 doi:10.20944/preprints202203.0258.v1

Article

Artificial Intelligence and Online Hate Speech Moderation: A Risky Match?

Natalie Alkiviadou [1]

[1] Justitia; natalie@justitia-int.org

Abstract: Artificial Intelligence is increasingly being used by social media platforms to tackle online hate speech. The sheer quantity of content, the speed at which is it developed and the enhanced pressure companies are facing by States to remove hate speech quickly from their platforms have led to a tricky situation. This commentary argues that automated mechanisms, which may have biased datasets and be unable to pick up on the nuances of language, should not be left unattended with hate speech as this can lead to issues of violating freedom of expression and the right to non-discrimination.

Keywords: hate speech; artificial intelligence; social media platforms; content moderation; freedom of expression; non-discrimination

1. Introduction

 doi:10.20944/preprints202203.0258.v1

In the mid-nineties, John Perry Barlow (1996) stated that the Internet would permit 'a world where anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity.' Although the Internet has been a massive information and communication revolution, Barlow's comment is not reflective of the *status quo* today. Social Media Platforms (SMPs), which are the primary traffic vehicle for communication and information, are coming under increasing pressure by States to remove content. There are approximately 2.9 billion monthly active users on Facebook [1], 2 billion on YouTube,[2] 1 billion on Instagram[3], 1 billion on TikTok[4], 430 million on Reddit[5] and 396.5 million on Twitter.[6] SMPs facilitate borderless communication, allow for, *inter alia*, political, ideological, cultural and artistic expression, give a voice to traditionally silenced groups, provide an alternative to mainstream avenues which may be State censored, permit an inflow of daily news and raise awareness on human rights violations. However, as noted by Mchangama *et al* (2021), individuals move onto such platforms, new visibility has been given to phenomena such as hate and abuse. Timofeeva (2003) notes that whilst hate existed before the Internet and social networks, the emergence of the Internet and the subsequent creation of social networks have added new dimensions to the already complex topic of hate speech. The use of SMPs has also been directly linked to horrific events such as the genocide in Myanmar. Cognizant of the dangers of violent speech with an imminent risk of violence, I argue that care must be taken when embracing the common rhetoric that hate speech is prevalent across social media, since empirical work has demonstrated the opposite. For example, a study conducted by Siegel et al (2019) assessing whether Trump's 2016 election campaign (and the six-month period following it) led to a rise in hate speech on Twitter. through an analysis of a sample of 1.2 billion tweets, found that between 0.001% and 0.003% of the tweets contained hate speech on any given day, 'a tiny fraction of both political language and general content produced by American Twitter users.' Either way, States are increasing pressure for platform regulation of hate speech which, as argued in this paper, has led to the dilution of the right to free speech and has directly contributed to silencing minority groups. The manner in which this

[1] Backlinko, Facebook Demographic Statistics 2022 <https://backlinko.com/facebook-users> [Accessed 30 December 2021]

[2] Backlinko, 'How Many People Use YouTube in 2022' <https://backlinko.com/youtube-users> [Accessed 3 January 2022]

[3] Backlinko, 'Instagram Demographic Statistics 2022' < https://backlinko.com/instagram-users> [Accessed 3 January 2022]

[4] Backlinko, 'TikTok User Statistics 2022' https://backlinko.com/tiktok-users [Accessed 3 January 2022]

[5] Backlinko, 'Reddit User and Growth Stats – updated October 2021' <https://backlinko.com/reddit-users

[6] https://backlinko.com/twitter-users> [Accessed 3 January 2022]

new reality is being tackled by States and institutions, such as the European Union, is of concern. For example, in 2017, Germany passed the 'Network Enforcement Act' (NetzDG), which seeks to counter illegal online speech such as insult, incitement and religious defamation. It obliges social media platforms with a minimum of 2 million users to remove illegal content – including hate speech and religious offence – within 24 hours, or risk steep fines of up to 50 million euros. This has become a prototype for Internet governance in authoritarian States. In two reports by Mchangama *et al*, one in 2019 and one in 2020, Justitia recorded the adoption of a NetzDG model in over 20 countries, with several of these countries being described by the Freedom House as 'not free' or 'partly free.' All countries require online platforms to remove vague categories of content that include 'false information,' 'blasphemy/religious insult' and 'hate speech.' Mchangama & Alkiviadou note that worryingly, 'few of these countries have in place the basic rule of law and free speech protections built into the German precedent.' The same template is currently being followed at a European Union (EU) level, with the Digital Services Act (DSA) currently being discussed at the time of writing (parliamentary plenary voting January 2022). The DSA includes, amongst others, an enhanced burden on platforms to remove 'illegal content' at the risk of fines. [7]

As a response to enhanced regulatory requirements, due to the risk of steep fines, platforms are prone to take the 'better safe than sorry approach' and regulate vivaciously. However, as noted by Llanso (2020) online communication on such platforms occurs on a 'massive scale,' rendering it impossible for human moderators to review all content before it is made available. The sheer quantity of online content also makes the job of reviewing, even reported content, a difficult task. As a response both to the need to dodge State fines and the technical aspect of content scale and quantity, SMPs have increasingly relied on Artificial Intelligence (AI) in the form of automated mechanisms that proactively or reactively tackle problematic content, including hate speech. In brief, as highlighted by Oliva *et al* (2021), AI provides SMPs with 'tools to police an enormous and ever-increasing flow of information – which comes in handy in the implementation of content policies.' Whilst this is necessary in areas involving, for example, child abuse and the non-consensual promotion of intimate acts amongst adults, the use of AI to regulate more contentious 'grey' areas of speech, such as hate

[7] The Digital Services Act < https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#documents> [Accessed 27 December 2021]

 doi:10.20944/preprints202203.0258.v1

speech, is complex. In light of these developments, this paper looks at the use of AI to regulate hate speech on SMPs, arguing that automated mechanisms, which may have biased datasets and be unable to pick up on the nuances of language, may lead to violations of the freedom of expression and the right to non-discrimination of minority groups, further silencing already marginalized groups. To achieve its objective, the paper commences with an overview of the semantical framework of hate speech. This is followed by a short insight into what is meant by AI in relation to content moderation, an insight into freedom of expression as provided for by institutions, namely, the United Nations and the Council of Europe, and the ramifications which the use of AI has on this freedom. Further, the paper looks at the doctrine of non-discrimination and how it is affected by the use of AI in the field of online moderation of hate speech.

2. Hate Speech: Semantics and Notions

Hate speech does not enjoy a universally accepted formulation, with most States and institutions adopting their own understanding of what hate speech entails[8] without defining it (Alkiviadou 2017). One of the few documents, albeit non-binding, which has sought to elucidate the meaning of hate speech, is the Recommendation of the Council of Europe Committee of Ministers on hate speech.[9] It provides that this term is to be:

'understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerant expression by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.'

Interestingly, the Recommendation incorporates the justification of hatred as well as its spreading, incitement and promotion, allowing for a broad spectrum of intentions to fall within its definition. Hate speech has also been mentioned, but not defined, by the European Court of Human Rights (ECtHR). For example, it found that hate speech entails 'all forms of expression which spread, incite, promote or justify hatred based on

[8] Council of Europe Committee of Experts for the Development of Human Rights Report (2007) Chapter IV, 123, para. 4.

[9] Council of Europe's Committee of Ministers Recommendation 97 (20) on Hate Speech.

 doi:10.20944/preprints202203.0258.v1

intolerance, including religious intolerance.'[10] The inclusion of merely justifying hatred demonstrates the low threshold attached to unacceptable speech. Further, the ECtHR has held that it is not necessary for the speech 'to directly recommend individuals to commit hateful acts'[11] since attacks on persons can be committed by 'insulting, holding up to ridicule or slandering specific groups of the population'[12] and that 'speech used in an irresponsible manner may not be worthy of protection.'[13] In this sense, the ECtHR has drawn the correlation between hate speech and the negative effects it can have on its victims, alleging that even violence-free speech amounting to mere insults has the potential to cause harm sufficient enough to limit free speech.

McGonagle (2013) argued that the fact that the Court has not yet offered a definition of hate speech has been characterized as 'unsatisfactory from the point of judicial interpretation, doctrinal development and general predictability and foreseeability.' It could also be argued that this has contributed to the rather controversial judgements which potentially go against the very essence of Article 10 of the European Convention on Human Rights (ECHR) protecting the freedom of expression.[14]

In addition, the EU's Fundamental Rights Agency has offered two separate formulations of hate speech, the first being that it 'refers to the incitement and encouragement of hatred, discrimination or hostility towards an individual that is motivated by prejudice against that person because of a particular characteristic.'[15] In its 2009 report on homophobia, the FRA held that the term hate speech, as used in that particular section, 'includes a broader spectrum of verbal acts including disrespectful public discourse.'[16] The particularly problematic part of this definition, is the broad reference to disrespectful public discourse, especially since

[10] Gűndűz v Turkey, Application no. 35071/97 (ECHR 4 December 2003) para. 40; Erbakan v Turkey, Application no. 59405/00 (6 July 2006) para. 56.

[11] Vejdeland and Others v Sweden, Application no. 1813/07 (ECHR 9 February 2012) para. 54.

[12] Ibid.

[13] Ibid. para. 55

[14] For more on this look at Jacob Mchangama & Natalie Alkiviadou, 'Hate Speech and the European Court of Human Rights, Whatever Happened to the Right to Offend, Shock or Disturb.' (2021) 21 *Human Rights Law Review* 4

[15] Fundamental Rights Agency, 'Hate Speech and Hate Crimes against LGBT Persons' (2009) 1.

[16] Fundamental Rights Agency, 'Homophobia and Discrimination on Grounds of Sexual Orientation and Gender Identity in the EU Member States: Part II - The Social Situation' (2009) 44.

 doi:10.20944/preprints202203.0258.v1

institutions, such as the ECtHR, extend the freedom of expression to ideas that 'shock, offend or disturb.'[17] This is the formal position of the Court, even though in relation to hate speech cases, as briefly noted above, it has rigorously adopted a very low threshold of what it is willing to accept as permissible speech.

In the framework of academic commentary, a plethora of definitions has been put forth to describe hate speech. In exploring its different formulations, Belavusau (2013) underlined that hate speech is 'deeply rooted in the ideologies of racism, sexism, religious intolerance, xenophobia, and homophobia.' In addition, he argues that pinpointing the grounds from which hate speech may arise is also a tricky task and poses the question of where limits are to be drawn. According to Matsuda *at al* (1993), hate speech which is discussed in the sphere of racism, contains three central elements: namely, that the message is 'of racial inferiority, the message is directed against historically oppressed groups and the message is persecutory, hateful and degrading.' McGonagle (2001) offers a broad interpretation of hate speech in terms of threshold but not in terms of content and target groups, arguing that 'virtually all racist and related declensions of noxious, identity-assailing expression could be brought within the wide embrace of the term.' Smolla (1990) defines it as a 'generic term that has come to embrace the use of speech attacks based on race, ethnicity, religion and sexual orientation or preference.' Although some common elements can be discerned from these approaches to hate speech and the variations therein, Kiska (2012) noted that 'hate speech seems to be whatever people choose it to mean.' What can be discerned from the various extrapolations of hate speech is that, as underlined by Sagle (2009) it 'singles out minorities for abuse and harassment.'

Turning to platforms themselves, while it is beyond the scope of this paper to assess all the guidelines/standards of SMPs, we look at two different approaches, Facebook and Instagram on the one hand (both owned by Facebook Inc) and Reddit on the other. The former[18] formulate their understanding of hate speech on three tiers, the first being violent and dehumanizing speech, the second being statements of inferiority, content, dismissal and other such as repulsion. Tier three includes statements pertaining to segregation and exclusion.

[17] The Observer and The Guardian v The United Kingdom, Application no 13585/88 (ECHR 26 November 1991) para. 59.

[18] Meta Transparency Centre, 'Hate Speech' <https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech> [Accessed 2 January 2022]

 doi:10.20944/preprints202203.0258.v1

The list of protected characteristics is broad such as race, ethnicity, religious affiliation, caste, sexual orientation and serious disease.[19] Reddit[20] takes a more speech protective approach, prohibiting incitement to violence and the promotion of hatred. The protected characteristics include, amongst others, race, colour, religion and pregnancy. It is noteworthy that all major platforms, including the ones above as well as Twitter,[21] YouTube[22] and TikTok[23], all incorporate the grounds of race and religion in the list of protected characteristics.

3. Artificial Intelligence

The use of AI is not only a response to issues of quantity. but also to increasing State pressure on social media platforms to remove hate speech quickly and efficiently. Examples of such pressure include, *inter alia*, the German Network Enforcement Act which provides for fines of up to 50 million euros in certain situations. As noted by Mchangama and Fiss, this template has been replicated around the globe, including in authoritarian states. SMPs also face pressure from other entities such as advertisers and their users. To be able to comply with such standards (and avoid hefty fines), companies use AI, alone or in conjunction with human moderation, to remove allegedly hateful content. As noted by Oliva (2020) such circumstances have prompted companies to 'act proactively in order to avoid liability…in an attempt to protect their business models.' Gorwa *et al* (2002) highlight that as "government pressure on major technology companies build, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation.'

The 'work from home' Covid situation has also led to enhanced reliance on AI (which came with errors in moderation). As announced by YouTube:

---

[19] Meta Transparency Centre, 'Hate Speech' <https://transparency.fb.com/policies/community-standards/hate-speech/#policy-details> [Accessed 2 January 2022]

[20] Reddit, Promoting Hate Based on Identity or Vulnerability <https://www.reddithelp.com/hc/en-us/articles/360045715951> [Accessed 2 January 2022]

[21] Twitter, Hateful Conduct Policy <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> [Accessed 2 January 2022]

[22] YouTube, Hate Speech Policy <https://support.google.com/youtube/answer/2801939?hl=en> [Accessed 2 January 2022]

[23] TikTok, Community Guidelines <https://www.tiktok.com/community-guidelines?lang=en#38> [Accessed 2 January 2022]

 doi:10.20944/preprints202203.0258.v1

'In response to COVID-19, we've taken steps to protect our extended workforce and reduce in-office staffing. As a result, we are temporarily relying more on technology to help with some of the work normally done by human reviewers, which means we are removing more content that may not be violative of our policies. This impacts some of the metrics in this report and will likely continue to impact metrics moving forward.'[24]

To exemplify the use of AI by social media platforms, some recent figures are given as follows. In its latest Community Standards Enforcement Report (for the third quarter of 2021), Facebook said that its proactive rate of removal for hate speech was 97.6%. During the reporting period, it removed 22.3 million pieces of hate speech. As noted in a post on the Transparency Centre, 'our technology proactively detects and removes the vast majority of violating content before anyone reports it.'[25]

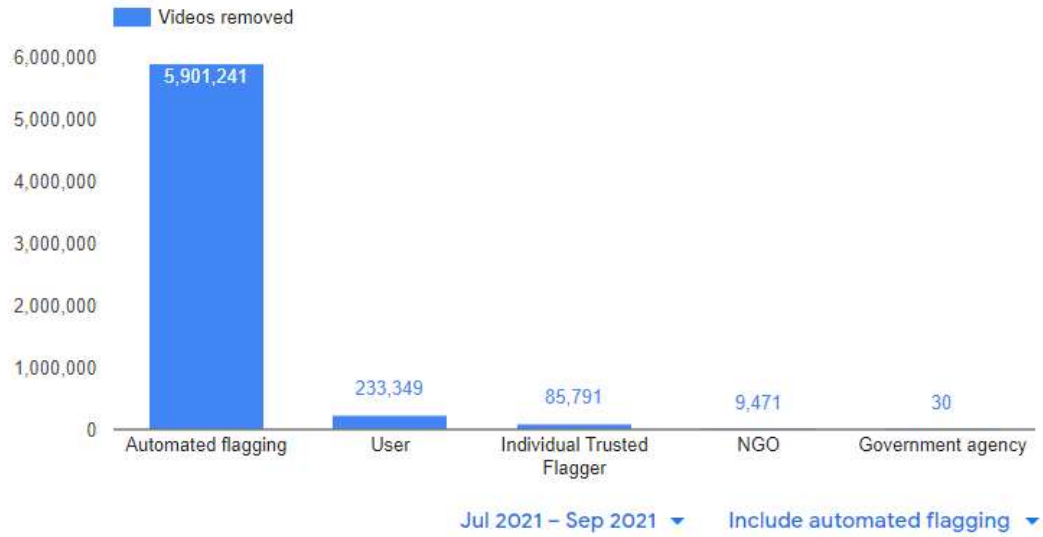
In its latest enforcement report[26] (Q3 of 2021), YouTube put forth the illustration below, demonstrating the percentage of human flagging and automated flagging across the board of removable content (not just hate speech):

[24] YouTube Community Guidelines Enforcement Report available at:
< https://transparencyreport.google.com/youtube-policy/removals?hl=en> [Accessed 27 December 2021]

[25] Meta, Transparency Centre, 'How Technology Detects Violations' <https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/> [Accessed 3 January 2022]

[26] https://transparencyreport.google.com/youtube-policy/removals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 March 2022 doi:10.20944/preprints202203.0258.v1



Oliva et al (2021) argue that the algorithms developed to achieve this automation are habitually customized for content type, such as pictures, videos, audio and text. As found by Duarte and Llanso (2017), current technologies detect harmful text through the use of Natural Language Processing and sentiment analysis and, even though its evolution is marked, the accuracy lies between 70-80 per cent. They argue that AI has 'limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker.' As such, these technologies 'still fail to understand context, thereby posing risks to users' free speech, access to information and equality.' Moreover, Oliva *et al* (2021) argue that going from policy to code may lead to changes in meaning, since machine language is more limited than its human counterpart.

With the power SMPs hold over today's marketplace of expression and information, and the growing need and trend to use AI to deal with external pressures for removal, as well as the quantity of material, Cowls *et al* (2020) argue that there is an urgent need to ensure that content moderation occurs in a manner where human rights are protected and public discourse ensured. In brief, large social media platforms face substantial pressure regarding their content moderation systems from public authorities, advertisers, and their users, and are increasingly relying on automated and proactive methods for detecting and evaluating user content. While it is comprehensible from a practical perspective, Llanso (2020) notes that the widespread use of proactive

 doi:10.20944/preprints202203.0258.v1

detection and automated evaluation in content moderation represents a 'significant shift in how we have conceived of speech regulation— and protection—under the international human rights framework.'

In light of the above, and with a focus on the contentious area of 'hate speech', this paper will examine the human rights risks that arise or may arise from the current *status quo*, namely, the increased reliance on AI by private profit-making companies, with a particular focus on freedom of expression and non-discrimination.

4. AI, Hate Speech and the Challenges to the Freedom of Expression

(i)The Freedom of Expression

Cato believed the freedom of expression is 'the great bulwark of liberty.'[27] Crossman (2012) described it as a 'cornerstone upon which the very existence of a democratic society rests.' On a United Nations level, Article 19 of the Universal Declaration of Human Rights (UDHR) provides that:

'Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.'

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for the freedom of opinion and expression. More particularly, it states that:

1. Everyone shall have the right to hold opinions without interference;
2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice;
3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

(a) For respect of the rights or reputations of others;

[27] John Trenchard & Thomas Gordon, '*Cato's letters*' (ed. Ronald Hamowy) (Liberty Fund, Carmel 1995) (Original: 1755).

 doi:10.20944/preprints202203.0258.v1

(b) For the protection of national security or of public order (ordre public), or of public health or morals.

Notwithstanding the fundamental position held by the freedom of expression in the international legal framework, Farrior (1996) underlined that 'this freedom does not enjoy such a position of primacy among rights that it trumps equality rights.' However, as noted by General Comment 34 of the Human Rights Committee (HRC) which monitors the implementation of the ICCPR, Article 19 of the ICCPR embraces 'even expression that may be regarded as deeply offensive.'[28]

As well as the limitations found in Article 19(3), Article 20 of the ICCPR contains a specific prohibition on two types of expression providing that:

1. Any propaganda for war shall be prohibited by law.
2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.[29]

On a Council of Europe level, Article 10 of the ECHR protects the right to freedom of expression. It states that:

1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.
2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder

[28] HRC General Comment 34, 'Article 19 – Freedom of Opinion and Expression' (2011) CCPR/C/GC/34, para. 11.

[29] This is also found in Article 13(3) of the CMW which states that the exercise of the freedom of expression 'may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For the purpose of preventing any propaganda for war; (b) For the purpose of preventing any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.

 doi:10.20944/preprints202203.0258.v1

or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

The article stipulates that 'everyone' has the right to freedom of expression, including, as recognized by Lester *et al* (2009), both natural and legal persons. As with Article 19 of the ICCPR, Article 10 recognizes that the freedom of expression, which includes the freedom to hold opinions and to receive and impart information and ideas, carries with it duties and responsibilities. As such, Article 10, unlike other articles of the ECHR, incorporates a further qualification in the form of duties and responsibilities when exercizing this right.

According to the jurisprudence of the ECtHR, the rights set out in Article 10(1) are to be interpreted and applied in a broad manner, given that, according to Williams & Cooper (1999), 'there is no room in general for an argument that Article 10 extends only to true information: opinions, speculations and criticism are all covered.' In *Handyside v The United Kingdom*, the Court held that:

'the freedom of expression constitutes one of the essential foundations of [a democratic society], one of the basic conditions for its progress and for the development of every man…It is applicable not only to information or ideas that are favourably received or regarded as inoffensive or as matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no democratic society.'[30]

The Court has repeatedly underlined the central position of the freedom of expression in a democratic and pluralist society. This is notwithstanding the fact that, in hate speech cases, for example, the Court has taken a rather restrictive approach to freedom of expression. [31]

[30] Handyside v The United Kingdom, Application no. 5493/72 (ECHR 7 December 1976) para. 49.

[31] For more on this look at Jacob Mchangama & Natalie Alkiviadou, 'Hate Speech and the European Court of Human Rights, Whatever Happened to the Right to Offend, Shock or Disturb.' (2021) 21 *Human Rights Law Review* 4

 doi:10.20944/preprints202203.0258.v1

(ii) AI and Challenges to the Freedom of Expression

As noted by Oliva (2020), relying on AI, even without human supervision, is a necessity when it comes to content that could never be ethically or legally justifiable, such as child abuse. However, the issue becomes complicated when it comes to contested areas of speech, such as hate speech, for which there is no universality in terms of ethical and legal positioning as to what it is and when (if at all) it should be removed. In the ambit of such speech, Llanso (2020) underlines that the use of AI raises "significant questions about the influence of AI on our information environment and, ultimately, on our rights to freedom of expression and access to information". As underlined by Llanso *et al* (2020) it poses 'distinct challenges for freedom of expression and access to information online.' As highlighted in a Council of Europe report, the use of AI for hate speech regulation directly impacts the freedom of expression, which raises concerns *vis-à-vis* the rule of law and, in particular, notions of legality, legitimacy and proportionality.[32] The Council of Europe noted that the enhanced use of AI for content moderation may result in over-blocking and consequently place the freedom of expression at risk.[33] Gorwa et al (2020) argue that the increased use of AI threatens to exacerbate already existing opacity of content moderation, further perplex the issue of justice online and "re-obscure the fundamentally political nature of speech decisions being executed at scale". Moreover, regardless of the technical specifications of a particular mechanism, proactive identification (and removal) of hate speech is a prior restraint of speech with all the legal issues that this entails. Specifically, Llanso *et al* (2020) argue that there is a 'strong presumption against the validity of prior censorship in international human rights law.' Former UN Special Rapporteur on the Freedom of Opinion and Expression, David Kaye, expressed his concern over the use of automated tools in terms of the potential of over-blocking, and argued that calls to expand upload filtering to terrorist related and other areas of content 'threaten to establish comprehensive and disproportionate regimes of pre-publication censorship.'[34]

5. AI, Hate Speech and the Challenges to Non-Discrimination

[32] 'Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (2017) Council of Europe, DGI(2017) 12, 18

[33] Ibid. 21

[34] Mandate of the Special Rapoprteur on the Promotion and Protection of the Right to Freedom of Expression (13 June 2018) <https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf> [Accessed 10 January 2022]

 doi:10.20944/preprints202203.0258.v1

(i) The Doctrine of Non-Discrimination

Shestack 1984) argued that the doctrines of equality and non-discrimination 'are central to the human rights movement.' The doctrine non-discrimination in International Human Rights Law (IHRL) can be traced back to the Charter of the United Nations which holds that the purposes of the UN are, amongst others, to 'achieve international cooperation … in promoting and encouraging respect for human rights and for fundamental freedoms for all without distinction as to race, sex, language, or religion.'[35]

Article 1 of the UDHR states that 'all human beings are born free and equal in dignity and rights.' Article 2 provides that 'everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.' Article 7 of the UDHR constitutes the first effort to incorporate incitement to discrimination, stipulating that '…all are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.' The non-discrimination clauses of the ICCPR and the International Covenant on Economic, Social and Cultural Rights (ICESCR), which are found in Article 2(2) of both covenants, adopt the same approach as Article 2 of the UDHR. The ICCPR incorporates a specialized non-discrimination clause in the form of Article 26, therein, which states that:

'all persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.'

General Comment 18 of the HRC, notes that non-discrimination should be understood:

'to imply any distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other

[35] Article 1.3 Charter of the United Nations 1945.

 doi:10.20944/preprints202203.0258.v1

status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms.[36]

In terms of racial discrimination, Article 1(1) of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) constitutes

'any distinction, exclusion, restriction or preference based on race, colour, descent or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.'

General Recommendation 35, of the Committee on the Elimination of All Forms of Racial Discrimination (CRD), refers to groups of people who may fall within the ambit of Article 1, namely 'indigenous people, descent-based groups, and immigrants or non-citizens, including migrant domestic workers, refugees and asylum seekers, as well as speech directed against women members of these and other vulnerable groups.'[37]

In relation to religious discrimination, in the CERD's General Recommendation 32 on Special Measures, the Committee underlined that the existing grounds of discrimination under the Convention are:

'extended in practice by the notion of intersectionality whereby the Committee addresses situations of double or multiple discrimination—such as discrimination on grounds of gender or religion—when discrimination on such a ground appears to exist in combination with a ground or grounds listed in Article 1 of the Convention.'[38]

[36] HRC General Comment 18: 'Non-Discrimination' (1994) HRI/GEN/1/Rev.1 at 26, para. 12.

[37] CERD General Recommendation 35: 'Combatting Racist Hate Speech' (2013) CERD/C/GC/35, para. 6, HRC General Comment 34: 'Article 19: Freedoms of Opinion and Expression' (2011) CCPR/C/GC/34, para. 6.

[38] CERD General Recommendation 32: 'The Meaning and Scope of Special Measures in the International Convention on the Elimination of Racial Discrimination' (2009) CERD/C/GC/32, para. 7.

 doi:10.20944/preprints202203.0258.v1

Intersectionality was referred to in two CERD cases, namely, *P.S.N. v Denmark* and *A.W.R.A.P. v Denmark*, which were declared inadmissible given that the respective claims were, according to the CERD, based on religious discrimination only and, as noted, 'Islam is not a religion practised solely by a particular group.'[39] The CERD summed up its position in relation to this issue by holding that 'religious questions are of relevance to the Committee when they are linked with issues of ethnicity and racial discrimination.'[40] Thus, in light of the principle of intersectionality, Islamophobic and/or other religiously motivated hate speech and activities, can be condemned and prohibited under the ICERD only if interlinked with one of the grounds expressly stipulated in Article 1, these being race, colour, descent, or national or ethnic origin.

On a Council of Europe level, non-discrimination is provided by Article 14 of the ECHR and its Protocol 12. The ECtHR has defined discrimination as 'treating differently, without an objective and reasonable justification, persons in relevantly similar situations.'[41] No objective and reasonable justification means that 'the distinction in issue does not pursue a legitimate aim or that there is not a reasonable relationship of proportionality between the means employed and the aim sought to be realized.'[42] Moreover, in order for an application to be successful under this article, the 'discriminatory intent or effect'[43] of the object or act or measure complained of must be established.

Article 14 of the Convention provides that:

'The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any grounds such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.'

[39] P.S.N v Denmark (2007) Communication no. 36/2006, CERD/C/71/D/36/2006, para. 6.3.

[40] CERD Report 66th session and 67th session (2005) A/60/18, para. 246.

[41] Willis v The United Kingdom, Application no. 36042/97 (ECHR 11 September 2002) para. 48.

[42] See, *inter alia*, Andrejeva v Latvia, Application no. 55707/00 (ECHR 18 February 2009) para. 81; Sejdić and Finci v Bosnia and Herzegovina, Application nos. 27996/06 and 34836/06 (22 December 2009) para. 42.

[43] Aksu v Turkey, Application nos. 4149/04 and 41029/04 (ECHR 15 March 2012) para. 45.

 doi:10.20944/preprints202203.0258.v1

Article 1 of Protocol 12 holds that:

'The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.'

However, the Protocol has only been adopted by eighteen of the forty-seven States Parties, and, thus, its actual impact remains limited. Nevertheless, with regard to the interrelationship between Protocol 12 and Article 14, the Court has underlined that they should be understood in a similar way, given that 'notwithstanding the difference in scope between those provisions, the meaning of this term in Article 1 of Protocol No. 12 was intended to be identical to that in Article 14.'[44]

From the time of the European Commission of Human Rights (EComHR), the particularly serious nature of racial discrimination has been underlined. In *3 East African Asians v The United Kingdom*, the Commission noted that 'discrimination based on race could, in certain circumstances, of itself amount to degrading treatment within the meaning of Article 3.'[45] This viewpoint was also adopted by the Court in *Timishev v Russia,* in which it was held that racial discrimination is a 'particularly invidious kind of discrimination'[46] with 'perilous consequences.'[47] This position was endorsed in *Aksu v Turkey*[48]*,* which dealt with Romaphobia. In *Sejdić and Finci v Bosnia and Herzegovina*, the Court held that, in the context where discrimination is based on race or ethnicity, the notion of differential treatment without an objective and reasonable justification, as referred to earlier, 'must be interpreted as strictly as possible.'[49]

(ii) AI and Challenges to Non-Discrimination

[44] Sejdić and Finci v Bosnia and Herzegovina, Application nos. 27996/06 and 34836/06 (ECHR 22 December 2009) para. 55.

[45] 3 East African Asians (British Protected Persons) v The United Kingdom, Application nos. 4715/70, 4783/71 and 4827/71 (EComHR 6 March 1978) para. 2

[46] Timishev v Russia, Application nos. 55762/00 and 55974/00 (ECHR 13 March 2006) para. 56.

[47] Ibid.

[48] Aksu v Turkey, Application nos. 414904 and 41029/04 (ECHR 15 March 2012) para. 43.

[49] Sejdić and Finci v Bosnia and Herzegovina, Application nos. 27996/06 and 34836/06 (ECHR 22 December 2009) para. 44.

 doi:10.20944/preprints202203.0258.v1

Oliva (2020) argues that AI may result from the biased enforcement of Terms of Service. This can be due to a lack of data and/or biased training datasets, leading to the potential silencing of members of minority communities.[50] This can lead to violations of the freedom of expression and the right to non-discrimination. In its report 'Mixed Messages: The Limits of Automated Social Content Analysis', the Centre for Democracy and Technology revealed that automated mechanisms may disproportionately impact the speech of marginalized groups.[51]Although technologies, such as natural language processing and sentiment analysis have been developed to detect harmful text without having to rely on specific words/phrases, research has shown that, as phrased by Oliva et al (2021), they are "still far from being able to grasp context or to detect the intent or motivation of the speaker". As noted by Oliva (2020), hash-matching, which is widely used to identify child sexual abuse content, is not easily transposed to other frameworks such as extremist content which 'typically requires assessment of context.'

Relevant to this, is Keller (2018) who noted that the decision of platforms to remove Islamic extremist content will 'systematically and unfairly burden innocent internet users who happen to be speaking Arabic, discussing Middle Eastern politics or talking about Islam.' She refers to the removal of a prayer (in Arabic) posted on Facebook, because it allegedly violated its Community Standards. The prayer read "God, before the end of this holy day forgive our sins, bless us and our loved ones in this life and the afterlife with your mercy almighty."

Further, as found by Oliva *et al* (2021), such technologies are just not cut out to pick up on the language used, for example, by the LGBTQ community whose 'mock impoliteness" and use of terms such as 'dyke', 'fag' and 'tranny' occurs as a form of reclamation of power and a means of preparing members of this community to 'cope with hostility.' *Oliva et al* (2021) give several reports from LGBTQ activists of content removal, such as the banning of a trans women from Facebook after she displayed a photograph of her new hairstyle and referred to herself as a 'tranny.' Another example used by Oliva (2020) is a research study which revealed that African American English tweets are twice as likely to be considered offensive compared to others,

[50] Llanso et al, 'Artificial Intelligence, Content Moderation and Freedom of Expression' (2020) Transatlantic Working Group, pg.9

[51] 'A Rights-Respecting Model of Online Content Regulation by Platforms' (2018) *Global Partners Digital* 22

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 March 2022 doi:10.20944/preprints202203.0258.v1

reflecting the infiltration of racial biases in technology. An assessment of AI tools for regulating harmful text found that African American English tweets are twice as likely to be labelled offensive compared to others.' Oliva et al (2021) pointed to the 'confounding effects of dialect' which need to be taken into account in order to avoid racial biases in hate speech detection. This reflects the significance of contextualizing speech, something which does not bode well with the design and enforcement of automated mechanisms and which could pose risks to the online participation of minority groups. Moreover, automated mechanisms fundamentally lack the ability to comprehend the nuance and context of language and human communication. For example, YouTube removed 6,000 videos documenting the Syrian conflict.[52] It shut down Qasioun News Agency,[53] an independent media group reporting on war crimes in Syria. Several videos were flagged as inappropriate by an automatic system designed to identify extremist content. As Oliva (2020) notes, other hash matching technologies, such as PhotoDNA, also seem to operate in 'context blindness' which could be the reason for the removal of those videos. Facebook banned the word 'kalar' in Myanmar.. Radicals have given this word a "derogatory connotation" used to attack the Rohingya people in Myanmar. The word was picked up through automated mechanisms, deleting posts which may use it in another context or with another meaning (including kalar oat – camel). This led to the removal of posts which condemned the fundamentalist movements in the country. For example, the below included the users' opinion that extreme nationalism and religious fundamentalism are negative factors:

[52]YouTube 'made wrong call' on Syria videos' (23 August 2017) BBC News, available at: <https://www.bbc.com/news/technology-41023234> [Accessed 15 December 2021]

[53]Ibid.

In light of the examples above, the problems of using AI *vis-à-vis* alleged hate speech, results not only in an infringement of the freedom of expression, due to over-blocking, but, also, to a violation of the right to non-discrimination.

5. Conclusions

The Council of Europe has proposed 10 recommendations which can be adopted to protect human rights when it comes to the use of AI. These include the establishment of a legal framework to carry out human rights impact assessments of AI systems in place, the evaluation of AI systems through public consultations, the obligation of Member States to facilitate the implementation of human rights standards in private companies (such as social media companies), a transparent and independent oversight of AI systems with special attention being granted to groups disproportionately impacted by AI, such as ethnic and religious minorities, due regard to human rights, particularly expression, the rule that AI must always remain under human control, with the possibility of remedies to victims of human rights violations arising from the functioning of AI as well as promotion of AI literacy. In relation to the latter, there is space for human rights training and capacity building for those who are directly or indirectly involved in the application of AI systems.[54]

The recommendations are indeed useful for improving the current landscape of using automated mechanisms to respond to online hate speech. However, social media companies must be wary of the structural issues when it comes to the use of such mechanisms for removal of hate speech. First and foremost, it must be underlined that, as noted by Llanso (2020), the above issues cannot be tackled with more sophisticated AI. Moreover, as noted by Perel and Elink-Koren (2016) 'the process of translating legal mandates into code inevitably embodies particular choices as to how the law is interpreted, which may be affected by a variety of extrajudicial considerations, including the conscious and unconscious professional assumptions of program developers, as well as various private business incentives.' Whilst automated mechanisms can be useful in assisting human moderators by picking up on potentially hateful speech, they should not be solely 'responsible'

[54] Council of Europe, 'Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights' (2019) available at: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 [Accessed 3 December 2021]

 doi:10.20944/preprints202203.0258.v1

for removing hate speech. Biased training data sets, the lack of relevant data and the lack of conceptualization of context and nuance, can lead to wrong decisions which can have dire effects on the ability of minority groups to function equally in the online sphere.

References

Adrian Marshall Williams & Jonathan Cooper, 'Hate Speech, Holocaust Denial and International Human Rights Law' (1999) 7 *European Human Rights Law Review* 593

Alexandra Siegel et al, 'Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath' (2019) <https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_et_al_election_hatespeech_qjps.pdf >[Accessed 5 January 2022]

Andrew Lester *et al*, '*Human Rights Law and Practice'* (3rd edn. LexisNexis, New York 2009)

Claudio Grossman, 'Challenges to Freedom of Expression within the Inter-American System: A Jurisprudential Analysis' 34 (2012) *Human Rights Quarterly*

Daphne Keller, 'Internet Platforms: Observations on Speech, Danger and Money' (2018) *Hoover Institution*

Emma Llanso 'No Amount of AI in Content Moderation Will Solve Filtering's Prior-Restraint Problem' (2020) *Big Data and Society*

Emma Llanso et al, 'Artificial Intelligence, Content Moderation and Freedom of Expression' (2020) Transatlantic Working Group

Jacob Mchangama & Natalie Alkiviadou, 'The Digital Berlin Wall: How Germany Built a Prototype for Online Censorship' (2020) *Euractiv* <https://www.euractiv.com/section/digital/opinion/the-digital-berlin-wall-how-germany-built-a-prototype-for-online-censorship/?fbclid=IwAR1fRPCtnP5ce_Glx77uaIB1sIS37BqqHdo-SliBiQWkYmGD3y7f8DaPOi4> [Accessed 4 January 2022]

Jacob Mchangama & Joelle Fiss, 'Digital Berlin Wall: How Germany (Accidentally Created a Prototype for Global Online Censorship' *Justitia* (2019)

Jacob Mchangama & Natalie Alkiviadou, 'Digital Berlin Wall: How Germany (Accidentally Created a Prototype for Global Online Censorship – Act Two' *Justitia* (2020)

Jacob Mchangama et al, 'A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of Platformization' (2021) *Justitia*

Jerome Shestack, 'The Jurisprudence of Human Rights,' in Theodor Meron (ed), *'Human Rights in International Law: Legal and Policy Issues'* (1st edn. Clarendon, Oxford 1984)

Joch Cowls et al, 'Freedom of Expression in the Digital Public Sphere' *AI and Platform Governance*

John Perry Barlow, A Declaration of the Independence of Cyberspace, 8 February 1996 <https://www.eff.org/cyberspace-independence> [Accessed 5 January 2021]

Maayan Perel & Niva Elkin-Koren, 'Accountability in Algorithmic Copyright Enforcement' (2016) 19 Stanford Technology Law Review 473 (2016)

Mari Matsuda al, '*Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*' (1993 Westview Press)

Mark Slagle, 'An Ethical Exploration of Free Expression and the Problem of Hate Speech' (2009) 24 *Journal of Mass Media Ethics* 4

Natalie Alkiviadou, 'Regulating Hate Speech in the EU' in Stavros Assimakopoulos, Fabienne H Baider & Sharon Millar (eds), '*Online Hate Speech in the EU: A Discourse Analytical Perspective*' (1st edn. Springer Briefs in Linguistics 2017).

Natasha Duarte & Emma Llanso 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (2017) *Centre for Democracy and Technology*

Perel M & Elkin-Koren N, 'Accountability in Algorithmic Copyright Enforcement' (2016) 19 Stanford Technology Law Review
Robert Gorwa et al, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) *Big Data & Society*

Rodney Smolla, 'Academic Freedom, Hate Speech and the Idea of a University.' 53 *Law and Contemporary Problems* 3

Roger Kiska, 'Hate Speech: A Comparison Between the European Court of Human Rights and the United States Supreme Court Jurisprudence' (2012) 25 *Regent University Law Review* 1

Stephanie Farrior, 'Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech.' (1996) 14 *Berkley Journal of International Law* 1

Tarlach McGonagle, 'The Council of Europe Against Online Hate Speech: Conundrums and Challenges' Expert Paper, Institute for Information Law, Faculty of Law (2013) <https://rm.coe.int/168059bfce> [Accessed 4 January 2022]

Tarlach McGonagle, 'Wresting Racial Equality from Tolerance of Hate Speech' (2001) 23 *Dublin University Law Journal* 21

Thiago Oliva Dias, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) *Human Rights Law Review* 20

Thiago Oliva Dias et al, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 *Sexuality & Culture*

Uladzislau Belavusau, *'Freedom of Speech: Importing European and US Constitutional Models in Transitional Democracies'* (1st edn. Routledge, London 2013)

Yulia A Timofeeva, 'Hate Speech Online: Restricted or Protected? Comparison of Regulations in the United States and Germany' (2003) 12 Journal of Transnational Law and Policy

 doi:10.20944/preprints202203.0258.v1