Regulation of online gender-based hate speech and international human rights law: Current status and challenges

Maria Sjöholm[*]

1. *Preliminary remarks*

Discussions of the merits and feasibility of regulating gender-based hate speech – at the domestic and international level – have come to the fore with the expansion of the Internet. Hate speech in general, and sexist speech in particular, has escalated exponentially with the development of the social web.[1] While gender inequality is the main underlying cause of gender-based hate speech,[2] numerous sociological and technical factors are considered causal to the growth of hateful speech online. This includes user anonymity, mob behaviour, the ease of rapidly disseminating information globally and the lackluster regulation of cyberspace.[3] In turn, non-regulation is reflective of both practical and ideological challenges. With large sections of the Internet *de facto* governed by private corporations, state control over user content is limited, further undermined by ineffective enforcement of state jurisdiction in cyberspace. Meanwhile, broad differences at the global level *vis-à-vis* restrictions of the freedom of expression entail that international regulation of harmful speech, including hate speech, is contentious *per se*. This is even more so on the Internet.

[*]Associate professor, School of Behavioural, Social and Legal Sciences, Örebro University, Sweden.

[1] CoE, 'Recommendation CM/Rec(2019)1 on Preventing and Combating Sexism' (adopted by the Committee of Ministers of the Council of Europe on 27 March 2019) 11.

[2] ibid.

[3] AM Major, 'Norm Origin and Development in Cyberspace: Models of Cybernorm Evolution' (2000) 78 Washington U L Quarterly 59, 76; A Barak, 'Sexual Harassment on the Internet' (2005) 23 Social Science Computer Rev 77, 89.

Similar issues arise in the treatment of gender-based hate speech in international human rights law (IHRL). Expanding hate speech to include gender as a protected ground is controversial. The process of defining gender-based hate speech and delineating the scope of intermediary liability and content moderation is also challenging. Nevertheless, while sexism is not a new phenomenon – and finds a range of expressions – its amplification through the Internet calls for a consideration in IHRL of contemporary forms of hate speech. This includes recognizing the systematic causes and harmful effects of gender-based hate speech at an individual-, group- and society-based level.

2. *Expanding hate speech to 'sex', 'gender' or 'women'*

There is no widely accepted definition of hate speech *per se* in IHRL. Rather, different forms of hateful and inciting speech are regulated in treaties,[4] or have been included by way of treaty interpretation.[5] Protection against hate speech mainly extends to select identity characteristics, such as race, ethnicity and nationality.[6] This is, in part, a result of human rights law treaties reflecting world events at the time of negotiation. The dilution of the concept of hate speech – and over-regulation of speech – has also been a concern. This was raised as an objection during treaty negotiations of the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), in response to a suggestion to include the elimination of sexist hate speech.[7] The UN Special Rapporteur on Freedom of Opinion and Expression has also noted the risk of

[4] Art 4 of International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entered into force 4 January 1969) 660 UNTS 195; art 20(2) of the International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171; the CoE Additional Protocol to the Convention on Cybercrime, Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed through Computer Systems (ETS No. 189) 28 January (Additional Protocol to the Budapest Convention).

[5] For example, *Carl Jóhann Lilliendahl v Iceland* App no 29297/18 (ECtHR, 12 May 2020) paras 33-39.

[6] Art 20(2) ICCPR (n 4); art 1 ICERD (n); art 2 (1) Additional Protocol to the Budapest Convention (n 4).

[7] Convention on the Elimination of All Forms of Discrimination against Women (adopted 18 December 1979, entered into force 3 September 1981) 1249 UNTS (CEDAW). *See* LA Rehof, *Guide to the Travaux Préparatoires of the United Nations*

devaluing the term hate speech, particularly in the digital age, as it may lead to excessive limitations on the freedom of expression.[8] Meanwhile, the impracticability of regulating sexist speech has been offered as a possible explanation for the reluctance to extend hate speech to 'sex', 'gender' or 'women', in view of the prevalence of misogynistic terminology in everyday language.[9] This is particularly the case online. Arguably, it may lead to excessive censoring and undermine the effective operation of the Internet. This accordingly implies that sexist speech is not as harmful as other types of hateful speech or, alternatively, is not sufficiently harmful to warrant regulation, in the balancing of interests.

Certain steps in regulating gender-based hate speech have nevertheless been taken, primarily at the regional level. This is reflective of a general expansion of the concept of hate speech in regional organisations and courts.[10] Most prominent is the proposed EU Directive on violence against women and its provision on cyber incitement to violence or hatred, as well as CoE recommendations on sexism and sexist hate speech.[11] UN bodies have not been on a similar trajectory. The reason may be that a more cohesive approach to acceptable limitations on the freedom of expression – and the concepts of 'sex' and 'gender' – can be found at the regional level.[12]

Convention on the Elimination of All Forms of Discrimination against Women (Martinus Nijhoff Publishers 1993) 78.

[8] UN, 'Companies "Failing" to Address Offline Harm Incited by Online Hate: UN expert' (21 October 2019) <https://news.un.org/en/story/2019/10/1049671>.

[9] A Brown, 'The Who Question in the Hate Speech Debate: Part 2: Functional and Democratic Approaches' (2017) 30 Canadian J L & Jurisprudence 23, 54.

[10] *See*, for example, IACmHR, 'Hate speech and Incitement to Violence against Lesbian, Gay, Bisexual, Trans and Intersex Persons in the Americas' Annual Report of the Inter-American Commission on Human Rights vol II (31 December 2015) OEA/Ser.L/V/II para 17; CoE 'Recommendation CM/Rec(2022)16 of the Committee of Ministers to Member States on Combating Hate Speech' (Adopted on 20 May 2022 at the 132nd Session of the Committee of Ministers) Appendix para 2.

[11] Art 10 of the European Commission, 'Proposal for a Directive of the European Parliament and of the Council on Combating Violence against Women and Domestic Violence', COM(2022) 105 final (8 March 2022); CoE, 'Recommendation CM/Rec(2019)1 on Preventing and Combating Sexism' (n 1).

[12] However, the potential regulation of hate speech in the CoE Budapest Convention and the concept of 'gender' in the CoE Istanbul Convention was also contentious.

So far, gender-based hate speech has been sporadically mentioned by UN bodies, albeit increasingly so. Broadly – and unconnected to a specific treaty – the UN Strategy and Plan of Action on Hate Speech includes 'gender' as a protected ground.[13] UN treaty bodies have also applied treaties in a manner that encompasses sexist hate speech. One path has been to adopt an intersectional approach, addressing hate speech against women of a certain ethnicity, nationality or religion, that is, connecting 'sex' to more widely accepted identity characteristics in the hate speech corpus.[14] An intersectional approach is essential as studies indicate that certain groups of women are particularly targeted. This includes women of certain ethnicities, LGBTQI+ and women active in the public sphere, such as journalists, human rights defenders and politicians.[15] However, intersectionality has mainly been applied in the form of adding 'women' to other statutes.

Another route has been to address 'women', 'sex' or 'gender' as standalone grounds for hate speech, in connection to a range of human rights. The rights applicable correlate with the purported harm of gender-based hate speech. From the perspective of the individual, hate speech may harm the integrity and dignity of the person, for example, involving the right to privacy.[16] More frequently, the group-based harm of hate speech is noted. Even when an individual is targeted, it is on the basis of his/her perceived membership in a protected group. Albeit the cause and effect of speech and social harm is contested,[17] the general view in IHRL is that hate speech undermines access to a range of human rights

[13] UN Strategy and Plan of Action on Hate Speech (2019) 2 <www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf>.

[14] See, for example, CERD, 'General Recommendation No. 35: Combating Racist Hate Speech' (26 September 2013) UN Doc CERD/C/GC/35 para 6.

[15] UN HRC, 'Promotion, Protection and Enjoyment of Human Rights on the Internet: Ways to Bridge the Gender Digital Divide from a Human Rights Perspective' (5 May 2017) UN Doc. A/HRC/35/9 para 36.

[16] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (30 July 2021) UN Doc A/76/258 para 23; *Aksu v Turkey* App nos 4149/04 and 41029/04 (ECtHR, 15 March 2012) para 58.

[17] I Gagliardone et al for UNESCO 'Countering Online Hate Speech' (2015) 54 <https://unesdoc.unesco.org/ark:/48223/pf0000233231>.

for specific groups – including the freedom of expression.[18] This in turn has consequences for democracy. The causal link between hate speech, social instability and hate crimes – including gender-based violence – is also often raised.[19] As a consequence of this approach to harm, the connection between discrimination and gender-based hate speech has been the most prominent. Given that one of the main aims of regulating hate speech is to ensure the principle of non-discrimination, and that the concept is commonly understood to include incitement to discrimination, this aligns with its purpose.[20]

At a general level, calls for correlating provisions on hateful and inciting speech with those on non-discrimination have been made. At times, this focuses on gender equality. For example, the UN Special Rapporteur on Freedom of Opinion and Expression has encouraged the inclusion of gender-based hate speech in Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR) – despite its textual limitations – in view of the gender equality clauses of the Convention.[21] The CEDAW Committee has in concluding observations primarily connected sexist hate speech to its provision on harmful gender stereotyping, which is considered both a cause and form of discrimination.[22] It has, to a more limited extent, categorised online harassment and hate speech as new forms of gender-based violence, drawing on General Recommendation No 35.[23] Sexist hate speech has also, briefly, been addressed in relation to gender-

[18] CERD, 'General Recommendation No. 35: Combating Racist Hate Speech' (n 14) para 26.

[19] See, for example, CoE GREVIO, 'General Recommendation No 1 on the Digital Dimension of Violence against Women' adopted on 20 October 2021 para 39.

[20] Art 2 (1) of the Additional Protocol to the Budapest Convention (n 4); UN HRC, 'Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence', Appendix (11 January 2013) UN Doc A/HRC/22/17/Add.4 para 29(a).

[21] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 70.

[22] CEDAW, 'Concluding Observations on the Seventh Periodic Report of Finland' (10 March 2014) UN Doc CEDAW/C/FIN/CO/7 para 14; CEDAW, 'Concluding Observations on the Ninth Periodic Report of Norway' (22 November 2017) UN Doc CEDAW/C/NOR/CO/9 para 22

[23] CEDAW, 'Concluding Observations on the Ninth Periodic Report of Spain' (31 May 2023) UN Doc CEDAW/C/ESP/CO/9 para 25 (g); CEDAW, 'Concluding Observations on the Ninth Periodic Report of Germany' (31 May 2023) UN Doc CEDAW/C/DEU/CO/9 paras 17 and 18(b).

based violence by the UN Special Rapporteur on Violence against Women and Girls, obliging states to regulate online fora advocating violence against women.[24] While the causal connection between harmful stereotypes and gender-based violence is not new – affirmed in a range of treaties, case law and soft law sources – the categorisation of sexist hate speech as a form of violence *per se* is important. This mirrors the linguistic concept of illocutionary speech acts – that certain forms of speech *constitute* discrimination or violence, as opposed to solely being causal.[25] This fits within the concept of gender-based violence in IHRL, which encompasses verbal acts.[26] Regional human rights law treaties on violence against women thus become applicable, in relation to both stereotyping and violence.

The effects of gender-based hate speech on women's freedom of expression have also been raised by the UN Special Rapporteur on Freedom of Opinion and Expression, in connection to non-discrimination.[27] Hate speech can silence vulnerable groups by causing them to retreat from public fora such as the Internet, and thus undermines democratic values. It is not only problematic from the standpoint of the individual but also limits the representation and visibility of alternative viewpoints and critique. The regulation of hate speech and the freedom of expression are accordingly understood as 'mutually supportive' from the standpoint of equality.[28]

Given the *ad hoc*-based inclusion of gender-based hate speech in existing provisions, different concepts are currently used by UN bodies.

[24] UNHRC, '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' (18 June 2018) UN Doc A/HRC/38/47 paras 31, 37.

[25] J Butler, *Excitable Speech: A Politics of the Performative* (Routledge 1997) 39.

[26] See, for example, CEDAW, 'General Recommendation No 35 on Gender-Based Violence against Women, Updating General Recommendation No 19' (14 July 2017) UN Doc CEDAW/C/GC/35 para 14; the CoE Convention on Preventing and Combating Violence against Women and Domestic Violence (the Istanbul Convention) (2011) CETS No 210 (entered into force 1 August 2014) para 3(a).

[27] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 70.

[28] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue' (7 September 2012) UN Doc A/67/357 para 3.

This includes 'sexist hate speech',[29] 'hate speech against women',[30] 'gendered hate speech',[31] 'cybermisogyny'[32] and 'anti-gender discourse'[33]. Often these concepts are not defined. Gender-based hate speech may also constitute or overlap with, *inter alia*, cyber harassment, cyber bullying and harmful gender stereotyping. Meanwhile, concepts such as 'hate speech', 'misogyny' and 'harassment' have different legal connotations. The importance of denoting certain types of sexist speech as hate speech is in part the use of a more substantial pre-existing legal framework in IHRL, compared with 'harassment' and 'misogyny', which are rather undeveloped.[34] Hate speech is also considered the most severe form of harmful speech – undermining core values of human rights – and requires more extensive measures by states. Although the scope of positive obligations *vis-à-vis* hate speech differs depending on the treaty, the categorisation has particular relevance on the Internet. In relation to Internet intermediary liability, a line has been drawn between hate speech and other types of harmful speech, with more far-reaching obligations *vis-à-vis* the former.[35] This does not mean that sexism not reaching the level of hate speech cannot be addressed through other means in IHRL, for instance, through provisions on harmful gender stereotyping.[36]

[29] CEDAW, 'Concluding Observations on the Sixth Periodic Report of Switzerland' (31 October 2022) UN Doc CEDAW/C/CHE/CO/6 para 38(f).

[30] CEDAW, 'Concluding Observations on the Seventh Periodic Report of Finland' (n 22) para 14.

[31] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 68.

[32] UN HRC, 'Right to Privacy: Report of the Special Rapporteur on the Right to Privacy' (16 October 2019) UN Doc A/HRC/40/63 para 73.

[33] CEDAW, 'Concluding Observations on the Combined Seventh and Eighth Periodic Reports of Japan' (10 March 2016) UN Doc CEDAW/C/JPN/CO/7-8 para 20(d).

[34] While harassment is regulated in a limited number of treaties, for example, in art 40 of the Istanbul Convention (n 26), misogyny is solely mentioned sporadically in soft law.

[35] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) para 12; *Delfi v Estonia* (2014) 58 EHRR 29 para 115; UNCHR, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (9 October 2019) UN Doc A/74/486.

[36] This includes art 5 of CEDAW (n 7) and art 12 (1) of the Istanbul Convention (n 26).

3. *Defining the elements of gender-based hate speech*

In terms of delineating the elements of hate speech more broadly, certain common features can be found in treaties, case law and soft law. The Rabat Plan of Action is often used as a reference in the UN context.[37] Primarily, hate speech involves speech that advocates, promotes or incites hatred, discrimination or violence.[38] It may also encompass pejorative speech or holding a group up to ridicule.[39] There are accordingly several categories of hate speech that warrant different types of obligations. Nevertheless, sexist and stereotyping speech more broadly is distinct from hate speech, unless it involves the requisite intent and harm. Whereas it is thus clear that gender-based hate speech encompasses, *inter alia*, incitement to violence and discrimination against women, problems may arise in the assessment of such criteria. Would comments objectifying women, incel communities and websites detailing fantasies of harming women rise to this level? The quandary of what constitutes, for instance, incitement to discrimination – as opposed to protected vulgar and offensive speech – is similar regardless of the type of hate speech. However, as argued by the CoE, sexist hate speech is often seen as acceptable and harmless, ingrained in social institutions and present in everyday communication.[40] It may thus not be viewed – by adjudicators and algorithms alike – as sufficiently severe as to constitute incitement to violence or discrimination.

Furthermore, there is an irregular use in UN sources of 'women', 'sex' and/or 'gender' as the protected identity characteristic *vis-à-vis* hate speech. The CEDAW Committee, for instance, fluctuates between the three concepts.[41] This is in part reflective of a broader development in IHRL, moving from a narrow focus on women and biological differences between the sexes, to acknowledging the impact also of social norms and

[37] *See*, for example, UNGA, 'Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 70.

[38] Art 20(2) ICCPR (n 4); art 4 ICERD (n 4); art 2 Additional Protocol to the Budapest Convention (n 4).

[39] Rabat Plan of Action (n 20) para 12; *Vejdeland and Others v Sweden* (2014) 58 EHRR 15 para 55.

[40] CoE, 'Seminar Combating Sexist Hate Speech: Report' (10-12 February) EYC Strasbourg 6.

[41] See, for example, CEDAW, 'Concluding Observations on the Seventh Periodic Report of Slovakia' (31 May 2023) UN Doc CEDAW/C/SVK/CO/7 paras 20-21.

gender roles on a person's access to human rights. For example, the CEDAW Committee in General Recommendation No 35 replaced the concept of 'violence against women' with 'gender-based violence against women', to better highlight its structural causes and consequences.[42] With the focus on 'gender' has come an increased use of gender-neutral language and recognition also of the rights of transgender and intersex persons.[43] The acknowledgement of 'gender' as separate but interconnected to 'sex' has, however, generated controversy. This was evident, for example, during treaty negotiations of the Rome Statute and the Istanbul Convention, where certain states pushed for the sole use of the term 'sex', or 'sex' and 'gender' as virtually synonymous.[44] However, even with an increased focus on 'gender', it may still be relevant to address the particular situation of women in certain instances. In fact, the CEDAW Committee has noted that a gender-neutral approach may undermine the adoption of effective measures, for example, to combat violence against women, by failing to acknowledge its causes and consequences.[45] Consequently, even when 'gender' is used as the operative word in UN sources, the focus in practice often lies on women. In view of this approach, 'gender-based hate speech' would be a timely concept, while a focus on women may be relevant, given that mainly women are the objects of such forms of speech. This does not mean that 'women' should be understood solely in the cisgender sense nor that the vulnerability of other groups should not also be recognised.

4. *Online hate speech*

Beyond defining these elements, an additional step is to consider whether gender-based hate speech online should be treated differently

[42] CEDAW, 'General Recommendation No. 35 on Gender-Based Violence against Women' (n 26) para. 9.

[43] *See*, for example, CEDAW, 'Concluding Observations on the Ninth Periodic Report of China' (31 May 2023) UN Doc CEDAW/C/CHN/CO/9 paras 55-56.

[44] Beate Rudolf et al, *The UN Convention on the Elimination of All Forms of Discrimination Against Women and Its Optional Protocol: Commentary* (OUP 2023) 26.

[45] CEDAW, 'Concluding Observations of the Committee on the Elimination of Discrimination against Women on Norway' (9 March 2012) UN Doc CEDAW/C/NOR/CO/8 para 9.

than offline speech. While treaties affecting rights in cyberspace are increasingly adopted – again primarily at the European level –[46] the general approach in IHRL is that the existing legal framework is equally applicable online.[47] General treaty provisions are thus being applied to the Internet, including on hate speech and gender stereotyping.[48] This does not mean that contextual differences are not recognised. Context is, for instance, relevant in relation to harm. Gender-based hate speech, whether online or offline, undermines gender equality and access to human rights. Meanwhile, *online* gender-based hate speech has particular consequences in that it limits women's access to an important public sphere, which has been used as an argument for strengthening content regulation.[49] The rapid and global reach of information – also difficult to permanently remove – is considered to augment both individual and group-based harm.[50] Gender stereotypes may gain credence through a broader audience and processes of radicalization.

Furthermore, the consequence for the effective operation of the Internet is an underlying concern when considering restrictions of speech, for example, in the balancing of rights or interests. This is a factor in both adjudication and in the development of Internet regulation. The Internet is frequently heralded by human rights bodies and courts as an essential aspect of the freedom of expression – enhancing democratic participation and individual autonomy – and a means to ensure a range of human

[46] See, for example, Convention on Cybercrime of the Council of Europe (Budapest Convention) (ETS No 185) 23 November 2001. Meanwhile, the development of a UN cybercrime treaty, set in motion through UNGA, 'Resolution 74/247 adopted by the General Assembly on 27 December 2019: Countering the Use of Information and Communications Technologies for Criminal Purposes (20 January 2020) UN Doc A/RES/74/247, is at a standstill.

[47] See, for example, UN HRC, 'The Promotion, Protection and Enjoyment of Human Rights on the Internet' (27 June 2016) UN Doc A/HRC/32/L.20 para 1

[48] CERD, 'Concluding Observations on Poland' (29 August 2019) UN Doc CERD/C/POL/CO/22-24 para 16 (b); CEDAW, 'Concluding Observations on the Seventh Periodic Report of Finland' (n 22) para 14.

[49] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 91

[50] See, for instance, European Commission (n 11) annex para 17; *Delfi v Estonia* (2014) 58 EHRR 29 para 133.

rights.[51] The practical consequences of regulating certain types of speech in the online sphere – including excessive censorship – are thus relevant and may override harm not considered sufficiently grave.

More specifically, the assessment of hate speech is *per se* contextual, considering, *inter alia*, the intention of the speaker and the effect on the audience.[52] The number and demographics of visitors on a website are relevant in assessing harm. Meanwhile, the tone and purpose of the website and previous comments are factors for evaluating intent. Neither human moderators nor algorithms used by social media companies to detect hate speech may be sensitive to such elements, despite increasingly advanced extra-linguistic assessments by AI. Often content is viewed in isolation and cultural differences and nuances may be over-looked. While the use of AI may avoid subjective and gendered assumptions of human moderators, gender bias is also a noted problem in algorithms.[53] Studies, for example, show that racist and homophobic tweets are more likely to be categorised as hate speech than sexist tweets, which are often classified as merely offensive.[54] The CEDAW Committee has even noted the necessity of eliminating gender bias in AI in order to effectively detect and regulate gender stereotypes, including hateful speech.[55]

The main difference between online/offline hate speech, however, lies in the content of state obligations. As noted, the scope of obligations *vis-à-vis* hate speech varies greatly depending on the treaty, with certain treaties allowing states to restrict expressions and others including obligations to prohibit speech. This is reflective of broad ideological differences on the legitimacy of restricting the freedom of expression and appropriate means to prevent harmful speech. In terms of UN treaties, obligations

[51] UNCHR, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye' (22 May 2015) UN Doc A/HRC/29/32 para 11.

[52] Rabat Plan of Action (n 20) para 29.

[53] De Streel et al for the European Parliament, Directorate-General for Internal Policies of the Union, 'Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform' (2020) 59 <https://op.europa.eu/en/publication-detail/-/publication/cd388309-cc89-11ea-adf7-01aa75ed71a1>.

[54] T Davidson et al, 'Automated Hate Speech Detection and the Problem of Offensive Language' (2017) 11 Proceedings of the International AAAI Conference on Web and Social Media 512.

[55] CEDAW, 'Concluding Observations on the Ninth Periodic Report of Spain' (n 23) para 23(c).

to prohibit hate speech can explicitly be found in the ICERD[56] and the ICCPR,[57] where 'women' have been addressed from an intersectional perspective. Obligations in CEDAW include measures to combat harmful gender stereotypes which, according to recent concluding observations, include the criminalisation of sexist hate speech.[58] Meanwhile, it is often emphasized by UN bodies that criminalisation should only be employed in the most egregious cases, for example, when speech presents a clear and imminent danger.[59] Other measures, such as education, may be more suitable.[60] Since the underlying causes of sexism include harmful social norms and gender stereotyping, this broad range of measures ensures a holistic approach.

These obligations remain the same online. The UN Special Rapporteur on Freedom of Opinion and Expression has in fact warned against the use of stricter penalties for individuals or excessively intrusive technology as a means of restricting online hate speech.[61] Where obligations may differ is in the regulation of intermediary liability and content moderation. According to EU law and under the ECHR, state obligations to ensure intermediary liability arise in instances of hate speech.[62] This type of content is considered 'illegal', in contrast to 'harmful' content, which may involve stereotyping that does not reach the threshold of hate speech. Modes of moderation range from the monitoring of websites and immediate removal of illegal content – be it by human moderators or AI – to notice-and-take-down systems.[63] There is no such cohesive approach in the UN, but rather brief recommendations by UN treaty bodies and

[56] Art 4(a) ICERD (n 4).

[57] Art 20(2) ICCPR (n 4).

[58] See, for example, CEDAW, 'Concluding Observations on the Seventh Periodic Report of Finland' (n 22) para 20(a); CEDAW, 'Concluding Observations on the Seventh Periodic Report of Slovakia' (n 41) para 21(b) and (c); CEDAW, 'Concluding Observations on the Ninth Periodic Report of Iceland' (31 May 2023) UN Doc CEDAW/C/ISL/CO/9 para 22.

[59] CERD, 'General Recommendation No 35: Combating Racist Hate Speech' (n 14) para 12; UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) para 70.

[60] CERD, 'General Recommendation No 35: Combating Racist Hate Speech' (n 14) para 30

[61] UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 35) para 29.

[62] See n 35.

[63] ibid.

special rapporteurs. This includes suggestions by the CEDAW Committee for states to strengthen self-regulation of intermediaries[64] and to hold '…social media companies accountable for discriminatory user-generated content'.[65] The Committee has also called on states to implement the proposed EU AI Act as a means to ensure certain standards in controlling online content.[66] In terms of direct obligations for intermediaries under IHRL, this does not go beyond obligations to respect, as developed in the UN guiding principles on business and human rights.[67] In practice, the terms of service of major social media platforms in fact prohibit sexist hate speech.[68] Issues are rather abstract definitions of gender-based hate speech – that may not align with international standards – and content moderation that fails to strike a balance between effectiveness and respect for the freedom of expression.

5. *Concluding remarks*

In moving forward, broader attention should be given to gender-based hate speech in IHRL, in particular by UN bodies. Although the discussed UN sources are solely soft law, they are indicative of an evolutive treaty interpretation open to the regulation of gender-based hate speech. However, this development is still sporadic and needs a more consistent approach. As a first step, it involves recognizing gender-based hate speech as a form of gender discrimination and gender-based violence, affecting a range of human rights. Secondly, it requires the development of a definition of gender-based hate speech and clarifying state

[64] CEDAW, 'Concluding Observations on the Seventh Periodic Report of Finland' (n 22) para 15 (c).

[65] CEDAW, 'Concluding Observations on the Ninth Periodic Report of Iceland' (n 58) para 22.

[66] CEDAW, 'Concluding Observations on the Ninth Periodic Report of Germany' (n 23) para 18(b); CEDAW, 'Concluding Observations on the Ninth Periodic Report of Spain' (n 23) para 23(c).

[67] UNHRC, 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and other Business Enterprises, John Ruggie: UN Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (21 March 2011) UN Doc A/HRC/17/31.

[68] See in UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (n 16) paras 77- 78.

obligations, including *vis-à-vis* intermediaries. Given the nature of the Internet, online hate speech is a global issue, and a global solution in the form of a workable legal framework for states and intermediaries alike is thus preferable. Nevertheless, as argued, it is challenging to find common ground at the international level in the regulation of hate speech *per se*, no less sexist speech on the Internet. The development of a cohesive approach may thus fare better at the regional level.