 13 March 2019, 2.89 Page 75, available here: https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panel

Note also that some features may not be derived from a formal business relationship but be through third party independent software such as services that allow a user to post to multiple social networks. Social media providers should also consider the risks of harms arising from VAWG arising from such software.

It is also recommended that Ofcom consider taking action against companies who have outsourced functions found to be in contravention of the Bill and the Guidelines, similar to enforcement taken by ICO.[61]

Section 11) Enforcement of criminal law

(1) Service providers must have in place a point of contact for law enforcement authorities in the UK. The contact is responsible for giving information about potentially criminal content to law enforcement authorities under para 2. This includes –

(a) information about the content;

(b) the details of the user, including location;

(c) details of enforcement action on the content undertaken by the provider; and

(d) other materials relevant to criminal investigations.

(2) Information requested by government and law enforcement authority in accordance with UK law should be delivered within the time frame specified by national rules or no later than one month of receiving the request. In exceptional circumstances this can be extended, with written approval from the relevant authorities placing the request, with a full expected time frame set out.

(3) Effective protections should be put in place by service providers to ensure flagging and court orders are not used for malign purposes by Government agencies or law enforcement of any kind to remove content they find objectionable, which is neither illegal nor harmful.

Commentary

Groups representing victims of VAWG have raised concerns that even in established markets such as the UK, social platforms were not complying swiftly with legitimate requests from law enforcement authorities.[62] Providers make decisions about the quantity and quality of resources they employ in services such as in investigation and remediation of potential human rights adverse, or even illegal, impacts. Providers should therefore ensure the adequate frameworks are in place to process law enforcement requests expeditiously, and resource them adequately.

61 Facebook fined £500,000 for Cambridge Analytica Scandal, October 2018 https://www.bbc.co.uk/news/technology-45976300

62 https://www.theguardian.com/uk-news/2022/mar/16/molly-russell-inquest-family-frustrated-by-wait-for-instagram-data

This does not mean automatically handing over information without consideration of the legitimacy of the request, or considering the privacy of the users involved, but that the question is given appropriate and timely attention by appropriately qualified staff, bearing in mind the general principle that applicable laws (which themselves respect international human rights laws) should be respected.

Section 12) Education and Training

(1) Regulated services should consider implementing appropriate education and training on VAWG for all staff and subcontractors involved in the content production and distribution chain. This includes senior executives, designers, developers, engineers, customer support and moderators, designed in consultation with VAWG experts.

(2) Materials used for such training must be made available to any Regulator, law enforcement authorities and Government agencies upon lawful request.

(3) The training of moderators and employees, including executives, must be driven by expertise from the specialist VAWG sector. Training on cultural competency and intersectionality to ensure a holistic understanding of VAWG and the way it impacts different minoritised groups must be included. Any expertise must be adequately remunerated.

(4) Within the service itself providers should ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms.

Commentary

Groups representing victims of VAWG have stated that a simple lack of staff training at tech companies can be a factor in further embedding harm. For example, a lack of understanding and training around domestic abuse and coercive behaviour, where many companies seem unable to account for the context and subjectivity of domestic abuse such as perpetrators posting images of survivors road signs etc.[63] The Interpretative Guide to the UNGPs stress repeatedly[64] the need for adequate staff skills in corporations to fulfil their duty to respect human rights.

63 Unsocial Spaces Refuge Report October 2021 https://www.refuge.org.uk/wp-content/uploads/2021/10/Unsocial-Spaces-for-web.pdf

64 See Q31 in response to UNGP 17: "It is important for all enterprises to ensure that the personnel responsible for human rights due diligence have the necessary skills and training opportunities.' Or in response to UNGP 19: "Can we build scenarios or decision trees for action across the company so that we are prepared to respond to the most likely or severe potential impact? Do staff need training and guidance on these issues?" Or in response to dealing with conflicting requirements Q83 "the more an enterprise has embedded respect for human rights into its values and the more it has prepared its personnel for ethical dilemmas, through training, scenarios, lessons learned, decision trees and similar processes, the more likely it will be able to identify appropriate and timely responses" The Corporate Responsibility to Respect Human Rights – an Interpretive Guide UNHCR HR/PUB/12/02; ibid.

Regulated services that have chosen to offer their service globally can neglect to ensure that moderators have been trained in the multitude of local issues of persecution of minoritised women and girls that might arise in markets far away from the corporate headquarters. Providers should also ensure that training and resources are kept up-to-date. An example of this could relate to how image-based sexual abuse can be understood as an attempt to control, subjugate and threaten victims by using, for example, (fear of) the shame associated with breaking perceived religious, cultural and faith boundaries, or by using faith and religion as a justification to pose for, send and share such images. Within a pattern of abuse offline, such intimate image abuse can be viewed as a tool of spiritual abuse.[65]

Section 13) Vigilance over Time

(1) Regulated services must have plans for ongoing review of their efforts in tackling VAWG and supporting review of risk assessments and mitigation plans. This might include engagement with relevant experts or organisations to advance policy development. The providers shall adapt internal processes accordingly, to drive continuous improvement and in particular shall regularly review and update when appropriate technical and organisational measures implemented under this code.

Commentary

The responsibility of companies is a continuous one. Where companies choose to offer services that develop rapidly, they put upon themselves an obligation of equally adaptable risk assessment processes. Cultural attitudes, social norms and behaviours are also in constant flux, reiterating the need for companies to consider their responsibilities within a universal human rights framework.

Risk management and mitigation should proceed in lock step with software and societal changes. This does not just include VAWG risk-assessments of new features, but involves continuing to risk assess for VAWG in the use or abuse of speech the use of older features.

65 Dr Lisa Oakley, 2018

This document has been produced to be put before the
Online Safety Bill Committee, May 2022

For any further information or enquiries please contact admin@evaw.org.uk

May 2022