v. 19 n. 32 São Paulo Dec. 2022

sur

international journal
on human rights

issue 32

ARTICLE

ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH MODERATION

Natalie Alkiviadou

• *A Risky Match?* •

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 it is developed and growing state pressure on companies 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 violations of the freedom of expression and the right to non-discrimination.

KEYWORDS

Freedom of expression | Hate speech | Artificial intelligence | Social media platforms

1 • Introduction

Social media platforms (SMPs) are the primary vehicle for communication and information. They 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 media, which may be state censored, permit the dissemination of daily news and raise awareness on human rights violations. However, as noted by Mchangama *et al.*[1], the massive use of SMPs gives new visibility to phenomena such as hate and abuse. 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, the author argues 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, Siegel *et al.* conducted a study to assess whether Trump's 2016 election campaign (and the six-month period following it) led to a rise in hate speech on Twitter.[2] Based on an analysis of a sample of 1.2 billion tweets, they 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."

Even so, state pressure for platform regulation of hate speech is increasing, which, as argued in this paper, has led to the dilution of the right to free speech and has directly contributed to the silencing of minority groups. The manner in which this 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, several of which were ranked by Freedom House as "not free" or "partly free".[3] All countries require online platforms to remove vague categories of content that include "false information", "blasphemy/religious insult" and "hate speech". Mchangama and Alkiviadou note worryingly that "few of these countries have in place the basic rule of law and free speech protections built into the German precedent."[4] A similar template is currently being followed at the European Union (EU) level with the Digital Services Act (DSA).[5]

As a response to enhanced regulatory requirements, due to the risk of steep fines, platforms are prone to taking the "better safe than sorry" approach and regulating content rigorously. However, as noted by Llanso,[6] 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. To respond to both 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 Dias *et al.*,[7] 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 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, thus further silencing already marginalized groups.

2 • Hate speech: semantics and notions

There is no universally accepted definition of hate speech. Most states and institutions are adopting their own understanding of what it entails,[8] without defining it.[9] One of the few, albeit non-binding, documents that has sought to elucidate the meaning of the term is the Recommendation of the Council of Europe's Committee of Ministers on hate speech.[10] 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.*

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 intolerance, including religious intolerance."[11] The inclusion of merely justifying hatred demonstrates the low threshold for speech to be considered unacceptable. Furthermore, in its rulings, the ECtHR has held that to be considered hate speech, it is not necessary for speech "to directly recommend individuals to commit hateful acts",[12] since attacks on persons can be committed by "insulting, holding up to ridicule or slandering specific groups of the population"[13] and that "speech used in an irresponsible manner may not be worthy of protection."[14] 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 sufficient harm to justify limiting free speech.

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 of the report, "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 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.

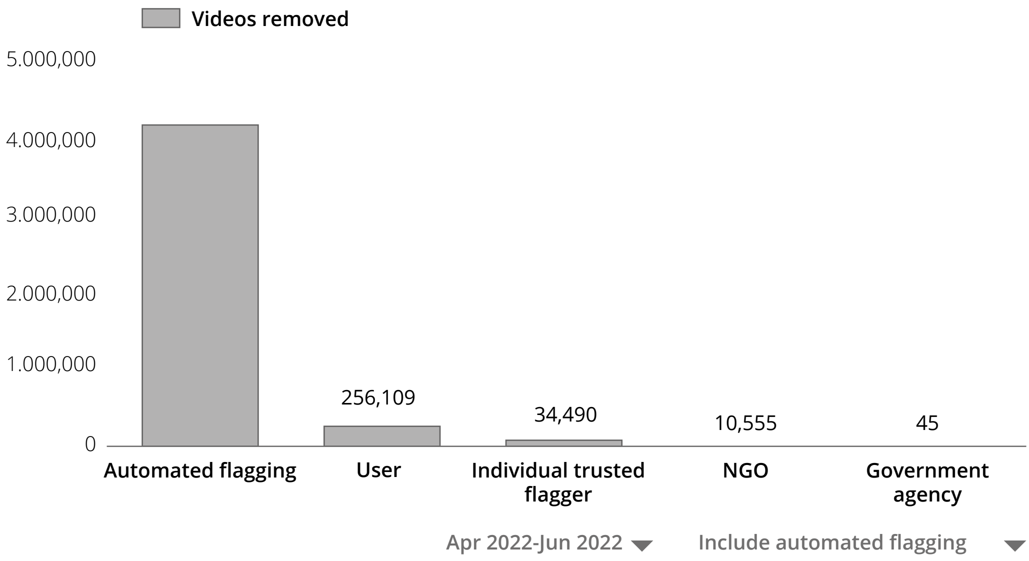
Turning now to the platforms themselves, while it is beyond the scope of this paper to assess all the guidelines and standards for SMPs, we look at two different approaches: Facebook and Instagram, on the one hand (both owned by Meta Platforms Inc..), and Reddit, on the other. The former[18] formulate their understanding of hate speech based on three tiers, the first being violent and dehumanizing speech and the second, statements of inferiority, contempt, dismissal and other forms of 'offence' such as repulsion. Tier three includes statements pertaining to segregation and exclusion. The list of protected characteristics is broad, including aspects 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 it uses include race, colour, religion and pregnancy, amongst others. It is noteworthy that all major platforms, including the ones above as well as Twitter,[21] YouTube,[22] and TikTok,[23] incorporate the grounds of race and religion in the list of protected characteristics.

3 • Artificial Intelligence

The use of AI is a response to increasing state pressure on social media platforms to remove hate speech quickly and efficiently. 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 Dias, such circumstances have prompted companies to "act proactively in order to avoid liability... in an attempt to protect their business models."[24]

To exemplify the use of AI by social media platforms, one can compare proactive rates of hate speech removal between the first quarter of 2018 (at 38%) and the second quarter of 2022 (at 95.6%) As noted in a post on the Transparency Center website, "our technology proactively detects and removes the vast majority of violating content before anyone reports it."[25]

In its latest enforcement report[26] (Q2 of 2022), YouTube put forth the illustration below, demonstrating the percentage of human and automated flagging across the board of removable content (not just hate speech):



Dias *et al.* argue that the algorithms developed to achieve this automation are habitually customized for content type, such as pictures, videos, audio and text.[27] As Duarte and Llanso found,[28] current technologies detect harmful text by using natural language processing and sentiment analysis and, even though they have evolved significantly, their 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, Dias *et al.* argue that going from policy to code may lead to changes in meaning, since machine language is more limited than its human counterpart.[29] Given the power that 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.* argue that there is an urgent need to ensure that content moderation occurs in a manner that safeguards human rights and public discourse.[30]

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, private profit-making companies' increased reliance on AI – while focusing on freedom of expression and non-discrimination.

4 • AI, hate speech and the challenges to the freedom of expression

Article 19 of the Universal Declaration of Human Rights (UDHR) provides that "[e] veryone 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."

The right to this freedom is also protected in other major documents such as Article 19 of the International Covenant on Civil and Political Rights (ICCPR) and Article 10 of the European Convention on Human Rights. Both articles include limitations to the freedom of expression, while Article 20 of the ICCPR stipulates 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.*

As noted by Dias,[31] 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 universal 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 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".[32] As Llanso *et al.* point out,[33] it poses "distinct challenges for freedom of expression and access to information online." A Council of Europe report highlights that the use of AI for hate speech regulation directly impacts the freedom of expression, which raises concerns about the rule of law and, in particular, notions of legality, legitimacy and proportionality.[34] 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.[35] Gorwa *et al.* 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".[36] Moreover, regardless of the technical specifications of a particular mechanism, proactive identification (and removal) of hate speech constitutes prior restraint of speech, with all the legal issues that this entails. Specifically, Llanso *et al.* argue that there is a "strong presumption against the validity of prior censorship in international human rights law."[37] Former UN Special Rapporteur on the Freedom of Opinion and Expression, David Kaye, expressed his concern about the use of automated tools in terms of potential 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."[38]

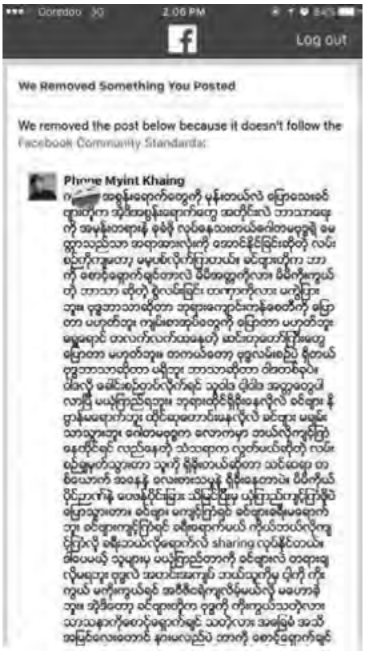
5 • AI and challenges to non-discrimination

Dias argues that the use of AI may result in the biased enforcement of companies' terms of service.[39] This can be due to a lack of data and/or biased training datasets, leading to the potential silencing of members of minority communities.[40] 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.[41] Although technologies such as natural language processing and sentiment analysis have been developed to detect harmful text without having to rely on specific words or phrases, research has shown that, as Dias *et al.* put it, they are "still far from being able to grasp context or to detect the intent or motivation of the speaker".[42] As noted by Dias,[43] although hash-matching is widely used to identify child sexual abuse content, it is not easily transposed to other cases such as extremist content, which "typically requires assessment of context."

In relation to this, Keller 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."[44] 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 Dias *et al.*,[45] such technologies are just not cut out to pick up on the language used by, for example, the LGBTQ community whose "mock impoliteness" and use of terms such as "dyke", "fag" and "tranny" are a way of reclaiming power and a means for preparing members of this community to "cope with hostility". Dias *et al.* give several reports from LGBTQ activists on content removal, such as the banning of a trans woman from Facebook after she displayed a photograph of her new hairstyle and referred to herself as a "tranny".[46] Another example used by Dias is a research study that revealed that African American English tweets are twice as likely to be considered offensive compared to others, thus reflecting the infiltration of racial biases in technology.[47] Dias *et al.* pointed to the "confounding effects of dialect" that need to be taken into account in order to avoid racial biases in hate speech detection.[48] This reflects the significance of contextualizing speech – something that does not bode well with the design and enforcement of automated mechanisms and that 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.[49] It shut down the Qasioun News Agency,[50] 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 Dias notes,[51] 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, as radicals had given this word a "derogatory connotation" and used it to attack the Rohingya people in Myanmar. The word was picked up through automated mechanisms that deleted posts which may have used it in another context or with another meaning (including kalar oat, which means camel). This led to the removal of posts condemning the fundamentalist movements in the country. For example, the post below included the user's opinion that extreme nationalism and religious fundamentalism are negative factors:



Ooredoo 3G 2:06 PM 84%

Log out

We Removed Something You Posted

We removed the post below because it doesn't follow the Facebook Community Standards:

Phone Myint Khaing

[illegible] sharing [illegible]

Source: Author's archive.

In light of the examples above, the problems of using AI to deal with alleged hate speech result not only in an infringement of the freedom of expression due to over-blocking, but also violations of the right to non-discrimination.

6 • Conclusions

The Council of Europe has proposed 10 recommendations that can be adopted to protect human rights when it comes to the use of AI. They include, for instance, 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); transparent and independent oversight of AI systems which gives special attention to groups disproportionately impacted by AI, such as ethnic and religious minorities; due regard to human rights, particularly the freedom of expression; the rule that AI must always remain under human control and states should offer effective access to remedy for victims of human rights violations arising from the way AI functions. It also refers to the promotion of AI literacy. In relation to the latter, there is space for offering human rights training and capacity-building to those who are directly or indirectly involved in the application of AI systems.[52]

These 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 structural issues arising from the deployment of such mechanisms for removal of hate speech. First and foremost, it must be underlined that, as noted by Llanso,[53] the above issues cannot be tackled with more sophisticated AI. Moreover, as noted by Perel and Elink-Koren, "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."[54] Whilst automated mechanisms can assist human moderators by picking up on potentially hateful speech, they should not be solely responsible 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.

NOTES

1 • Jacob Mchangama *et al.*, "A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of Platformization," Justitia, November 2021, accessed November 25, 2022, https://futurefreespeech.com/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf.

2 • Alexandra Siegel *et al.*, "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath." Alexandra Siegel, March 6, 2019, accessed January 5, 2022, https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_et_al_election_hatespeech_qjps.pdf.

3 • Jacob Mchangama *et al.*, "A Framework of First Reference," November 2021.

4 • Jacob Mchangama and Natalie Alkiviadou, "The Digital Berlin Wall: How Germany Built a Prototype for Online Censorship." Euractiv, October 8, 2020, accessed January 4, 2022, https://www.euractiv.com/section/digital/opinion/the-digital-berlin-wall-how-germany-built-a-prototype-for-online-censorship/?fbclid=IwAR1fRPCtnP5ce_Glx77uaIB1sIS37BqqHdo-SliBiQWkYmGD3y7f8DaPOi4.

5 • "The Digital Services Act: ensuring a safe and accountable online environment," European Commission, 2022, accessed October 17, 2022, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#documents.

6 • Emma J. Llansó, "No Amount of AI in Content Moderation Will Solve Filtering's Prior-Restraint Problem," *Big Data & Society* 7, no. 1 (2020).

7 • Thiago Oliva Dias et al., "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online," *Sexuality & Culture* 25 (2021): 700-732.

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

9 • Natalie Alkiviadou, "Regulating Hate Speech in the EU," in *Online Hate Speech in the EU: A Discourse Analytical Perspective*, 1st ed., eds. Stavros Assimakopoulos, Fabienne H. Baider, and Sharon

Millar (Springer Cham, 2017).

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

[11] 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.

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

[13] *Ibid*.

[14] *Ibid*. para. 55.

[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.

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

[18] “Hate Speech,” Meta Transparency Center, 2022, accessed October 25, 2022, https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech.

[19] “Hate Speech,” Meta Transparency Center, 2022, accessed November 2, 2022, https://transparency.fb.com/policies/community-standards/hate-speech/#policy-details.

[20] “Promoting Hate Based on Identity or Vulnerability,” Reddit, 2020, accessed November 2, 2022, https://www.reddithelp.com/hc/en-us/articles/360045715951.

[21] “Hateful Conduct Policy,” Twitter, 2016, accessed January 2, 2022, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[22] “Hate Speech Policy,” YouTube, 2019, accessed January 2, 2022, https://support.google.com/youtube/answer/2801939?hl=en.

[23] “Community Guidelines,” TikTok, 2022, accessed October 2, 2022, https://www.tiktok.com/community-guidelines?lang=en#38.

[24] Thiago Oliva Dias, “Content Moderation Technologies: Applying *Human Rights Standards to Protect Freedom of Expression,” Human Rights Law Review* 20, no. 4 (2020): 607-640.

[25] “How Technology Detects Violations,” Meta Transparency Center, January 19, 2022, accessed November 3, 2022, https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/.

[26] “YouTube Community Guidelines enforcement,” YouTube, 2022, accessed November 3, 2022, https://transparencyreport.google.com/youtube-policy/removals.

[27] Thiago Oliva Dias *et al.*, “Fighting Hate Speech, Silencing Drag Queens?,” (2021).

[28] Natasha Duarte and Emma J. Llansó, “Mixed Messages? The Limits of Automated Social Media Content Analysis.” Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81 (2018): 106-106.

[29] Thiago Oliva Dias *et al.*, “Fighting Hate Speech, Silencing Drag Queens?,” (2021).

[30] Joch Cowls *et al*., “Freedom of Expression in the Digital Public Sphere,” AI and Platform Governance, 2020, accessed November 25, 2022, https://doi.org/10.5281/zenodo.4292408.

[31] Thiago Oliva Dias, “Content Moderation Technologies,” (2020).

[32] Emma J. Llansó, “No Amount of AI…,” 2020.

[33] Emma Llanso *et al.*, “Artificial Intelligence, Content Moderation and Freedom of Expression.” Transatlantic Working Group, 2020, accessed November 23, 2022, https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

[34] “Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications”, Council of Europe, DGI (2017) 12, 2017, accessed November 23, 2022, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, 18.

[35] *Ibid.* 21.

[36] Robert Gorwa et al., “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance,” *Big Data*

& Society 7, no. 1 (2020).

37 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation...," 2020.

38 • "Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression," OHCHR, June 13, 2018, accessed November 10, 2022, https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

39 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

40 • Emma Llanso *et al.*, "Artificial Intelligence, Content Moderation...," 2020.

41 • Natasha Duarte and Emma J. Llansó, "Mixed Messages?...," 2018.

42 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

43 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

44 • Daphne Keller, "Internet Platforms: Observations on Speech, Danger and Money," *Hoover Institution's Aegis Paper Series*, no. 1807 (2018).

45 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

46 • *Ibid.*

47 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

48 • Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

49 • "YouTube 'made wrong call' on Syria videos'," BBC News, August 23, 2017, accessed October 2022, *https://www.bbc.com/news/technology-41023234*.

50 • *Ibid*.

51 • Thiago Oliva Dias, "Content Moderation Technologies," (2020).

52 • "Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights," Council of Europe, 2019, accessed November 23, 2022, https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

53 • Emma J. Llansó, "No Amount of AI...," 2020.

54 • Maayan Perel and Niva Elkin-Koren, "Accountability in Algorithmic Copyright Enforcement," *19 Stanford Technology Law Review* 473 (2016), accessed November 25, 2022, https://law.stanford.edu/wp-content/uploads/2016/10/Accountability-in-Algorithmic-Copyright-Enforcement.pdf.

NATALIE ALKIVIADOU – *Cyprus/Denmark*

Natalie Alkiviadou is Senior Research Fellow at Justitia (Denmark). Her interests lie in the freedom of expression, the far-right, hate speech and hate crime. She has published three monographs and a range of peer-reviewed articles. Alkiviadou is a fellow of the Information Society Law Centre of the Università degli Studi di Milano.

email: *natalie@justitia-int.org*

Received in September 2022.
Original in English.