The Opportunities and Challenges of Addressing Hate Speech with Artificial Intelligence

Submission to the OSCE Representative on Freedom of the Media
#SAIFE Public Consultation

Katie Pentney & Tarlach McGonagle*

September 2020

* Katie Pentney, Adv. LL.M. (Leiden University) is an independent researcher. Prof. dr. Tarlach McGonagle holds the chair in Media Law & Information Society at Leiden University. The authors are currently developing this line of research on the implications of AI for freedom of expression and hate speech.

Introduction

The #SAIFE consultation provides an excellent opportunity to address pressing freedom of expression issues throughout OSCE participating States, and how artificial intelligence (AI) can shape or re-shape those issues. This submission examines the complex relationship between AI and hate speech: how AI can be used to create and disseminate – but also to detect and curb – hate speech. This complex relationship is appropriately one of the main areas of concern in the #SAIFE project. This submission explores a range of opportunities and challenges concerning the interplay between free speech, hate speech and AI, and builds upon our ongoing research in this area.

Hate speech has been a kind of gordian knot, an intractable problem for states to address for a host of reasons: from difficulties defining it in a way that respects freedom of expression, to applying this definition in a consistent manner; from discerning how best to regulate and eradicate hate speech, to how to remedy violations of overzealous (or underwhelming) application. At the same time, States must ensure that the targets of hate speech are adequately protected against the varying harms it causes and that they have effective remedies whenever their rights or dignity are violated. These challenges are not new. However, they are magnified by technological innovations, such as the internet and social media platforms which enable and facilitate the rapid and global dissemination of hate speech; its amplification and its enduring presence online. Parallel advancements in AI technology may be a partial answer to this problem – not by unravelling or eradicating hate speech itself, but by stemming its tide.

The #SAIFE Paper identifies several concerns in relation to hate speech and AI, namely: the lack of uniform definition in international human rights law, the contextual analysis required to identify and remove hate speech which has proven difficult for AI systems, and disproportionate effects on minority groups as a result of bias in AI design. In addition to these well-founded areas of concern, we wish to highlight three further challenges and opportunities which require further study and reflection:

(i) the differentiated roles and responsibilities of State and private actors in a broader, human rights-driven and multi-stakeholder approach;
(ii) the need for non-discriminatory policies and practices in the design and implementation of measures to block/remove hate speech; and
(iii) the broader effects on public debate where AI is used to moderate and censor expression.

The first section provides a brief background of how AI is being used to address hate speech at present, and outlines standard-setting and policy-making initiatives to date at the international and European levels. The second section addresses the three challenges and opportunities outlined above and proposes areas of further research and study. The final section sets out preliminary general conclusions, supplemented by specific priority issues and lines of enquiry that could fruitfully be pursued in the next phase of the #SAIFE project.

Background

Hate speech finds itself at the frontier between the right to freedom of expression – which includes information and ideas that 'offend, shock or disturb'[1] – and the imperative of upholding the rights and freedoms of those on the receiving end of such expression as well as the public at large.[2] States are ultimately responsible for protecting these competing rights and interests – whether it is through policing hate speech, battling disinformation, or combatting extremist threats in streets and parks within their borders, or in the global 'marketplace of ideas' proffered by the internet.[3]

Increasing attention has been paid in recent years to the rise of hate speech on the internet – particularly on social media platforms such as Facebook, Twitter and Reddit – and on how AI technology such as machine learning can address it.[4] States and regional bodies, such as the European Commission, have exerted increasing pressure on these intermediaries to address hate speech on their platforms. For instance, Germany and France both passed laws requiring social media sites to remove content within twenty-four hours of notification.[5] The European Commission launched a Code of Conduct on Countering Illegal Hate Speech Online in 2016 together with Facebook, Microsoft, Twitter and YouTube, which has made strides in improving response times to flagged content and in removal of content deemed to be illegal hate speech.[6]

[1] *Handyside v. United Kingdom*, European Court of Human Rights (ECtHR) judgment of 7 December 1976, at para. 49. The right is recognized in international and regional treaties, as well as domestic laws and constitutions. See, e.g., International Covenant on Civil and Political Rights (ICCPR), Article 19; European Convention on Human Rights (ECHR), Article 10. See, e.g., Section 2(b) of the Canadian Charter of Rights and Freedoms; the First Amendment of the Constitution of the United States of America; Article 7 of the Constitution of the Kingdom of the Netherlands.

[2] These imperatives are reflected in limitation clauses (such as Article 19(3) ICCPR and Article 10(2) ECHR) as well as abuse of rights clauses (Article 5 ICCPR and Article 17 ECHR).

[3] *Mouvement Raëlien Suisse*, Grand Chamber Judgment of the European Court of Human Rights of 13 July 2012 (Dissenting opinion of Judge Pinto de Albuquerque, who remarked: "If the streets and parks of a city are the historical quintessential public fora, the Internet is today's global marketplace of ideas").

[4] Machine learning refers to "computer algorithms that have the ability to 'learn' or improve in performance over time" including through detecting patterns to automate complex tasks or make predictions. See H. Surden, "Machine Learning and Law," *Washington Law Review*, Vol. 88, at 88-9.

[5] Germany passed the Network Enforcement Law (NetzDG) in 2017. The French 'Avia' law was recently declared unconstitutional. See Article 19, "France: Constitutional Council declares French hate speech 'Avia' law unconstitutional," 18 June 2020, at https://www.article19.org/resources/france-constitutional-council-declares-french-hate-speech-avialaw-unconstitutional/.

[6] The Code of Conduct is available at:
http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf. 9 IT companies have now agreed to adhere to it, with the addition of Instagram, Dailymotion, Snapchat, Jexuvideo.com and (most recently) TikTok. European Commission, "The Code of conduct on countering illegal hate speech online," 22 June 2020, http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf;
European Commission, "The Eu Code of conduct on countering illegal hate speech online," at
https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

Social media platforms appear to be rising to the challenge. In the first three months of 2020 alone, Facebook removed 9.6 million pieces of content that it qualified as hate speech.[7] Advancements in the AI techniques used meant that nearly 90% of the hate speech removed during this period was automatically detected without human intervention (in the form of flagging the content as a violation of the platform's guidelines).[8] However, concerns about over-capture (for instance, removing 'counter-speech',[9] political speech and satire,[10] and documented human rights violations[11]) and anomalous results[12] remain. Moreover, the emphasis has been on filtering, blocking and removal, which can be seen as "fire-fighting" activities. A comprehensive approach to countering hate speech – in all of its manifestations – necessarily includes a range of "fire-prevention" strategies: educational, informational, awareness-raising measures and investment in capacity-building for minority groups and the creation of (inter-group) dialogical opportunities and fora. The aforementioned Code of Conduct contains commitments to such strategies, but there has been relatively little scrutiny or assessment of how and to what extent those commitments are actually being fulfilled.

In light of the rapid-fire spread of networked hate speech, and the promise of AI to address it, the topics of hate speech and AI – sometimes addressed independently, sometimes in

---

[7] J. Kahn, "Facebook makes strides using A.I. to automatically find hate speech and COVID-19 misinformation," Fortune, May 12, 2020, https://fortune.com/2020/05/12/facebook-a-i-hate-speech-covid-19-misinformation/; Facebook Community Standards Enforcement Report (August 2020), "Hate Speech," https://transparency.facebook.com/community-standards-enforcement#hate-speech.

[8] Kahn, *id.* These gains are due in large part to the XLM-R system developed by Facebook, which "trained on two terrabytes of data, or about the equivalent of all the words in half a million 300-page books. It learns the statistical map of all of those words across multiple languages at once. The idea is that conceptual commonalities between hate speech in any language will mean the statistical maps of hate speech will look similar across every language even if the words themselves are completely different" (*id*).

[9] "Counter-speech" refers to "a response to hate speech that may include the same offensive terms", a phenomenon that is particularly difficult for AI to correctly classify because it shares so many characteristics with the hate speech it is attacking. See Facebook AI, "AI advances to better detect hate speech," 12 May 2020, https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/. For a rich theoretical exposition of counter-speech and the question of its suitability as a remedy for hate speech, see: K. Gelber, Speaking Back: The Free Speech Versus Hate Speech Debate (Amsterdam/Philadelphia, John Benjamins Publishing Company, 2002), and K. Gelber, "Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia)", in M. Herz & P. Molnar (eds.), *Content and Context: Rethinking Regulation and Remedies for Hate Speech*, (New York, Cambridge University Press, 2012) at pp. 198-216.

[10] Reuters, "German opposition calls for abolition of online hate speech law," 7 January 2018, https://www.reuters.com/article/us-germany-hatecrime/german-opposition-calls-for-abolition-of-online-hate-speech-law-idUSKBN1EW0Q9; M. Schaake, MEP, "When YouTube took down my video," 7 October 2016, https://www.marietjeschaake.eu/en/when-youtube-took-down-my-video.

[11] M. Browne, "YouTube Removes Videos Showing Atrocities in Syria," The New York Times, 22 August 2017, https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.

[12] B. Wagner, "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making," *Policy & Internet*, Vol. 11(1) (2019), at 112-4; Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy, "Exploring the Role of Algorithms in Online Harmful Speech," 10 August 2017, https://shorensteincenter.org/exploring-role-algorithms-online-harmful-speech/; A. Shahani, "From Hate Speech to Fake News: The Content Crisis Facing Mark Zuckerberg," NPR, 17 November 2016, https://www.npr.org/sections/alltechconsidered/2016/11/17/495827410/from-hate-speech-to-fake-news-the-content-crisis-facing-mark-zuckerberg?t=1601217474162.

conjunction – have gained increasing attention in recent years at the international, regional and national levels. For instance, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has released separate reports addressing AI and online hate speech (in 2018 and 2019, respectively).[13]

In Europe, initiatives are underway at the Council of Europe (CoE) and the European Union (EU). The CoE Ad hoc Committee on Artificial Intelligence is examining the feasibility and design of a rights-compliant legal framework for AI,[14] and the Committee of Experts on Combating Hate Speech is preparing a Recommendation to be adopted by the Committee of Ministers on a comprehensive (and human rights-based) approach to hate speech, including in an online environment.[15] The Committee of Ministers has already adopted Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems[16] and a Declaration on the manipulative capabilities of algorithmic processes.[17] AI and human rights is a thematic area of focus for the CoE Commissioner for Human Rights,[18] who issued last year a Recommendation on preventing or mitigating the negative impacts of AI-systems on human rights.[19]

In the EU, the European Commission has adopted a Communication on Artificial Intelligence for Europe.[20] In 2018, it created a High-Level Expert Group on AI, which published its Ethics Guidelines for Trustworthy AI and Policy and investment recommendations for trustworthy AI in 2019.[21] The EU Agency for Fundamental Rights also convened a roundtable discussion in 2019

[13] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Artificial Intelligence technologies and implications for the information environment (A/73/348), 29 August 2018, https://undocs.org/en/A/73/348; Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Online hate speech (A/74/486), 9 October 2019, https://undocs.org/A/74/486.

[14] Council of Europe, "Publication of the first progress report of the Ad hoc Committee on Artificial Intelligence (CAHAI)," 29 September 2020, https://www.coe.int/en/web/artificial-intelligence/-/publication-of-the-first-progress-report-of-the-ad-hoc-committee-on-artificial-intelligence-cahai-.

[15] Council of Europe, ADI/MSI-DIS Committee of Experts on Combating Hate Speech, https://www.coe.int/en/web/committee-on-combatting-hate-speech/home.

[16] Adopted on 8 April 2020.

[17] Adopted on 13 February 2019.

[18] Council of Europe Commissioner for Human Rights, "Artificial Intelligence and Human Rights," https://www.coe.int/en/web/commissioner/thematic-work/artificial-intelligence.

[19] Council of Europe Commissioner for Human Rights, "Unboxing Artificial Intelligence: 10 steps to protect Human Rights," Recommendation (May 2019), https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

[20] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, COM(2018) 237 final, 25 April 2018.

[21] European Commission, High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

focusing on AI and online hate speech.[22] Parallel initiatives are ongoing at the domestic level in a number of OSCE participating States.[23]

Challenges and Opportunities

Many of the aforementioned initiatives have focused on technical requirements for the private sector – such as detection and notification, effective removal, and prevention of reappearance requirements[24] – as well as the legal challenges of using AI to address hate speech.[25]

This submission encourages further study and consideration of three questions:

(i) the differentiated roles and responsibilities of State and private actors in a broader, human rights-driven and multi-stakeholder approach;
(ii) the need for non-discriminatory policies and practices in the design and implementation of measures to block/remove hate speech; and
(iii) the broader effects on public debate where AI is used to moderate and censor expression.

Each of these issues is briefly introduced here, with a view to opening up space for further research and development.

(i) The Differentiated Roles of State and Private Actors

Countering hate speech traditionally fell to the state – from the legislature framing and/or defining it, to the police detecting it, to the courts adjudicating it. Importantly, states are obliged

[22] European Union Agency for Fundamental Rights, Opening address – Artificial intelligence and online hate speech, delivered by Michael O'Flaherty, 31 January 2019, https://fra.europa.eu/en/speech/2019/opening-address-artificial-intelligence-and-online-hate-speech.

[23] For instance, in the United Kingdom, the Alan Turing Institute's Hate Speech: Measures & Counter-measures project, focusing on measuring, analysing and countering hate speech with advanced computational methods (https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures). In the United States, universities such as the Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy are leading the charge on the role of algorithms in online hate speech (https://medium.com/berkman-klein-center/exploring-the-role-of-algorithms-in-online-harmful-speech-1b804936f279). These efforts also extend across borders: for instance, the Canada-UK Artificial Intelligence Initiative has awarded funding to projects which include cross-disciplinary approaches to detect and counter abusive language and hate speech online (https://sciencebusiness.net/news/canada-and-uk-pick-winners-joint-c136m-ai-research-competition).

[24] See, e.g., European Commission, "Security Union: Commission steps up efforts to tackle illegal content online," 28 September 2017 at https://ec.europa.eu/commission/presscorner/detail/en/IP_17_3493; N. Lomas, "Tech giants pressured to auto-flag 'illegal' content in Europe," Tech Crunch, 28 September 2017, at https://techcrunch.com/2017/09/28/tech-giants-pressured-to-auto-flag-illegal-content-in-europe/?_ga=2.142159601.2103368534.1601142340-1600936047.1601142340.

[25] Council of Europe Committee of Experts on Internet Intermediaries, "Algorithms and Human Rights: Study on the human rights dimensions of data processing techniques and possible regulatory implications," DGI(2017)12, at 19-22, https://rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of-aut/1680796d10.

to fulfil their human rights obligations – including freedom of expression, non-discrimination and the right to an effective remedy – in doing so.

In the era of social media, networked hate speech, and AI technology, two new considerations arise in states' exercise of their role policing hate speech.

First, with AI technology, states now have additional (and powerful) tools in their arsenal to combat hate speech, which makes it all the more urgent that the definitions employed, the detection and monitoring functions used, and the removals and remedies undertaken are human rights compliant. Overbroad definitions of hate speech leave open the door for misuse and abuse, including removal of content critical of government or that of political opponents.[26] While the European Court of Human Rights (ECtHR) has resoundingly criticized overt censorship and suppression of activist groups, there is a danger that AI may allow states to accomplish the same ends in a more secretive and insidious way.[27]

This requires renewed focus on states' own use of AI to combat hate speech – separate and apart from their oversight of private corporations doing the same. States are increasingly turning to AI to moderate content online and assist in policing its deleterious impacts in real life. For instance, the UK Home Office has developed machine learning technology to detect terrorist content online,[28] and the promise of AI in police prediction and prevention of hate crimes is an area of recent study.[29] Consideration should be given to the requirements of transparency and accountability which ought to accompany states' use of AI. This should include focus on states' negative obligations (for instance, refraining from prior censorship or suppression of dissent, avoiding discriminatory practices, and ensuring effective remedies for rights violations) as well as their positive obligations (including creating a favourable environment for public discourse and debate[30]). The OSCE and parallel regional organisations play a critical role in setting standards and ensuring States respect their human rights commitments in this new frontier.

---

[26] For analogous cases of suppression of critical speech, see *Ragip Zarakolu v. Turkey*, ECtHR judgment of 15 September 2020; *Castells v. Spain*, ECtHR judgment of 23 April 1992; *Kablis v. Russia*, ECtHR judgment of 30 April 2019.

[27] See, e.g., *Sunday Times v. United Kingdom (no. 1)*, ECtHR judgment of 26 April 1979; *Ahmet Yildirim v. Turkey*, ECtHR judgment of 18 December 2012; *Khadija Ismayilova v. Azerbaijan (no. 2)*, ECtHR judgment of 27 February 2020.

[28] United Kingdom Home Office, *Press Release*, "New technology revealed to help fight terrorist content online," 13 February 2018, https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online.

[29] R. Kelly, "Could Artificial Intelligence Help Police Predict Hate Crimes?" *Digit*, 15 October 2019, https://digit.fyi/artificial-intelligence-hate-crimes-police-hatelab/; D. Lu, "UK police are using AI to spot spikes in Brexit-related hate crimes," *New Scientist*, 28 August 2019, https://www.newscientist.com/article/mg24332453-500-uk-police-are-using-ai-to-spot-spikes-in-brexit-related-hate-crimes/; M.L. Williams et. al, "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crimes," *British Journal of Criminology* (2020) 60, 93-117.

[30] See, for instance, European Court of Human Rights, *Dink v. Turkey*, ECtHR judgment of 14 September 2010.

Second, it is possible to observe a recent and ongoing shift in existing regulatory patterns, which entails a significant move towards foisting greater liability and responsibility on internet intermediaries for illegal third-party content hosted by them or distributed via their services or networks. There is an emergent preference for self-regulatory codes of conduct as a regulatory technique.[31] States have thus increasingly shifted their responsibility for combatting hate speech to private corporations such as Facebook, Twitter and YouTube.[32] In effect, "the private sector has gained unprecedented influence over individuals' right to freedom of expression and access to information".[33] This presents certain advantages for freedom of expression, including practical ones: social media platforms enable and facilitate the rapid and global dissemination of hate speech, as well as its amplification and enduring presence online. Increasing buy-in from private actors to combat hate speech on their platforms will have significant effects on the scope and scale of networked hate speech.

However, states cannot absolve themselves of responsibility by delegating their obligations to private corporations.[34] The buck ultimately stops with the state to ensure that such private entities comply with international and domestic human rights laws in using AI (or any other mechanism) to detect and block/remove networked hate speech. To date, the focus of regulation by states and regional organisations has largely been on technical requirements – including platform guidelines, notification requirements and removal periods. These standards are an important first step in the direction of human rights-compliant AI technology. But greater attention should also be paid to preventive – rather than simply reactive – measures, including awareness raising and civil society engagement.

Moreover, internet platforms are largely left to self-regulate in when and how they use AI to detect and block/remove hate speech. Their content moderation decisions – including those pertaining to hate speech – are largely made on the basis of their terms of service, rather than states' laws.[35] While the UN Guiding Principles on Business and Human Rights (the "Ruggie Principles") may provide guidance on appropriate methods of self-regulation, they are not the complete answer. Unlike hard and fast rules for child labour laws, working hours, or greenhouse emission rates, minimum thresholds are not particularly apt to address hate speech, an inherently contextual exercise. Moreover, it is increasingly clear that profit and reputational

---

[31] T. McGonagle, 'Free expression and internet intermediaries: the changing geometry of European regulation', in: G. Frosio, *The Oxford handbook of online intermediary liability*, Oxford: Oxford University Press 2020; G. Frosio, "Why Keep a Dog and Bark Yourself? From Intermediary Liability to Responsibility", *International Journal of Law and Information Technology*, Vol. 26, Issue 1, 1 March 2018, pp. 1-33, p. 8.

[32] Council of Europe, "Algorithms and Human Rights," Council of Europe Study DGI(2017)12, March 2018, at 20, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5.

[33] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2011), https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf, p. 13. For a detailed overview of the challenges this poses, see D. Kaye, *Speech Police: The Global Struggle to Govern the Internet* (New York: Columbia Global Reports, 2019).

[34] T. McGonagle, "The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing," in R. Frank Jørgensen (ed.), *Human Rights in the Age of Platforms*, The MIT Press: 2019, at 234, https://www.ivir.nl/publicaties/download/CoE_and_internet_intermediaries.pdf.

[35] Wagner, *supra*, at 112.

motivations largely govern such corporations' willingness to address hate speech on their platforms.[36] This may lead to problems of over- or under-inclusion, with worrisome consequences for freedom of expression and non-discrimination.[37] Given the sheer scale and positions of entrenched dominance enjoyed by the so-called "tech giants", it is no longer tenable that governance and content policies continue to be shaped by their CEOs' discretion. Informal regulation "by raised eyebrow",[38] or more accurately by the furrowed brows of advertisers, needs to make way for external, independent oversight mechanisms to ensure public scrutiny.

Further guidance on substantive issues is necessary – including how hate speech is defined to respect freedom of expression, how datasets and trainings must be formulated to avoid discriminatory inputs or effects, and the levels of transparency, accountability and oversight by users and the state required. At each stage, there is a need for enforceable transparency around who is making these decisions, how they were reached, and what remedies exist.

(ii) Eradicating Bias in Design and Implementation

AI has the potential to remove bias and provide neutrality in decisions that affect individuals' day-to-day lives.[39] However, significant concerns have been raised around the creation, replication or amplification of bias by AI systems in other areas, from policing and parole to medicine and employment.[40] The #SAIFE Paper reflects on the risks of bias in AI detection of hate speech – from the disproportionate removal of minority groups' content, to trainings which prioritize certain cultures or groups at the expense of others.[41] These are long-standing concerns;[42] systemic problems that tend to be addressed in piecemeal and reactionary fashion. Successive controversies have prompted criticism of input biases in,[43] and/or racist outcomes of,

[36] H. Ziady, "Facebook and YouTube accept hate speech audits to keep advertisers happy," CNN Business, 24 September 2020, https://edition.cnn.com/2020/09/23/tech/facebook-youtube-advertisers/index.html.

[37] Such profit motivations and reputational interests either support a hands-off approach to hate speech (evident in Facebook's approach during the last United States election and in Myanmar, over the protests of many human rights advocates) or an over-inclusive one where campaigns such as the #StopHateForProfit take hold with advertisers.

[38] Y. Benkler, "A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate," *Harvard Civil Rights Civil Liberties Law Review*, Vol. 46(1), 2011, at 367.

[39] See generally: F.J. Zuiderveen Borgesius, "Discrimination, artificial intelligence, and algorithmic decision-making", Study for the Council of Europe (2018), https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

[40] B. Buchanan & T. Miller, "Machine Learning for Policy Makers: What It Is and Why It Matters," Harvard Kennedy School Belfer Center for Science and International Affairs (June 2017) at 33-4, https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf;
S. Larsson, "The Socio-Legal Relevance of Artificial Intelligence," *Droit et société* 103 (2019), at 578; C. Criado Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men*. London, UK: Penguin Random House, at 163-7.

[41] OSCE Representative on Freedom of the Media, "Spotlight on Artificial Intelligence and Freedom of Expression," (2020), at 57-9.

[42] I. Lapowski, "Google Autocomplete Still Makes Vile Suggestions", *Wired*, 2 December 2018, https://www.wired.com/story/google-autocomplete-vile-suggestions/. See generally (not specifically hate speech): N. Kayser-Bril, "Ten years on, search auto-complete still suggests slander and disinformation", Algorithm Watch, 3 June 2020, https://algorithmwatch.org/en/story/auto-completion-disinformation/.

[43] A. Allen, "The 'three black teenagers' search shows it is society, not Google, that is racist", *The Guardian*, 10 June 2016, https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet.

various algorithm-based services, such as Google's search function,[44] autocomplete function[45] and image-labelling service,[46] and Twitter's image-cropping function.[47]

There is increasing recognition of the need to reflect upon the values, norms and biases we bring to AI development, as well as the impacts AI has on society at large in reproducing and reinforcing certain values, cultures and power dynamics.[48] States and regional organisations, such as the OSCE, should lead the charge in this regard to provide standards and guidance on non-discrimination in the use of AI in combatting networked hate speech. Two areas in particular warrant further study.

First, datasets on which AI is trained should be reflective of the gender, race, religion, and language of the populations affected.[49] Algorithms modelled on imbalanced or unrepresentative datasets may learn incorrect patterns, or become oversensitive to group identifiers, such that individuals who fall within these groups are blocked or their content censored inappropriately.[50] States have a role to play in ensuring the diversity and representativeness of these datasets so that non-dominant groups are not disproportionately excluded or targeted.

Second, standards should be put in place to ensure that trainings do not replicate or entrench existing hierarchies or human biases. For instance, traditional scholarship and jurisprudence has focused on racist and ethnic hate speech. Attention should also be paid to hate speech on other grounds – such as sexual orientation, gender, gender identity, immigration status and disability – and the ways in which this may manifest differently (and require different responses).[51] While there appears to be growing recognition of the need to address hate speech on the basis of sexual orientation – indeed, it is the most commonly reported ground of hate speech in the EU, followed

[44] A. Mahdawi, 'Can Googling be racist?', *The Guardian*, 5 February 2013, https://www.theguardian.com/commentisfree/2013/feb/05/can-googling-be-racist.

[45] C. Cadwalladr, "Google, democracy and the truth about internet search", *The Observer*, 4 December 2016, https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook.

[46] N. Kayser-Bril, "Google apologizes after its Vision AI produced racist results", Algorithm Watch, 7 April 2020, https://algorithmwatch.org/en/story/google-vision-racism/.

[47] A. Hern, "Twitter apologises for 'racist' image-cropping algorithm", *The Guardian*, 21 September 2020, https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm.

[48] Larsson, *supra*, at 578.

[49] Buchanan & Miller, *supra*, at 33-4.

[50] A. McFarland, "New Study Attempts to Improve Hate Speech Detection Algorithms," Unite AI, 12 July 2020, https://www.unite.ai/new-study-attempts-to-improve-hate-speech-detection-algorithms/.

[51] Efforts in this direction are already underway. The ECtHR recognised homophobic hate speech for the first time in *Vejdeland & others v. Sweden*, Judgment of 9 February 2012. The Committee of Ministers and Parliamentary Assembly of the Council of Europe have also adopted recommendations and resolutions concerning discrimination on the basis of sexual orientation or gender identity, and concerning sexism: Recommendation CM/Rec(2010)5 of the Committee of Ministers to member States on measures to combat discrimination on grounds of sexual orientation or gender identity, 31 March 2010; Parliamentary Assembly of the Council of Europe, Resolution 1728 (2010), "Discrimination on the basis of sexual orientation and gender identity", 29 April 2010, and Parliamentary Assembly of the Council of Europe, Recommendation 1915 (2010), "Discrimination on the basis of sexual orientation and gender identity," 29 April 2010; Recommendation CM/Rec(2019)1 of the Committee of Ministers to member States on preventing and combating sexism, 27 March 2019.

by xenophobia[52] – research shows that racist and homophobic tweets are more likely to be classified by human content moderators as hate speech than sexist ones, which are classified as "simply offensive".[53]

Ensuring that algorithms and human content moderators are trained to detect and remove hate speech on all grounds is critical. So too is ensuring appropriate oversight, transparency and accountability at all stages – from the training through processing through outputs – to expose not only the decisions themselves but also how they were arrived at. This could include, *inter alia*, categorisations of the grounds of hate speech detected and removed, the diversity and representation of the coders/developers, the training provided, the breakdown of users' content affected, and audits of expected versus actual results. Enforceable transparency, explainable AI and human rights impact assessments may be of particular relevance in this context and should be further explored,[54] including along the lines set out in Guidelines 10, 11 and 37 of the OSCE High Commissioner on National Minorities' Tallinn Guidelines on National Minorities and the Media in the Digital Age.[55]

(iii) The Broader Effects on Public Debate

The freedom of political debate is "at the very core of the concept of a democratic society".[56] Nowadays, political and public debate increasingly take place online, via privately-owned and controlled platforms, some of which have achieved positions of global dominance. Such platforms have emerged as powerful gatekeepers that control the flow of information, content and debate online. Heavy reliance on algorithmic and AI-driven content moderation is key to their success. AI may thus be an unseen moderator in the contents and form of such debate, being waged on an unprecedented stage. Further research and consideration are needed to assess the

[52] European Commission, 5th Evaluation of the Code of Conduct, *supra*.

[53] Centre on Regulation in Europe (CERRE), "Issue Paper: Artificial Intelligence and Online Hate Speech," (January 2019), at 7, https://cerre.eu/wp-content/uploads/2020/05/CERRE_Hate-Speech-and-AI_IssuePaper.pdf, citing T. Davidson et al., "Automated Hate Speech Detection and the Problem of Offensive Language," (2017), https://arxiv.org/pdf/1703.04009.pdf. See also Amnesty International, "Twitter still failing women over online violence and abuse," 22 September 2020, https://www.amnesty.org/en/latest/news/2020/09/twitter-failing-women-over-online-violence-and-abuse/.

[54] CERRE, *supra*, at 10; Public Policy Forum, "Democracy Divided: Countering Disinformation and Hate in the Digital Public Sphere," University of British Columbia (August 2018) at 15, https://ppforum.ca/wp-content/uploads/2018/08/DemocracyDivided-PPF-AUG2018-EN.pdf; Council of Europe Commissioner for Human Rights, "Unboxing Artificial Intelligence: 10 steps to protect Human Rights," *supra*; Access Now, "Human Rights in the Age of Artificial Intelligence," (2018), https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf; Access Now, "Submission to the Consultation on the 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust," June 2020, at 3-4, https://www.accessnow.org/cms/assets/uploads/2020/06/EU-white-paper-consultation_Access_Now_June2020.pdf.

[55] See also the accompanying explanations and references in the Explanatory Note to the Guidelines: pp. 39-42 and 68.

[56] *Lingens v. Austria*, ECtHR judgment of 8 July 1986, at para 42.

ways that AI may steer public discourse by promoting or suppressing certain contents, including hate speech.

Similarly, social media have proved an invaluable tool in the arsenal of pro-democracy movements and human rights activists around the world, from the Arab Spring to women's marches to days of action by climate protesters. The use of AI could have widespread (and pernicious) effects on freedom of assembly by "removing groups, pages and content that facilitate organization of in-person gatherings and collaboration".[57] Further research should be undertaken to investigate whether and to what extent AI interferes with organisers' freedom of expression and related rights (such as freedom of association and assembly), and what risks it may pose to democracy.[58]

Algorithms in Facebook determine the contents of users' newsfeeds, while Google indexes content and ranks search results – decisions which significantly affect the news individuals receive and shape the world as they know it.[59] Social media employ algorithms to enable – and indeed incentivize – users' personalization of content and news. Societal proclivities for personalized content and news can lead to the emergence of so-called "filter bubbles",[60] deliberative "fragmentation and cybercascades".[61] Algorithms which sort, filter and recommend certain content may "form the 'terrain' on which harmful speech occurs. Even as these algorithms may have been designed to promote the 'best' content, they can also empower, surface, and aggregate harmful content in ways that platform designers may not have anticipated".[62] Moreover, they may operate as echo chambers, exploiting deliberative fragmentation, driving polarisation and furthering the spread of hateful rhetoric.[63]

Greater attention is now being paid to the role of social media in changing behaviours in insidious and unforeseen ways – so-called "AI-enabled behavioural nudges".[64] The ways that AI may shape the nature of public debate by amplifying or suppressing speech should be the subject of similar scrutiny. The push and pull of algorithmic moderation of content on dominant global platforms has the potential to affect public debate and democratic discourse in profound, if hidden, ways.[65]

---

[57] Access Now (2018), *supra*, at 23.

[58] *Taranenko v. Russia*, ECtHR judgment of 15 May 2014, at para. 70.

[59] Access Now (2018), *supra*, at 23; Council of Europe Committee of Experts on Internet Intermediaries, *supra*, at 10-12.

[60] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You (New York, Penguin Books, 2011).

[61] C. R. Sunstein, *republic.com* (Princeton, Princeton University Press, 2001), Chapter 3 – "Fragmentation and Cybercascades".

[62] *id*.

[63] *id.* See also T. McGonagle, "Minority rights, freedom of expression and of the media: dynamics and dilemmas," Intersentia, Cambridge: 2011; M. Cormack and N. Hourigan (eds.), *Minority Language Media: Concepts, Critiques and Case Studie*s, Multilingual Matters Ltd. (2007), at 157.

[64] Public Policy Forum, *supra*, at 11. See, for instance, "The Social Dilemma" – a Netflix documentary which outlines in worrisome detail the nature and scale of such behaviour modification.

[65] See generally: M. Moore and D. Tambini (eds.), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (Oxford, Oxford University Press, 2018); R. Frank Jørgensen (ed.), *Human Rights in the Age of Platforms* (Cambridge, Massachusetts and London, England, The MIT Press, 2019).

It is particularly worrisome as political speech becomes increasingly polarised – with 'alt-right' groups spouting dangerous content including hate speech, extremist content, and disinformation. Moreover, private corporations – now responsible for defining, detecting and blocking/removing hate speech on their platforms – may be pulled in opposing directions where moderating content is incompatible with profit motivations and the need to drive content.[66]

Conclusion & Recommendations

AI presents a great opportunity to neutralise existing biases and counter the dissemination of networked hate speech. But further work is required if it is to untie the gordian knot of hate speech. The #SAIFE consultation is a timely step in this direction.

This submission explored a range of opportunities and challenges concerning the interplay between free speech, hate speech and AI, and called for further research into three particular areas: (i) the differentiated roles and responsibilities of State and private actors in a broader, human rights-driven and multi-stakeholder approach; (ii) the need for non-discriminatory policies and practices in the design and implementation of measures to detect and block/remove hate speech; and (iii) the broader effects on public debate where AI is used to moderate and censor expression.

Several preliminary general conclusions are evident. First, in light of the technical and substantive hurdles associated with AI usage to combat hate speech, a multidisciplinary and multi-stakeholder approach is required. Stakeholders from the fields of law, technology, policy development and civil society should have a seat at the table to ensure that diverse interests and perspectives are incorporated into the decisions made and standards set. Second, significant work has been done to date which examines the promise and pitfalls of countering hate speech with AI. This should be drawn on – and built upon – to incorporate best practices and lessons learned in any standards or guidance to be developed by the OSCE. Finally, in each of the areas discussed, greater (enforceable) transparency, accountability and oversight are imperative. AI allows for greater secrecy in how expression is moderated and censored; this must be countered in efforts to combat hate speech so that justice is not only done, but is *seen* to be done.

More specifically, further exploration and development of issues raised in this submission may be warranted in the next stage of the #SAIFE project. With respect to the **differentiated roles and responsibilities** of State and private actors, these priority issues concern:

- The contours of states' **negative and positive obligations** in using AI to counter hate speech;

- The **standards, safeguards and limitations** necessary to ensure respect for freedom of expression where private corporations use AI to moderate content on their platforms.

[66] BBC News, "Facebook 'profits from hate' claims engineer who quit," 9 September 2020, https://www.bbc.com/news/technology-54086598.

With respect to the **eradication of bias**, attention should be paid to the following priority issues:

- The policies and practices necessary to ensure that **bias is eradicated at each stage of the lifecycle** of AI;

- The implementation of **enforceable transparency and explainable AI** across all subcategories of hate speech.

Finally, the next stage of the project should explore and develop the following priority issues concerning AI's **broader ramifications**:

- The **impacts on public debate** of using AI to address hate speech by both State and private actors;

- The best and promising practices to, on the one hand, **mitigate these effects** on public discourse and debate and, on the other hand, ensure **effective protection of minorities** against hate speech.