Protecting freedom of speech by tackling online gender-based violence

"Ensuring an internet free from gender-based violence enhances freedom of expression as it allows women to fully participate in all areas of life and is integral to women's empowerment".

UN Special Rapporteur on Violence against Women, Dubravka Simonoci[1]

[1] In "UN experts urge States and companies to address online gender-based abuse but warn against censorship", Geneva 8 March 2017, available at: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21317&LangID=E

Executive summary

This paper analyses to what extent very large online platforms should be incentivised to take more meaningful action to address gender-based violence and harassment online as legislative proposals on their obligations are being debated at the European Union level.

Women are subject to massive and persistent abuse online, with one in 10 women in the EU reporting that they have experienced some form of 'cyber-harassment' since the age of 15. Although the phenomenon of online gender-based violence is not new, the reluctance of social media platforms to take decisive actions to reduce it is increasingly having real and significant impacts not only on the victims themselves but also on democracy, freedom of expression, gender equality and society. Women who are subjected to gender-based violence and harassment online are withdrawing from public discourse, stepping down from taking an active part in society and democracy (including those in public office) and withdrawing from public spaces in a way which raises significant concerns for the protection of fundamental rights, in particular freedom of expression and gender equality.

But this is entirely preventable. Today, online platforms' commercial incentives to amplify and spread 'engaging' content that captures attention (and therefore advertising revenue) coupled with a deficient EU regulatory framework that fails to address the specific threats posed by online gender-based violence have converged to create a hostile online environment for women who want to play an active part in public debate.

This paper explores ways to change this situation by encouraging legislative measures which would incentivise social media platforms to take a more proactive approach to addressing and preventing gender-based violence on their platforms.

The proposed approach looks at possible ways to force platforms to 'price-in' societal risk to their business operations so that the decision as to what to build and how to build it is no longer based solely on profit maximisation driven by growing engagement and monopolising attention to generate advertising revenue, whatever the cost to society. This would create a legal obligation for social media platforms to mitigate the risk of gender-based violence taking place on their products and being amplified by them.[2]

[2] Prepared by AWO for MEP Karen Melchior, 15/05/2021.

renew europe.

Table of Contents

renew europe.

1. Introduction: consistent evidence demonstrates that women are disproportionately subject to a wide variety of online harassment techniques

During the pandemic most of our social and professional interactions moved online, and social media platforms became an even more critical space to exercise freedom of expression. While these platforms have the potential to contribute to the democratisation of the public sphere and have offered new possibilities to exercise freedom of expression, they are also used by people to threaten, harass and abuse others online.

There is a growing body of research suggesting that women in particular are a target of online attacks.[3] Studies show that in 10 women in the European Union report have experienced 'cyber-harassment' since the age of 15[4]. Participants of a seminar on combating sexist hate speech organised by the Council of Europe summarised the issue as follows: "any woman, publicly known or not, is a potential target of sexist hate speech"[5]. This number significantly increases for women who are active in the public sphere: rampant online gendered harassment against female journalists, politicians and academics has become more visible and more coordinated in recent years and has occurred on the basis of their reporting, research, political work or purely on the basis of being women operating in the public sphere.

Numerous reports have highlighted how female journalists are impacted by sexist comments that criticize, attack, marginalize, stereotype, or threaten them based on their gender or sexuality. [6] This directly influences how female journalists do their job; causing them to decrease their visibility online, censor themselves, or to stop covering stories[7]. As such there is a chilling effect even to non-victims of such violence.

According to a 2016 Inter-Parliamentary Union study, 82 percent of female parliamentarians in 39 countries across five global regions have experienced some form of psychological violence (remarks, gestures and images of a sexist or humiliating sexual nature made against them or threats and/or mobbing) while serving their terms. They cited social media as the main channel through which such psychological violence is perpetrated; nearly half of those surveyed (44 percent) reported having received death, rape, assault or abduction threats towards them or their families[8]. 61.5 percent of those who had been subjected to sexist behaviour and/or violence believed those acts had been intended primarily to dissuade them and their female colleagues from continuing in politics[9]. In fact, a 2019 study found that 28

---

[3] Ging D and Norman JO (2016) Cyberbullying, conflict management or just messing? Teenage girls' understanding and experiences of gender, friendship, and conflict on Facebook in an Irish second-level school. Feminist Media Studies 16(5): 805–821.
[4] https://fra.europa.eu/sites/default/files/fra-2014-vaw-survey-at-a-glance-oct14_en.pdf at 29
[5] Report, Seminar Combating sexist hate speech, 10-12 February 2016, European Youth Centre, available at: https://rm.coe.int/16806cac1f
[6] See for example: "New Challenges to Freedom of Expression: Countering Online Abuse of Female Journalists", Representative on Freedom of the Media Organization for Security and Co-operation in Europe (OSCE), 2016, available at: https://www.osce.org/files/f/documents/c/3/220411.pdf ; "Attacks and Harassment The Impact on Female Journalists and Their Reporting", International Women's Media Foundation and Trollbusters, available at: https://www.iwmf.org/wp-content/uploads/2018/09/Attacks-and-Harassment.pdf
[7] Chen et al; 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. In Journalism, 2018, pp. 1-19.
[8] http://archive.ipu.org/pdf/publications/issuesbrief-e.pdf at 3
[9] Idem at 6.

percent of municipal officials in Finland targeted with online hate speech said they were less willing to participate in decision-making as a result[10].

Research undertaken by the Institute of Strategic Dialogue to analyse the scale and nature of online abuse targeting Congressional candidates during the 2020 US presidential campaign found that female candidates were up to three times more likely than men to be abused on Twitter.[11] A recent study by the Wilson Center sheds light on the volume of abusive content some of these candidates face. The researchers found a total of 336,000 pieces of abusive content shared by over 190,000 users over a two-month period in an analysis of online conversations about 13 female politicians across six social media platforms. The report focused on an emerging category of gendered and sexualized disinformation that is "a subset of online gendered abuse that uses false or misleading gender and sex-based narratives against women, often with some degree of coordination, aimed at deterring women from participating in the public sphere"[12].

Unfortunately, the pandemic seems to have exacerbated the practice of gender based online violence (GBOV). A study from EVAW and Glitch (2020) describes an 'epidemic' of online abuse during COVID-19, with 29% of women and non-binary individuals and 38% of black and Minoritised women and non-binary individuals in the UK reporting that they experienced more abuse during the pandemic than before.[13]

2. A common definition of gender-based online violence (GBOV) is needed to capture a wide range of behaviours and speech in future policy initiatives aimed at tackling this issue

The examples above immediately make clear that there is a wide range of terms and definitions that are used to describe a wide range of behaviours and speech that are directed at women. A number of international organisations, including the EU, have acknowledged this phenomenon by attempting to link some of these harassment techniques under the banner of GBOV. The European Commission for instance defines GBOV as

> an umbrella term used to describe all sorts of illegal or harmful behaviours against women in the online space. They can be linked to experiences of violence in real life or be limited to the online environment only. They can include illegal threats, stalking or incitement to violence, unwanted, offensive or sexually explicit emails or messages, sharing of private images or videos without consent, or inappropriate advances on social networking sites.[14]

[10] Leonie Cater, Finland's women-led government targeted by online harassment. Politico, 17 March 2021.
[11] Cecile Guerin, Eisha Maharasingam-Sha, Public Figures,Public Rage, Candidate abuse on social media. ISD, 2020 at 3.
[12] Nina Jankowicz et al., Malign Creativity – How gender, sex, and lies are weaponized against women online. Wilson Center, 2021.
[13] "The Ripple Effect: Covid-19 and the epidemic of online abuse" (September 2020), End Violence Against Women and Glitch, available at: https://www.endviolenceagainstwomen.org.uk/wp-content/uploads/Glitch-and-EVAW-The-Ripple-Effect-Online-abuse-during-COVID-19-Sept-2020.pdf
[14] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/gender-based-violence/what-gender-based-violence_en

As with 'cyber violence' and 'online hate speech', the term 'GBOV' has not been fully conceptualised yet, nor has it been adopted in binding legislation. Many studies and reports[15] recall the variety of players, institutions and committees which have looked at the question of defining cyber violence, or online hate speech with a gender perspective, and present overviews of such definitions.

The Istanbul Convention on preventing and combatting violence against women and domestic violence (2011)[16], which is the only legally binding European instrument targeting gender-based violence does not, for instance, explicitly refer to cyber violence. However, it is often quoted by EU institutions and is understood to include cyber violence as it applies "to all forms of violence against women"[17]. As such, it is important to stress that such violence is part of a continuum of violence against women and not merely a virtual phenomenon that can be neatly separated from violence in real life.

Multilateral organisations, including the UN, the Council of Europe and the EU, as well as Member States, have provided elements of definitions which differ in scope and in granularity[18]. The UN Special Rapporteur on violence against women defines the practice in general as "any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of ICT, such as mobile phones and smartphones, the internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately".[19] The European Commission's Advisory Committee on Equal Opportunities for Women and Men, in its opinion on combatting online violence against women focused more on the wide range of harms that can be the result of such violence, defining 'cyber violence against women' as an "act of gender-based violence perpetrated directly or indirectly through information and communication technologies that results in, or is likely to result in, physical, sexual, psychological or economic harm or suffering to women and girls, including threats of such acts, whether occurring in public or private life, or hindrances to the use of their fundamental rights and freedoms". [20]As a result, the EU institutions refer to definitions enshrined in the Council of Europe treaties, in UN resolutions or to definitions used in certain Member States without much consistency.[21]

---

[15] See for example, Study on Cyber violence and hate speech online against women, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the FEMM Committee, 2018, available at https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604979/IPOL_STU(2018)604979_EN.pdf

[16] "Declaration on the Elimination of Violence against Women", General Assembly resolution 48/104, 1993; Council of Europe Convention on preventing and combating violence against women and domestic violence, CETS no. 210, signed in Istanbul on 11.5.2011.

[17] Article 2, "This Convention shall apply to all forms of violence against women".

[18] See "Violence against women: an EU-wide survey", European Union Agency for Fundamental Rights (FRA), 2014, available at https://fra.europa.eu/en/publication/2014/violence-against-women-eu-wide-survey-main-results-report ; "Cyber violence against women and girls", European Institute for Gender Equality (EIGE), 2017, available at: https://eige.europa.eu/publications/cyber-violence-against-women-and-girls

[19] Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective A/HRC/38/47, UN Human Rights Council, 2018, available at https://www.ohchr.org/EN/HRBodies/HRC/.../Session38/.../A_HRC_38_47_EN.docx

[20] Opinion on combatting online violence against women, European Commission Advisory Committee on Equal Opportunities for Women and Men, April 2020, available at: https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/opinion_online_violence_against_women_2020_en.pdf

[21] Study on Cyber violence and hate speech online against women, European Parliament, 2018.

The lack of a common definition in EU law does not mean there is not a broad understanding of the type of activities that can be seen as GBOV. There are numerous studies, glossaries and categorisations available of activities that can be considered as such.

Based on over 150 interdisciplinary research papers Thomas et al helpfully distinguish 7 distinct categories of online activities that are especially relevant to counter.[22] .

Toxic content	Attacks involving media sent to a target or audience without the necessity of more advanced capabilities, including bullying, trolling, threats of violence and sexual harassment. They explain that "toxic content can be used to violate availability, preventing victims from properly taking advantage of an online community and even forcing them to leave it."
Content leakage	When an attacker leaks (or threatens to leak) sensitive, private information. This includes doxing, outing, deadnaming or non-consensual intimate imagery exposure. Often, the attacker's intent is "either to embarrass, threaten, intimidate, or punish the target".
Overloading	When an attacker forces a target to triage hundreds of notifications or comments via amplification, or otherwise makes it technically infeasible for the target to participate online due to jamming a channel. Examples include organised trolling activity orchestrated through social media, the use of 'SMS bombers' to send thousands of text messages to a target, 'zoombombing' to disrupt a video conference, 'brigading' where a large group of people overwhelm the comment feed of a targeted group or individual in coordinated 'raids'. These attacks can lead to "frustration, fatigue, and a reduced sense of emotional safety".
False reporting	When an attacker deceives a reporting system or emergency service to falsely accuse a target of abusive behaviour. This includes 'SWATing' or falsified abuse flagging (flagging a piece of content or an account as abusive)
Impersonation	When an attacker relies on deception of an audience to assume the online persona of a target in order to create content that will damage the target's reputation or inflict emotional harm. In addition to reputation harm and isolation, attackers may also use impersonation to physically and sexually threaten targets. This includes fake profiles and websites, as well as 'deepfakes' and the photoshopping of pictures.
Impersonation	When an attacker relies on deception of an audience to assume the online persona of a target in order to create content that will damage the target's reputation or inflict emotional harm. In addition to reputation harm and isolation, attackers may also use impersonation to physically and sexually threaten targets. This includes fake

[22] Thomas et al., "SoK: Hate, Harassment, and the Changing Landscape of Online Abuse", 2021 IEEE Symposium on Security & Privacy (i.e., Oakland 2021), available at: https://rist.tech.cornell.edu/papers/sok-abuse.pdf

	profiles and websites, as well as 'deepfakes' and the photoshopping of pictures.
Surveillance	When an attacker is leveraging privileged access to a target's devices or accounts to monitor the target's activities, location, or communication.
Lock-out and control	Scenarios where an attacker leverages privileged access to a target's account or device—including computers, phones, or IoT devices to gaslight the target or interfere with how they engage with the world.

From this typology, and the attempts by different entities to define GBOV, it remains clear that GBOV is not limited to illegal activities that can – and should – be criminalised. GBOV also includes a number of practices which are closely linked to, or even facilitated by, specific design features and product characteristics of online platforms such as those listed above. The impact of these behaviours has the potential to lead to the same outcome as illegal activities, including the victim's withdrawal from social media and public discourse, which can in turn undermine their overall network of relationships and have a larger societal impact due to restricting one group of the population's participation in democratic life [23].

3. The current EU legal framework fails to sufficiently address the threats posed by GBOV

The EU has adopted a wide range of instruments that regulate the online and electronic environment, including, but not limited to, the General Data Protection Regulation (GDPR), the e-Commerce Directive, the Audiovisual Media Services Directive (AVMSD), and the Code of Conduct on Countering Illegal Hate Speech Online. These instruments impose some obligations on platforms to deal with illegal content, but since they do not specifically address cyberviolence and the specific harm women experience online, their effectiveness to counter GBOV appears to remain limited. They would also not incentivise platforms to think through the impact of their design decisions in facilitating GBOV or to take action against the abuse of their services by perpetrators of GBOV. (see infra section 4)

The European treaties still offer a limited legal basis for the EU to take action in the area of criminal law, including action on illegal hate speech. However, at national level, the disparity of Member States legislation is still wide, which makes it difficult to know which cyberviolence or harassment techniques would be considered a crime or illegal hate speech. Article 19 explains that

> States take very different positions on whether different forms of online harassment and abuse should be a criminal offence, and even where there is general agreement, they may disagree on precisely how these offences should be defined and where the threshold for criminal liability might arise. Even when the legislation sets a certain severity threshold, there is a lack of comprehensive guidelines on when such a threshold is reached."[24]

[23] Combating gender-based violence: Cyber violence European added value assessment, March 2021, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf
[24] Available at: https://www.article19.org/wp-content/uploads/2020/03/Guiliano-v-Hungary-Submission-2-March-2.pdf

The Advisory Committee on Equal Opportunities for Women and Men notes that

> "cyberviolence is still a blind spot in many Member States, where online offences only exist as an excrescence of their offline counterpart. Studies show that such legal frameworks are ineffective, as the police struggle to respond to online harassment when they are left with antiquated laws and procedures. As a result, victims' interactions with the authorities and other public services often prove frustrating and women tend to refrain from reporting offences – a significant amount of women do not even acknowledge their experience as a crime".[25]

3.1 Existing EU rules require platforms to take some action against illegal content online without any specific reference to GBOV

The GDPR is the primary tool that can be invoked by victims to provide a remedy against the sharing of content to which the victim has not consented, while at the same time it empowers users with rights on control, erasure, rectification and objection to personal data processing. This would be of particular importance for victims for non-consensual pornography or revenge porn. However, it appears that in practice the GDPR has had little impact, as lawyers and researchers have called it 'not fit for purpose' when it comes to protecting the data of these victims[26].

The e-commerce directive addresses illegal content and the liability of online service providers, and it can oblige service providers to remove or disable access to illegal content hosted on their platforms. This law is currently under revision and could result in harmonising the ways platforms should react to notices of illegal content. (see section 7 below)

The AVMSD obliges Member States to ensure that audiovisual media services do not contain any "incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter", which includes discrimination based on sex.[27] Although the AVMSD does not explicitly reference 'illegal content', it mentions three categories of criminal offences under EU law, namely: public provocation to commit a terrorist offence, offences concerning child pornography and offences concerning racism and xenophobia.

Under the Code of Conduct on countering illegal speech online[28] platforms agree to have in place (1) effective mechanisms to review notifications regarding illegal hate speech (2)

---

[25] *Opinion on combatting online violence against women*, April 2020, available at: https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/opinion_online_violence_against_women_2020_en.pdf

[26] https://www.politico.eu/article/how-europe-privacy-laws-are-failing-victims-of-sexual-abuse/ "With the onus being on users to chase their private images across the Internet, to contact the individual data controllers and persuade reluctant platforms to act on the flagged material, the process can be exacting — and expensive if one chooses to outsource the work to a lawyer ».

[27] Article 21.1 of the Charter of Fundamental Rights: "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

[28] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

community guidelines that prohibit the promotion of incitement to violence and hateful conduct. They also promise to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours, and remove or disable access to such content, if necessary.[29] However, this Code does not include any explicit reference to gender-based violence and the latest monitoring report indicates that only 3.7% of the content reported within the scope of this Code was 'gender-based hate speech'[30].

3.2 Lack of harmonization of national criminal law approaches to illegal content

In terms of criminal law, there are four types of illegal content defined by EU law: child sexual abuse material; racist and xenophobic hate speech; terrorist content; and content infringing intellectual property rights. Outside of these four categories, there is no EU harmonisation of illegal content online[31]. The most interesting directives to flag for the purposes of this paper are the Framework Decision on racism and the Victims' Rights Directive.

The Framework Decision[32] establishes a common criminal law approach to racist and xenophobic hate speech and hate crimes. Its aim is to ensure "that racist and xenophobic offences are sanctioned in all Member States by at least a minimum level of effective, proportionate and dissuasive criminal penalties". The Decision defines "all conduct publicly inciting to violence or hatred directed against a group of people or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin" as an 'offence concerning racism and xenophobia' which should be punishable in national law. While it does not provide for full harmonisation of criminal laws in this field, it seeks to establish "the minimum approximation necessary to ensure that national legislation is sufficiently comprehensive".[33] The guidance note on the Framework Decision states that this approach leaves Member States free to include other protected characteristics in national legislation, "such as sexual orientation, gender identity, disability, sex/gender, social status, etc". [34]

The Victims' Rights Directive[35] protect victims of crime in the EU and provides a minimum level of rights, protection, support, access to justice and restoration. However, although the

[29] For criticism of the Code see https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/EC_Code_of_Conduct_TWG_Bukovska_May_2019.pdf
[30] https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf
[31] Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform, Study, European Parliament, June 2020, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf
[32] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, OJ L 328, 6.12.2008, available at: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32008F0913
[33] See the explanatory memorandum accompanying the proposal, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52001PC0664&from=it
[34] EU High Level Group on combating racism, xenophobia and other forms of intolerance, "Guidance note on the practical application of council Framework Decision 2008/913/JHA on Combating Certain Forms and Expressions of Racism and Xenophobia by means of Criminal Law", November 2018, available at: https://ec.europa.eu/newsroom/just/document.cfm?doc_id=55607
[35] "Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime", available at https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32012L0029

Directive directly refers to gender-based violence, observers have noted that some provisions "do not account for the specific nature of gender-based violence at all" and that it it does "not regulate the issues of support and protection for victims of gender-based violence in an optimum manner" [36] [37] The 2020 EU Victims Rights Strategy Strategy does announce the launch of an EU network on "the prevention of gender-based violence", which will "take actions to protect the safety of victims of gender-based cybercrime in particular by facilitating the development of a framework for cooperation between internet platforms and other stakeholders".[38]

3.2.1 Member States take divergent approaches to GBOV

A thorough overview of the 27 Member States' legislation is beyond the scope of this paper. At this point it is relevant to note the conclusion of a study for the European Parliament, which assessed the legislation of 12 Member States. [39] Their analysis concluded that in most cases, existing criminal law approaches are used to address GBOV although they are not specifically designed to do so.[40] They identified four different types of legislation addressing GBOV and noted that some Member States use a mix of the four approaches.

Some Member States criminalise GBOV, such as Romania and France. Romania is the only country surveyed which has implemented a general definition of GBOV which lists specific forms of GBOV and includes online harassment, hate speech, stalking, threats, publication of information or intimate content without consent and interception of communications. France passed a law in 2018 strengthening action against sexual and gender-based violence to provide better support for victims of gender-based violence, including cyber violence.

Other Member States criminalise specific types of 'cyber violence' without addressing a specific gender angle, such as Belgium, Czechia, Spain and France. For example, Italy specifically addresses 'cyberbullying' ;and 'cyber harassment' is considered a crime in Austria. Hate speech without a specific gender component is criminalised in Spain, the Netherlands, Bulgaria, Greece, Croatia, Portugal and Malta.

In a third category are Member States that use existing provisions not specific to online crimes, such as Finland, the Netherlands, Germany or Spain. Germany applies criminal provisions that cover stalking, harassing, threatening, abusing or insulting to the online environment. Spain punishes "all forms of harassment or stalking", both online and offline.

Some Member States work on the basis of non-criminal provisions: Germany introduced the NetzDG legislation which enforces obligations for social networks to manage complaints and remove non-consensual photographs.

---

[36] An analysis of the Victims' Rights Directive from a gender perspective, 19 April 2016, available at https://eige.europa.eu/publications/analysis-victims-rights-directive-gender-perspective

[37] A Union of Equality: Gender Equality Strategy 2020-2025, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A152%3AFIN

[38] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0258

[39] Belgium, Czech Republic, Germany, Spain, Finland, France, Italy, Lithuania, The Netherlands, Poland, Romania and Sweden.

[40] Combating gender-based violence: Cyber violence European added value assessment, March 2021, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf

3.3 Potential future EU initiatives could pave the way to a more harmonized approach to countering GBOV

The European Commission's 2021 Work Programme announced an intention to create a new legislative proposal in Q4 to prevent and combat "specific forms of gender-based violence against women and domestic violence". This initiative could potentially lead to a new European directive which could introduce a binding definition of the term GBOV across the EU. It also could harmonise the criminalization of a number of specific acts of GBOV. This initiative will be complemented by extending the list of so-called 'Eurocrimes' under Article 83 TFEU to cover "all forms of hate crime and hate speech".[41]

This new directive would be an interesting exercise given that section 3.2 has demonstrated that there remains significant divergences in national law to provide protection for victims of GBOV.[42] Such a directive is also likely to improve the current situation in which law enforcement and police services have routinely dismissed violence against women and girls, including sexual assault, intimate partner violence, and TFGBV[43]. However, negotiations on these initiatives might take a lot of time, given Member States' general reluctance to extend EU competence in criminal law [44] and the tensions around the concepts of gender equality[45] and hate speech.

More fundamentally, there will always be a range of behaviours and activities (as identified in section 2) that will not be – and probably should not be – subject to criminal law. In the Canadian context, Khoo makes it clear that "many forms of harassment and abuse may not amount to causes of action that would succeed in a civil lawsuit or chargeable crimes under criminal law".

> The facts of a particular situation may not meet the elements of any legal test, and the law has not yet caught up to the many new ways in which technology enables what should be actionable or chargeable acts of violence, abuse, or harassment. Moreover, pursuing a legal claim or criminal charge would require being able to identify the specific wrongdoer to sue or lay charges against. This may be difficult in the case of anonymous abusers, or in the case of mob-style abuse where hundreds or thousands of individuals may only send one or two messages insufficient to ground legal action, but results in a level of impact and harm that warrants legal recognition and redress for the victim.

Indeed – as the examples show in section 4 and section 6 – these practices can have a serious impact on the rights of the victims of these attacks, in particular their right to freedom of speech. Hence, it is important to assess the role that platforms can play in enabling,

---

[41] Commission Work Programme 2021 - A Union of vitality in a world of fragility, 2020, available at: https://eur-lex.europa.eu/resource.html?uri=cellar%3A91ce5c0f-12b6-11eb-9a54-01aa75ed71a1.0001.02/DOC_1&format=PDF

[42] See also https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf at 12

[43] Cynthia Khoo, Deplatforming misogyny (2021) at 51 https://www.leaf.ca/publication/deplatforming-misogyny/

[44] See the policy debate on "The future of EU substantive criminal law" – Report from the Romanian Presidency (2019), available at: https://data.consilium.europa.eu/doc/document/ST-9726-2019-INIT/en/pdf

[45] See for example, "Poland and Hungary battle to eradicate 'gender' in EU policies", Eszter Zalan, EUObserver, 16 December 2020, available here: https://euobserver.com/political/150395

facilitating but also preventing GBOV through their products and design decisions, and how the Digital Services Act in particular can change the incentives for platforms to tackle these problems.

4. Platforms' design choices enable and strengthen GBOV

It is often assumed that GBOV is primarily the result of human nature: hate speech is the natural product of hateful people. However, this approach neglects how this behaviour is shaped – and sometimes even encouraged – by (visual) design decisions and proprietary algorithms, which can contribute to polarising, impulsive, or antagonistic behaviours. As Munn points out: "Just as the design of urban space influences the practices within it, the design of platforms, apps and technical environments shapes our behaviour in digital space."[46] He quotes Gillespie in arguing that

> these platforms are designed to invite and shape participation toward particular ends. This includes what kind of participation they invite and encourage; what gets displayed first or most prominently; how the platforms design navigation from content to user to exchange… and how they organise information through algorithmic sorting, privileging some content over others in opaque ways. And it includes what is not permitted, and how and why they police objectionable content and behaviour.

Nir Eyal, the author of Hooked: How to Build Habit-Forming Products, has explained how certain design features from big tech companies are designed to exploit negative emotions that can act as 'triggers'.[47] Facebook's metrics deeply shape the kinds of content we post. As Michael Caulfield, the Director of blended and networked learning at Washington State University Vancouver states: "If you look at the interface of Facebook, it's completely designed to get you to share as quickly as possible with as many people as possible without leaving the site, without doing any deep analysis".[48]

The incentive structures and social cues of algorithm-driven social media sites amplify the anger of users over time until they "arrive at hate speech".[49] Mark Zuckerberg acknowledged this in his own post on content moderation in which he observed that content near but not over the edge of what Facebook allows gets more engagement.[50] Munn has argued how "based on engagement, Facebook's Feed drives clicks and views, but also privileges incendiary content, setting up a stimulus–response loop where outrage expression becomes easier and even normalized".[51]

---

[46] Munn, L. Angry by design: toxic communication and technical architectures. *Humanit Soc Sci Commun* 7**,** 53 (2020 at 1, available at https://www.nature.com/articles/s41599-020-00550-7.pdf

[47] Quoted in https://www.theguardian.com/technology/2017/oct/05/smartphone-addiction-silicon-valley-dystopia "Feelings of boredom, loneliness, frustration, confusion and indecisiveness often instigate a slight pain or irritation and prompt an almost instantaneous and often mindless action to quell the negative sensation," Eyal writes.

[48] https://www.theringer.com/2017/2/15/16038024/how-the-like-button-took-over-the-internet-ebe778be2459

[49] https://www.nytimes.com/2018/04/25/world/asia/facebook-extremism.html

[50] https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

[51] Munn at 8.

Munn further argues that Facebook and YouTube's affordances can induce experiences that are "stressful and impulsive, establishing some of the key conditions necessary for angry communication".[52] Design decisions in this sense work to reduce the barrier of outrage expression.[53] Similarly, research has demonstrated how the design of YouTube's comments system greatly exacerbates toxic comments. As Munn states, "For instance, YouTube comments can be upvoted or downvoted, but downvoting doesn't lower the number of upvotes. This suggests a design logic that favours any kind of engagement, whether positive or negative. The result is that provocative, controversial, or generally polarising comments seem to appear towards the top of the page on every video." [54]

Khoo quotes Becca Lewis, who has demonstrated how platform characteristics facilitate hate speech: "Easy feedback systems.. lead to discursive loops, in which influencers build audiences that ask for, or reward, certain types of content. Such audiences can in turn drive political influencers to deliver ever more extreme content, resulting in self-reinforcing feedback loops of content increasingly bordering or crossing into hate speech".[55]

These feedback systems can be manipulated by audiences and groups to amplify hate speech and silence other voices. Khoo lists examples of coordinated campaigns that exploit simple platform engagement features, including:

- mass upvoting or positively boosting extremist or hate-based content so that it appears on homepages, is promoted under trending topics, or is recommended to more users;
- mass downvoting feminist content or the content of marginalised individuals exposing or bringing attention to abuses so that their speech is effectively buried;
- falsely mass flagging or reporting feminist, anti-racist, or LGTBQ+ users, posts, and pages for violating community standards in order to have their accounts or pages suspended, banned, or deleted;
- upvoting, liking, and engaging with harmful content posted without consent so the algorithms will pick up and assist in further disseminating the content, including intimate photos or personal information that comes from doxing such as the person's home address

Media coverage has revealed how executives at both YouTube and Facebook ignored or shelved internal research demonstrating the platforms' propensity to systematically amplify and promote abusive speech and hate-based rhetoric. The Wall Street Journal[56] obtained internal research from Facebook from 2016 which states that "64% of all extremist group joins are due to our recommendation tools" and that most of the activity came from the platform's "Groups You Should Join" and "Discover" algorithms. When a task force was created to propose potential fixes, such as tweaking the recommendation algorithms to suggest a more diverse range of groups for people to join, the ideas were deemed "antigrowth" and were either abandoned or weakened.

---

[52] Munn at 3
[53] Munn at 5
[54] Munn at 8
[55] Khoo at 57
[56] Jeff Horwitz and Deepa Seetharaman, Facebook Executives Shut Down Efforts to Make the Site Less Divisive. Wall Street Journal, 26 May 2020 https://www.wsj.com/articles/facebook-knows-it-encourages-division-top- executives-nixed-solutions-11590507499

A similar pattern was revealed by MIT Technology Review this year, when internal research from 2018 demonstrated that internal "study after study" proved that Facebook's business model based on maximising engagement increases polarisation. When the company's internal research team proposed to tweak the content-ranking models so that, for instance, "users with a tendency to post or engage with melancholy content – a possible sign of depression" weren't recommended "increasingly negative material that risked further worsening their mental health", their suggestions weren't followed up on because such measures could reduce engagement[57].

According to Bloomberg, corporate leadership at YouTube was unable or unwilling to act on internal alarms about the role of the platform in spreading a "mass of false, incendiary and toxic content" for fear of throttling engagement, prioritising profits over the safety of its users. "When people at YouTube realised that outrage equals attention, instead of revamping the recommendation engine, YouTube doubled down to achieve its goal of keeping people on the platform"[58].

All of this has a direct impact on the proliferation of GBOV, especially when it manifests in publicly shared content and/or leverages some of the techniques to 'game' the platform's recommendation engines. Perpetrators of gender-based violence can use these techniques to amplify content which marginalises and attacks women, whilst employing the same methods to silence or target the voices of women (especially female politicians) and those who support them.

5. Current voluntary platform mechanisms to counter GBOV are not adequate

In a groundbreaking study Khoo argues that despite the widely documented proliferation of GBOV on digital platforms and its impacts on women and girls, as well as online abuse perpetrated against platform users with other marginalized identities, "platform companies have on the whole done poorly to address the issue". [59] She states that

> platforms have also displayed a certain degree of selective attentiveness and double standards to the extent they have developed and applied their content moderation policies, such as determining that abusive content does not violate any policies while removing the content or suspending the account of users who were the victims of abuse, or for pointing out the existence of such abuse

She is also critical about the effectiveness of user flagging and reporting – two of the major tools nearly all platforms use in their content moderation to counter GBOV and hate speech. She argues that users often misunderstand these tools, and they can even be purposely gamed to have the opposite effect by silencing members of marginalised communities. More fundamentally, she argues that addressing GBOV through user flagging and reporting constitutes

[57] Karen Hao, How Facebook got addicted to spreading misinformation. MIT Technology Review, 11 March 2021, https://www .technologyreview .com/2021/03/11/1020600/facebook-responsible-ai-misinformation/
[58] Mark Bergen, YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant. Bloomberg, 2 April 2019. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic- videos-run-rampant
[59] Khoo at 50.

> a complaint-driven system that addresses GBOV entirely on a one-off, case-by-case, reactive basis, assuming a user has the wherewithal to report content in the first place, rather than a systemic, platform-wide, and proactive response that would mitigate or prevent the proliferation of online abuse to begin with. Such a system
>
> 'responsibilises' users - and disproportionately marginalised users at that - who are targets of speech-based abuse, and forces them to bear the burden of seeking accountability, redress, and functioning content moderation in addition to living through the impacts of the abuse itself. [60]

In addition, the platforms' policies and processes to tackle abuse are complicated to navigate and, in most cases, not specific to GBOV. Article 19 found[61] that Facebook, Twitter and YouTube's regulations on gender-based harassment and abuse, including the terms of services and community guidelines and their enforcement, were drafted in broad terms, giving companies flexibility to interpret them according to their own needs, which results in "inconsistent and sometimes apparently biased outcomes". Because these policies apply to all users and are not geared towards women in particular, Article 19 found it "difficult to know what content actually gets removed from these platforms", which is particularly true for "gender-based harassment and abuse that women journalists face on these platforms."

On enforcement, Article 19 points out that enforcement mechanisms are "not always easy to find" and that "significant shortfalls" remain, with the companies' responses being often "lacking" or "inconsistent with the stated objectives of the policies". Overall, they mention the lack of transparency over the actions taken in response to gender-based harassment and abuse, and especially the important shortcomings of the platforms' transparency reports.

Fundamentally, the onus is on the victim to regulate the attention their profiles receive and report abusive comments, which is not only burdensome for the victim but also does not make women feel supported in their fight against the abuse they face. It is time for a paradigm shift that does not see GBOV as a responsibility for the target to defend herself against, but as an infringement of fundamental rights, not only to the Charter Right to gender equality (article 23) but also as an interference with freedom of speech, that must be prevented.

6. Allowing GBOV to proliferate on social media undermines, rather than promotes, freedom of speech

Platforms need to be incentivised to take mitigating measures against the risks their products and services pose in facilitating GBOV in order to protect the freedom of speech of the victims of these attacks.

To meaningfully and effectively enjoy the right to freedom of expression, it should be exercised freely and without fear of violence and abuse. Faced with inadequate responses to

[60] Khoo at 77.
[61] Online harassment and abuse against women journalists and major social media platforms, Article 19, 2020, available at: https://www.article19.org/wp-content/uploads/2020/11/Gender-Paper-Brief-2.pdf

the GBOV they experience, a number of studies (see section 1) show that victims resort to changing their behaviour, from self-censorship to leaving the online sphere altogether. [62].

Amnesty International has been documenting the scale and the nature of online abuse and violence against women since 2017. Its Toxic Twitter project is unequivocal in finding that GBOV leads to a chilling effect on women speaking out online, with "far-reaching and harmful repercussions on how younger women, women from marginalised communities, and future generations fully exercise their right to participate in public life and freely express themselves online".[63] The report succinctly summarises the harms that are the result of GBOV.

> For many women, the inability to fully participate and express themselves equally online means that they are absent from public conversations they would like to be part of, and sometimes, *need* to be part of. To not engage or comment on an issue out of fear of violence and abuse means that certain women's voices are not represented on Twitter and that women are no longer part of the debate. For women in the public eye, in particular, this can have a detrimental effect on their career and building networks. The silencing effect of online abuse on women, including on Twitter, may also send a worrying message to younger generations that women's voices are not welcome.[64]

This one-sided abuse can not only lead to a loss of diversity in public debates but even has an impact on the nature of the democratic process itself. Di Meco and Brechenmacher have explained how the targeting of female politicians and activists has actually discouraged women from running for office or has led them to "disengage from online political discourse in ways that harms their political effectiveness". [65] Even more, they state, "for those women who persevere, the abuse can cause psychological harm and waste significant energy and time, particularly if politicians struggle to verify whether or when online threats pose real-life dangers to their safety".[66]

Despite these effects GBOV is often trivialised. Victims are often told to 'ignore the trolls', with commentators considering their experience as a mere inconvenience or dismissing it as 'pranks' or 'locker-room talks'[67]. This is partly because some of the behaviour is seen as legally insignificant in the same way - as Citron has pointed out - "as prosecutors once refused to file charges in cases involving gender-specific sexual assaults such as domestic violence and rape"[68].

---

[62] Danielle K. Citron, Law's Expressive Value in Combating Cyber Gender Harassment, 108 MICH. L. REV. 373 (2009). Available at: https://repository.law.umich.edu/mlr/vol108/iss3/3

[63] Available at https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-5/ , see also "Unsocial Media: The Real Toll of Online Abuse against Women (2017)", available at: https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4 and the Troll Patrol project: https://decoders.amnesty.org/projects/troll-patrol/findings

[64] Idem.

[65] "Tackling Online Abuse and Disinformation Targeting Women in Politics", Lucina Di Merco and Saskia Brechenmacher, November 2020, Carnegie Endowment for International Peace, available at: https://carnegieendowment.org/2020/11/30/tackling-online-abuse-and-disinformation-targeting-women-in-politics-pub-83331

[66] Ibid.

[67] Law's Expressive Value in Combating Cyber Gender Harassment, Danielle Keats Citron, Michigan Law Review, Volume 108, issue 3, 2009, https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1300&context=mlr

[68] Ibid

Often caution is urged when governments or platforms want to mitigate these abusive practices online as they can be seen as leading to a 'chilling effect' on freedom of speech online. Yet such claims neglect the chilling effect such practices have on the freedom of speech of the target of these practices. Allowing individuals to engage in GBOV protects the freedom of speech of the speaker, but this comes at the expense of the target's freedom of expression. Arguments to allow harmful but legal GBOV rests on the idea that such expressions are ultimately valuable because it leads to the 'discovery and acceptance of truth' given that counterspeech will ultimately bring up the best arguments on a given topic. But this argument does not hold up in a GBOV context.

Ultimately, it needs to be stressed that the freedom of expression has never been an absolute right in the EU, and can legitimately be restricted if it means protecting the rights of others. In the Charter of Fundamental Rights, article 23 provides for the equality between women and men. The right to equality must reflect substantive equality i.e. the notion that to achieve true equality, people in different positions may have to be treated differently. A detached, marketplace-of-ideas libertarian philosophy might elevate freedom of speech over other rights in a US-context (the infamous "gravitational pull of the 1st Amendment"), but this approach does not work in the European Union. While the European Court of Human Rights has not yet looked at the issue of sexist or misogynist hate speech, Judge Pinto de Albuquerque has argued that "the full effet utile of the (Convention) can only be achieved with a gender-sensitive interpretation and application of its provisions which takes into account the factual inequalities between women and men and the way they impact on women's lives."[69] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression David Kaye recommended a substantive equality approach to online content moderation in his 2018 thematic report on content regulation to the UN Human Rights Council:

> Meaningful guarantees of non-discrimination require companies to transcend formalistic approaches that treat all protected characteristics as equally vulnerable to abuse, harassment and other forms of censorship [….] Instead, when companies develop or modify policies or products, they should actively seek and take into account the concerns of communities historically at risk of censorship and discrimination

As a result, restricting GBOV can actually promote freedom of expression. As Khoo has argued

> The choice here isn't between free speech and censorship, it's between who will and won't be heard. Indeed, the choice is not even between who will and will not be heard, because those whose expression would be chilled by laws restricting online abuse can still be heard, since they remain as free to engage in non-abusive expression as they have been able to all along. Thus, the choice is only whether or not women, and other marginalised and vulnerable groups and individuals who are silenced by TFGBV (gender-based violence), will be heard or not[70].

---

[69] Cyberviolence, domestic abuse and lack of a gender-sensitive approach – Reflections on Buturuga versus Romania, Strasbourg Observers, March 2020, available at: https://strasbourgobservers.com/2020/03/11/cyberviolence-domestic-abuse-and-lack-of-a-gender-sensitive-approach-reflections-on-buturuga-versus-romania/#more-4542

[70] Khoo at 192

In the EU, governments can actually be seen as having a positive obligation to protect their citizens from this intrusion on their freedom of speech rights. By imposing new rules on tech companies they can legitimately enhance the fundamental freedoms for

> those whose ability to participate has been impeded by privately imposed forms of bigotry and control. In fact, for those whose democratic participation has been limited by non-state actors' discriminatory conduct, equality may appear as the matrix from which the substantive ability to experience other rights and freedoms flows.

7. How can the Digital Services Act encourage platforms to prevent and restrict GBOV?

In December 2020, the European Commission published its proposal for the Digital Services Act (DSA), consolidating and updating different pieces of legislation on illegal content and intermediary liability. The DSA will regulate how to deal with content which is illegal either according to EU or national laws and includes cross-border removal orders between authorities and services established in different EU countries.

It also contains other provisions which could be relevant for victims of GBOV such as the obligation to establish points of contact and legal representatives, a complaint and redress mechanism, out of court dispute settlement, trusted flaggers, measures against abusive notices and counter-notices, codes of conduct, and crisis response cooperation[71].

The DSA clarifies responsibilities and accountability for providers of intermediary services, and in particular online platforms such as social media. It maintains existing key principles, such as the exemption of liability for intermediaries for the content users post, and imposes a duty, upon obtaining actual knowledge or awareness of illegal content, to act expeditiously to remove or to disable access to that content. The prohibition on general monitoring of content is also maintained, while allowing platforms to undertake activities aimed at detecting, identifying and acting against illegal content.

The DSA regulates the process of content removal but not the content itself and so, similarly, it does not regulate speech. While the harmonisation of procedures around the removal of illegal content is a good step in the fight against GBOV, section 3.2.1 above highlights that national legislations do not take a harmonised approach to defining which aspects of GBOV are illegal content and which ones are not.

The DSA can be amended in order to ensure platforms take proportional and effective measures to prevent GBOV from taking place and being amplified. This paper identifies two categories of amendments *not* to pursue which we believe would be detrimental from a human rights perspective (a ban on online anonymity and including harmful but legal categories of GBOV in the DSA), and one category of amendments that would achieve the goal of preventing GBOV from being shared and amplified online while respecting freedom of speech.

[71] European added-value assessment, 2021.

7.1 What not to do (I): ban online anonymity

On a practical and technical level, the online environment, coupled with an online disinhibition effect provides for an easier ground for large scale abuse. Some argue that online anonymity plays a large role in GBOV and have advocated for a ban on online anonymity as a way to tackle such violence by enabling the perpetrators to be 'unmasked' and held to account. There have been discussions on introducing new legislations in Member States[72] to force internet users to identify themselves, using their real names or registering on platforms with official IDs.

These kinds of proposals come from a misconception about online anonymity and the roots of online hate speech and GBOV, all of which are complex concepts and phenomena.

First, anonymity can be a key enabler for freedom of speech, especially for historically marginalised groups as it allows them to freely seek information, engage in political debate and develop ideas without fear of reprisals.[73] Online anonymity or pseudonymisation can be a key protective measure against GBOV. According to one study, 43% of online harassment victims changed their contact information to protect themselves from abuse.[74]

Secondly, according to research there is no conclusive evidence that displaying names and identities will reliably reduce social problems[75]. In fact, Matias argues that "forcing real names in online communities could also increase discrimination and worsen harassment". He explains that "Gender- and race-based harassment are only possible if people know a person's gender and/or race, and real names often give strong indications around both of these categories."

Thirdly, anonymity has been recognised as being linked with fundamental rights such as the right to privacy, data protection and freedom of expression. The United Nations Special Rapporteur on freedom of expression, David Kaye, in his 2015 report[76], recognised that encryption and anonymity, as leading instruments for online security, enable people to exercise their rights to freedom of opinion and expression and the right to privacy in the digital sphere.

7.2 What not to do (II): impose liability rules for harmful content

As stated in section 2, many aspects of GBOV would not amount to illegal content and as such it should not be defined in the DSA, nor be linked to the liability regime of the DSA. The Commission has correctly explained that there is "a general agreement among stakeholders that 'harmful' (yet not, or at least not necessarily, illegal) content should not be

---

[72] See for example, https://www.politico.eu/article/austrian-conservatives-want-to-end-online-anonymity-and-journalists-are-worried/

[73] https://www.theguardian.com/technology/2020/apr/04/social-media-giants-must-tackle-trolls-or-face-charges-poll

[74] Amanda Lenhart, Michelle Ybarra, Kathryn Zickuhr, and Myeshia Prive-Feeney. Online Harassment, Digital Abuse, and Cyberstalking in America. Report, Data & Society Institute, November 2016.

[75] J. Nathan Matias, the Real Name Fallacy, 2017, https://coralproject.net/blog/the-real-name-fallacy/

[76] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 2015, available at: https://www.undocs.org/A/HRC/29/32

defined in the Digital Services Act and should not be subject to removal obligations, as this is a delicate area with severe implications for the protection of freedom of expression"[77].

7.3 Improve the risk-based approach for VLOPs in the DSA in order to mitigate GBOV

The DSA sets up a "supervised risk management approach" in which certain substantive obligations are limited only to "very large online platforms" (VLOPs)[78], which due to their reach have acquired a central, systemic role in facilitating the public debate[79]. This would cover services who have an average of 45 million monthly active users, which corresponds to roughly 10% of the EU's population.

The European Commission argues that the way these VLOPS design their services is generally "optimised to benefit their often advertising-driven business models"; which can cause "societal concerns". In the absence of effective regulation and enforcement, these VLOPS "can set the rules of the game, without effectively identifying and mitigating the risks and the societal and economic harm they can cause"[80].

As a result, the DSA imposes an obligation on VLOPs to assess the systemic risks that stem from the functioning and use of their service. Platforms then need to take risk mitigating measures [81] which can be audited. If an audit report recommends operational recommendations that the platform doesn't implement, it can ultimately be sanctioned by the EC.

This process-based approach is the best rights-respecting approach to deal with societal risks such as GBOV. While the current draft of the Commission is a good starting point, several tweaks could strengthen the proposal even more.

7.3.1 Ensure that the DSA forces platforms to 'price in' societal risks to their business operations

The current draft does not sufficiently incentivise VLOPs to assess and manage the *probability* of risk occurring through their services and operations. Including such a reference would force VLOPs to 'price-in' societal risk to their business operations, so that the decision as to what to build and how to build it is not merely based on profit maximisation, growth and increased engagement considerations that take no account of unforeseen downside risks such as enabling GBOV.

The current draft only focuses on managing the *severity* of risks, which incentivises interventions that occur only *after* the risk has materialised. While obviously necessary, these *ex-post* mitigation measures are not sufficient. Platforms need to be forced to take into account the ways in which design choices and operational approaches can influence and increase the risks as defined in article 26.1.(a)-(c).

---

[77] Explanatory memorandum, DSA proposal, available at: https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital
[78] DSA article 25
[79] DSA recital 53.
[80] DSA recital 56.
[81] DSA article 27.

The EC hints at the importance of this in article 26.2 but it is crucial to get this wording right in 26.1 because the definition of risks has a direct impact on other provisions, such as article 31 on access to data for researchers, which provides that vetted researchers can be granted access "for the sole purpose of conducting research contributing to the identification and understanding of systemic risks as set out in art. 26.1".

Finally, since risks need to be assessed throughout the product development cycle, risk assessments need to be done on an ongoing basis

Suggested amendment

Article 26.1
1. Very large online platforms shall identify, analyse and assess, from the date of application referred to in the second subparagraph of Article 25(4), ~~*at least once a year thereafter*~~, **on an ongoing basis, the probability and severity of** any significant systemic risks stemming from the functioning and use made of their services in the Union. This risk assessment shall be specific to their services and shall include the following systemic risks:

7.3.2 Widen the scope of article 26.1.b and include references to gender equality throughout the text

The current language of article 26(1)b is in contrast with recital 57, which expressly clarifies that the list of rights is not exhaustive. Indeed, recital 57 explains that the "second category concerns the impact of the service on the exercise of fundamental rights, as protected by the Charter of Fundamental Rights, *including* the freedom of expression and information, the right to private life, the right to non-discrimination and the rights of the child". Therefore article 26(1)b should be amended to include other fundamental rights and to explicitly clarify that the list is not exhaustive, and the assessment can cover any other relevant risk potentially impacting fundamental rights. As demonstrated by this paper, it would be advisable to explicitly include a reference to article 23 of the Charter as well.

Suggested amendments

Article 26.1(b)
Any negative effects for the exercise of **any of the fundamental rights listed in the Charter, in particular on** the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination, **the right to gender equality** and the rights of the child, as enshrined in Articles 7, 11, 21, **23** and 24 of the Charter respectively.

The same language would also need to be included in recitals 41 and 57

Recital 34, as it mentions the 'safety and trust of the recipients of the service', should be amended to include 'women' to allow for a gender perspective in this public policy objective. For a more inclusive amendment, the mention of vulnerable users 'with protected characteristics under Article 21 of the Charter' would allow for the consideration of users safety in relation to race, religion, sexual orientation etc.

Suggested amendment

Recital (34): In order to achieve the objectives of this Regulation, and in particular to improve the functioning of the internal market and ensure a safe and transparent online environment, it is necessary to establish a clear and balanced set of harmonised due diligence obligations for providers of intermediary services. Those obligations should aim in particular to guarantee different public policy objectives such as the safety and trust of the recipients of the service, including minors, **women** and vulnerable users, **such as those with protected characteristics under Article 21 of the Charter**, protect the relevant fundamental rights enshrined in the Charter, to ensure meaningful accountability of those providers and to empower recipients and other affected parties, whilst facilitating the necessary oversight by competent authorities.

7.3.3 Preserve the original wording of article 26.1.c

This wording is crucial to force very large online platforms to take risk mitigating measures against online harassment campaigns, for instance those as described in section 4.

7.3.4 Create a duty to justify the choice of risk mitigating measures taken

The DSA provides a wide degree of flexibility to VLOPs to assess how they can best mitigate the risks they identify. This flexibility makes sense given the different design features and business models that underlie the operations of different VLOPs. However, this flexibility cannot be an excuse to pick and choose the least intrusive measure from this sensible list of general risk mitigating measures. A platform would need to argue why it does not choose one or more of the proposed measures in articles 27(1)a-e in order to provide more information to the auditors, which can then better assess the necessity, effectiveness and proportionality of the (combination of) risk mitigating measures. This amendment would require a new addition to article 27(1) (see below).

Suggested Amendment

Article 27

1. Very large online platforms shall put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 26. Such measures ~~may~~ **shall** include, where applicable:

(a) adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions;

(b) targeted measures aimed at limiting the display of advertisements in association with the service they provide;

(c) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk;

(d) initiating or adjusting cooperation with trusted flaggers in accordance with Article 19;

(e) initiating or adjusting cooperation with other online platforms through the codes of conduct and the crisis protocols referred to in Article 35 and 37 respectively.

1. a Where a very large online platform decides not to put in place any of the mitigating measures listed in article 27.1, it shall provide a written explanation that describes the reasons why those measures were not put in place, which shall be provided to the independent auditors in order to prepare the audit report in article 28.3

7.3.5 Ensure the supervised risk management approach can be overseen by third parties

The current draft runs the risk of being interpreted too narrowly as a mere tick-box exercise. The obligation for a platform in article 26 is to conduct a risk assessment and take risk mitigating measures as a response to that assessment. For an audit to have added value it needs to be able to assess the quality of the risk assessment, and the necessity, proportionality and effectiveness of the risk mitigating measures referred to in Article 27. This is also a standard practice in other sectors.

Suggested amendment

Article 28

1. Very large online platforms shall be subject, at their own expense and at least once a year, to audits to assess compliance with the following:

(a) the obligations set out in Chapter III; **in particular the quality of the identification, analysis and assessment of the risks referred to in Article 26, and the necessity, proportionality and effectiveness of the risk mitigation measures referred to in Article 27**

7.3.6 Ensure the European Commission has a duty to enforce these Chapter III obligations vis-à-vis VLOPs

Any regulation is only as good as its enforcement. The European Commission ultimately has the power to decide whether to investigate and start proceedings when a very large online platform has infringed any of the provisions of Section 4 of Chapter III, including articles 26 and 27. However, the current draft does not make such an investigation and the initiation of the relevant proceedings mandatory, which would be crucial for this provision to be effective.

Suggested amendments

Article 50

The Commission acting on its own initiative, or the Board acting on its own initiative or upon request of at least three Digital Services Coordinators of destination, ~~**may**~~ **shall,** where it has reasons to suspect that a very large online platform infringed any of those provisions, recommend the Digital Services Coordinator of establishment to investigate the suspected infringement with a view to that Digital Services Coordinator adopting such a decision ~~**within a reasonable time period.**~~ **without undue delay and in any event within two months**.

Article 51

1. The Commission, acting either upon the Board's recommendation or on its own initiative after consulting the Board, ~~**may**~~ **shall** initiate proceedings in view of the possible adoption of decisions pursuant to Articles 58 and 59 in respect of the relevant conduct by the very large online platform that:

(a) is suspected of having infringed any of the provisions of this Regulation and the Digital Services Coordinator of establishment did not take any investigatory or enforcement measures, pursuant to the request of the Commission referred to in Article 45(7), upon the expiry of the time period set in that request;

(b) is suspected of having infringed any of the provisions of this Regulation and the Digital Services Coordinator of establishment requested the Commission to intervene in accordance with Article 46(2), upon the reception of that request;

(c) has been found to have infringed any of the provisions of Section 4 of Chapter III, upon the expiry of the relevant time period for the communication referred to in Article 50(4).

2.~~**Where**~~ **When** the Commission ~~**decides to**~~ initiate**s** proceedings pursuant to paragraph 1, it shall notify all Digital Services Coordinators, the Board and the very large online platform concerned.