cerre Centre on Regulation in Europe

ISSUE PAPER

January 2019

Michèle Finck

ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH

The event, for which this Issue Paper has been prepared, has received the support and/or input of the following CERRE members: AGCOM, Facebook and Ofcom. As provided for in CERRE's by-laws, this Issue Paper has been prepared in strict academic independence. At all times during the development process, the author, the CERRE Academic Team and the Director General remain the sole decision-makers concerning all content in the Paper.

ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH
Michèle Finck

January 2019

info@cerre.eu
www.cerre.eu

ARTIFICIAL INTELLIGENCE
AND ONLINE HATE SPEECH

I. INTRODUCTION

Online hate speech is widely recognised as a societal problem.[1] Yet, defining what exactly amounts to hate speech is no easy task. There are no clear legal criteria to distinguish between speech that might be offensive or hurtful but protected under freedom of expression, and speech that is unlawful because it, in fact, qualifies as hate speech. In the absence of a clear-cut legal definition, the identification of speech as hate speech is fraught with difficulty, as evidenced by the judicial assessment of such matters in past decades, particularly as the amount of content shared online is growing steadily and expected to continue to grow over the coming years.

To date, the identification of hate speech online and removal thereof by human content moderators has been burdensome. In light of the sheer amount of data to be reviewed, the required investments in human resources are significant. Further, in light of the lack of clear criteria allowing for the identification of hate speech, human discretion means that oftentimes speech that is merely offensive but not hate speech is banned from platforms in contravention of the right to freedom of expression. Empirical research has moreover underlined the considerable psychological strain weighing on human content moderators.

Against the background of a more general period of Artificial Intelligence ('AI') enthusiasm, AI, in particular machine- and deep-learning are widely perceived as desirable innovations in this domain. The automated detection (and maybe also deletion) of hate speech would be a scalable solution to manage ever-growing amounts of online content, reduce costs and decrease human discretion in this process.

This Issue Paper provides an overview of related opportunities and challenges. It first documents the problem of online hate speech and the shortcomings of current forms of human-based content moderation processes before introducing the potential of machine- and deep-learning, highlighting that AI may trigger important efficiency gains in this area. At the same time, however, there are also considerable weaknesses associated with current forms of AI, most importantly its over-inclusiveness which causes considerable problems from the freedom of expression perspective. The paper will consider if and how future developments in artificial intelligence may address some of these issues and the paper closes with suggestions of themes for future discussion.

[1] https://www.europol.europa.eu/activities-services/main-reports/european-union-terrorism-situation-and-trend-report-2018-tesat-2018 For an overview, see Kai Kaspar et al, *Online Hate Speech* (Kopaed Verlag 2017).

II. ONLINE HATE SPEECH

Online hate speech can, in essence, be defined as online speech that attacks a person or a group on the basis of certain attributes such as race, ethnic origin, religious affiliation, disability, gender or sexual orientation.[2] As the amount of online content continues to rapidly increase, so does online hate speech. Recent research and media attention have highlighted related problems. It is, for instance, increasingly apparent that social media are effectively used to promote extremist causes.[3] Research has moreover shown that Twitter is used for jihadist hate speech as well as right-wing hate speech.[4] For example, a UN Independent Fact-Finding Mission on Myanmar reported in 2018 that social media, in particular Facebook, had played a 'determining role' in the human rights violations committed against the Rohingya population in spreading disinformation and hate speech.[5]

Where definitions of hate speech do exist, they are broad and under-inclusive, such as Article 20(1) of the ICCPR according to which '[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law'. Hate speech, however, denotes a 'broad spectrum of extremely negative discourse stretching from hatred and incitement to hatred; to abusive expression and vilification; and arguably also to extreme forms of prejudice and bias'.[6] A consequence of the lack of a uniform definition of hate speech in the EU is that there are no clear-cut parameters that the private sector can rely on to identify what is and isn't hate speech, resulting in considerable uncertainty and discretion in law-enforcement.[7] It is worth noting that even in jurisdictions that have adopted specific legislation to combat online hate speech, the definitional problem remains. The German *Netzwerkdurchsetzungsgesetz* ('NetzDG') for instance refers to '*Hasskriminalität*' (hate crime) and '*offensichtlich rechtswidrigen Inhalte*' ('manifestly illegal content').

Given that hate speech is unlawful, online platforms are required to remove related content where it has been identified. It is important to note that there is no general obligation for platform providers to systematically screen content in light of the E-Commerce Directive's hosting exemption.[8] Under Article 14 of the E-Commerce Directive, information society service providers that simply store information provided by users are not liable for such content if they (i) do not have active knowledge of illegal activity or information, and (ii) act expeditiously to remove or to disable access to the information after becoming aware of it.[9] This means that while platforms are not required to systematically monitor content, they must verify whether certain content amounts

[2] See further EU Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178
[3] https://www.hsgac.senate.gov/imo/media/doc/Bergen%20Testimony_PSI%202016-07-06.pdf
[4] See Tom De Smedt et al, Automatic Detection of Online Jihadist Hate Speech (2018) CLiPS Technical Report 7, https://www.uantwerpen.be/images/uantwerpen/container2712/files/hate-speech-detection.pdf and Sylvia Jaki and Tom De Smedt, Right-Wing German Hate Speech on Twitter: Analysis and Automatic Detection (2018), http://organisms.be/downloads/jaki2018.pdf
[5] "Fact-finding Mission on Myanmar: concrete and overwhelming information points to international crimes." March 12, 2018. http://www.ohchr.org/EN/HRBodies/HRC/Pages/NewsDetail.aspx?NewsID=22794&LangID=E
[6] Tarlach McGonagle, The Council of Europe Against Online Hate Speech: Conundrums and Challenges Expert Paper (2013), 4. See also Robert Post, "Hate Speech", in Ivan Hare and James Weinstein, Eds., Extreme Speech and Democracy (New York, Oxford University Press, 2009) 123.
[7] For an overview of EU guidance on this matter, see further https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms.
[8] See also Article 15(1) of the E-Commerce Directive.
[9] See also Alexandre de Streel, Miriam Buiten and Martin Peitz, Liability of Online Hosting Platforms. Should Exceptionalism End? (CERRE 2018). https://www.cerre.eu/publications/liability-online-hosting-platforms-should-exceptionalism-end

to hate speech where it is flagged as such by others, and must remove it where they conclude that the identified speech in fact crosses the hate speech threshold. [10]

Traditionally, human content moderators have been in charge of verifying whether reported content is in fact illegal. This has not been without difficulty. First, the amount of content produced continues to grow steadily, having forced economic operators to make significant investments in human resources. To illustrate, Facebook alone has been reported to have hired over 20,000 workers to detect hate speech on its platform.[11] YouTube is said to be employing over 10,000 people that check whether content violates its policies.[12] Related costs are of particular concern for smaller players that may struggle to deal with the resulting budgetary strain. Particularly in light of the lack of a legal definition of hate speech, human discretion can lead to the false labelling of content. Opaque criteria are used by humans to determine what is and isn't hate speech.[13] Content moderation is, moreover, usually carried out by subcontractors in low-wage jurisdictions that apply appropriateness criteria 'that are often ambiguous and culturally-specific'.[14] In light of these difficulties and the steadily increasing volume of online content, artificial intelligence is increasingly explored as a complement or even replacement of human analysis.[15]

III. USING AI TO IDENTIFY ONLINE HATE SPEECH

In essence, machine learning refers to a process whereby an algorithm is trained on training data to identify patterns in datasets. The resulting model can then be applied to new data to detect the same patterns. In relation to hate speech, an algorithm could be trained on an existing data set where hate speech is clearly labelled, and then applied to new datasets (newly generated content) to determine whether similar instances of hate speech are present. The idea is that through the use of machine learning models, such as natural language processing, hate speech can be automatically detected.[16] Deep-learning analyses, such as convolutional neutral networks, are also used to make advances in this field.[17] This has led all major platforms to experiment with AI for content moderation purposes.[18] Public actors also increasingly see the appeal of such solutions – in the UK, a machine learning tool is used to automatically detect propaganda produced by the Islamic State terror group with the idea that such content could be blocked before it is uploaded to platforms.[19]

---

[10] See also Commission Recommendation 2018/334 L 63/50 (2018), in particular paras 20 and 28.
[11] Issie Lapowsky, 'Facebook Moves to Limit Toxic Content as Scandal Swirls' (Wired, 15 November 2018), https://www.wired.com/story/facebook-limits-hate-speech-toxic-content/
[12] Sam Levin, Google to Hire Thousands of Moderators after Outcry over YouTube Abuse Videos (The Guardian, 5 December 2017), https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos
[13] Andrew Arsht and Daniel Etcovitch, The Human Cost of Online Content Moderation, Harvard Law Review Online (March 2, 2018) https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation
[14] Ibid.
[15] Note, however, the legal limits to replace human with technical decision-making under Article 22 GDPR. See also Para 20 of Commission Recommendation 2018/334.
[16] Anna Schmidt & Michael Wiegand, 'A Survey on Hate Speech Detection Using Natural Language Processing', Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (2017) http://www.aclweb.org/anthology/W17-1101
[17] See, by way of example, Björn Gambäcl and Utpal Kumar Sikdar, 'Using Convolutional Neural Networks to Classify Hate-Speech' (2017) Proceedings of the First Workshop on Abusive Language Online 85.
[18] See, by way of example, https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html
[19] https://techcrunch.com/2018/02/13/uk-outs-extremism-blocking-tool-and-could-force-tech-firms-to-use-it/

Appropriately designed models offer various advantages for the detection of hate speech. Unlike humans, these technical systems are scalable and moreover bring the promise of absolving workers from the psychological strain that comes with content moderation.[20] This could result in considerable cost savings and speedier decisions. Further, while there is always discretion in model design and the underlying training data, inter-personal forms of discretion in a system where different humans evaluate content would be removed as the same model could be applied for all forms of content in a specific language and jurisdiction.

At the same time, a number of shortcomings can be identified in relation to current forms of AI. Whereas machine learning models are good at spotting nudity or sexual activity, they have proven to be much less efficient in detecting hate speech.[21] Of the 2.5 million pieces of hate speech removed from Facebook in Q1 2018, only 38% was flagged by technology beforehand.[22] This underlines that algorithmic tools are not yet able to understand context. Indeed, while for nudity or sexual activity context doesn't matter (it is per se prohibited on most platforms), for hate speech it does, as the same words can have vastly divergent meanings depending on context. Machine learning models are unable to understand irony or satire or realise that hate speech can be used to raise awareness (such as where someone describes hate speech they have witnessed or been subject to). To illustrate, commentators have noted that Google's 'Perspective API' (an API using machine learning to identify hate speech) identified expressions such as 'garbage truck' or 'few Muslims are a terrorist treat', 'you are no racist' as well as 'I fucking love you man. Happy birthday' as toxic speech.[23]

This highlights that there is a significant risk of over-blocking in relation to current forms of AI. Models are susceptible to false positives as they fail to distinguish between hate speech and offensive ordinary speech, or simply words that can be used to offend, but also in entirely harmless manners.[24] Beyond this, detection techniques have proven brittle against adversaries that (automatically) insert typos, change word boundaries or add innocuous words to hate speech.[25] Research has indeed highlighted that simply adding the word 'love' to expressions that otherwise qualify as hate speech makes them go undetected by machine learning models.[26]

A further complication stems from the fact that hate speech changes over time.[27] Models are however trained on historical data that may not yet be able to catch current forms of hate speech. Language is also an important factor, as separate learning models need to be trained for each language. There is, however, more training data for some languages than for others and some have in the past been neglected.[28] AI hate speech detection tools are accordingly harder to develop

[20] Tarleton Gillespie *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (*Yale University Press 2018)
[21] See further https://newsroom.fb.com/news/2018/05/enforcement-numbers/
[22] Ibid.
[23] https://www.perspectiveapi.com/#/; https://towardsdatascience.com/why-alphabets-ai-cannot-fix-hate-speech-8d352892cdba
[24] Tommi Gröndahl et al, 'All You Need is « Love »: Evading Hate Speech Detection (2018) https://arxiv.org/pdf/1808.09115.pdf, 1.
[25] Ibid.
[26] Ibid.
[27] https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/
[28] To prevent further instances of this problem, Facebook announced that it would develop its AI and hire more Burmese-language editors. See further Andy Sullivan, Yimou Lee. "Myanmar activists welcome Zuckerberg's 24-hour target to block hate speech on Facebook." April 10, 2018. https://www.reuters.com/article/us-facebook-privacy-myanmar/myanmar-activists-welcome-zuckerbergs-24-hour-target-to- block-hate-speech-on-facebook-idUSKBN1HI028

in languages less used on the platform.[29] A further complication is the application of different legal standards, even where the same language is used. Indeed, it might be unwise for companies to simply train one model to detect hate speech on all English-language expressions as freedom of expression covers a much broader range of offensive speech in the United States than in the United Kingdom.[30]

In light of the highly contextual nature of hate speech it has proven particularly difficult to appropriately label training data as hate speech (to train the model, which would then be able to identify hate speech when applied to new datasets).[31] This results in models being under-inclusive or over-inclusive. Where they are under-inclusive, they fail to detect speech that in fact is hate speech, which is likely to trigger liability for the platform concerned. Where a model is over-inclusive, this is particularly problematic from the perspective of freedom of expression.

IV. THE RISKS OF UNDER-OR OVER-INCLUSIVE AI MODELS

Where machine learning models are calibrated to be under- or over-inclusive, different sets of problems emerge. Where algorithmic detection tools are under-inclusive, instances of hate speech will not be qualified as such. This is problematic from users' perspective as unlawful behaviour goes unpunished and the addressees of hate speech have to continue suffering the consequences thereof. Under-inclusiveness is also problematic from the perspective of the operator as they are likely to suffer reputational damage and also financial consequences where enforcement action is thereafter taken in the judicial system.

This explains why the private sector is exposed to considerable reputational and financial incentives to design automated detection models in an over-inclusive fashion. This, however, is problematic from the perspective of freedom of expression. Research has indeed pinpointed that supervised machine learning often fails to distinguish between offensive speech and hate speech.[32] Further, where crowd-sourcing has been used to label data appropriately, racist and homophobic tweets were more likely to be classified as hate speech than sexist tweets, which were rather classified as simply offensive.[33]

This problem has been highlighted in relation to the German NetzDG legislation that requires platforms to remove unlawful content within seven days where this is signalled by users, something that must happen within 24 hours where the content is 'manifestly' unlawful. Companies face fines up to €50 million where they fail to comply.[34] It has been stressed that the threat of reputational and financial damage incentivises platforms to remove content in case of doubt, with negative implications on freedom of expression. Even though fines only apply in instances of systemic non-compliance, this has been said to push companies towards shaping their systems in a manner that is generally restrictive of many forms of speech, not just hate speech.[35] Indeed, in implementing the German NetzDG, Twitter blocked a satirical magazine that had parodied anti-

[29] https://newsroom.fb.com/news/2018/05/enforcement-numbers/
[30] This is due to the expansive scope of the First Amendment to the United States constitution.
[31] Njagi Dennis Gitari et al, A Lexicon-based Approach for Hate Speech Detection (2015) 10 International Journal of Multimedia and Ubiquitous Engineering 215, 216.
[32] Thomas Davidson et al, 'Automated Hate Speech Detection and the Problem of Offensive Language' (2017) https://arxiv.org/pdf/1703.04009.pdf 1.
[33] Ibid.
[34] It is worth noting that small social networks are excluded from its scope of application.
[35] https://www.uni-muenster.de/news/view.php?cmdid=9429

Muslim comments made by the AFD.[36] The resulting takedowns of content not only impact freedom of expression but also the right to maintain a private life and private communications.[37] Even where an algorithm can detect certain information with a 99.995 percent accuracy and thus only has a false positive rate of 0.0005 percent, this would mean that out of the more than 1 million pieces of content produced on Facebook each day, 15,000 would be falsely flagged.[38]

As such content would then be removed, significant concerns emerge from the perspective of the right to freedom of expression as it is protected in Member States' national legal orders, Article 11 of the EU Charter of Fundamental Rights and the European Convention on Human Rights. Article 10 ECHR (to be read in conjunction with Article 17 ECHR) relates to the freedom of expression. Article 10(1) clearly states that this right includes the 'freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers'. This freedom can be limited for reasons prescribed law and necessary in a democratic society under Article 10(2) ECHR, such as the prevention of hate speech. However, European human rights law is also clear on the fact that expression that may 'offend, shock or disturb' is covered by the right to freedom of expression.[39] As a consequence, a nuanced understanding of contextual settings is necessary to determine what crosses the tipping point to hate speech, something that neither algorithms nor humans without the opportunity to carefully review can do.

It has been suggested that 'AI can't understand the context of speech and, since most categories for problematic speech are poorly defined [by necessity], having humans determine context is not only necessary but desirable'.[40] As linguistic nuances continue to exceed AI's current capabilities, related techniques may be used as a tool in helping to detect violations rather than do so on their own. Unlike humans, algorithms cannot read between the lines (just as humans that don't have sufficient contextual understanding or adequate time to make such assessments). As current solutions work on the basis of flagging certain words, they should not, at this moment in time, be used as a standalone solution to online hate speech.

To illustrate, Google's Perspective API, a tool developed to combat hate speech, has been trained only to identify 'toxic' speech.[41] The word 'moron' is thus seen as toxic speech although it likely wouldn't qualify as hate speech.[42] The lack of clear-cut criteria that can be applied to determine whether something is constitutive of hate speech thus burdens AI developers' task. It also results in significant discretion for the private sector. When it comes to content moderation, private and often non-transparent decision-making processes determine what speech is lawful and what speech isn't.

Platforms asked to identify and take down content that amounts to hate speech must fulfil traditional State functions in two manners. First, in the absence of a legal definition of hate speech

[36] https://techcrunch.com/2018/01/09/europe-keeps-up-the-pressure-on-social-media-over-illegal-content-takedowns/
[37] https://privacyinternational.org/blog/1111/two-sides-same-coin-right-privacy-and-freedom-expression
[38] https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/
[39] Handyside v. the United Kingdom, Judgment of the European Court of Human Rights of 7 December 1976, Series A, No. 24, para. 49.
[40] Drew Harwell. "AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how." April 11, 2018. https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark- zuckerberg-says-just-dont-ask-when-or-how/?utm_term=.85a6138cdb26; Larry Greenemeier. "Can AI Really Solve Facebook's Problems?" April 13, 2018. https://www.scientificamerican.com/article/can-ai-really-solve-facebooks-problems1/.
[41] https://www.perspectiveapi.com/#/
[42] https://www.perspectiveapi.com/#/

they are compelled to independently define this concept – and label training data accordingly. Second, they need to adjudicate whether something matches that definition and must accordingly be removed where content has been flagged by users. This matches the traditional State prerogatives of law-making and enforcement. The resulting power-shift has been widely criticised.

To illustrate, Facebook has adopted its own definition of hate speech as 'a direct attack on people based on protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, and serious disability or disease. We also provide some protections for immigration status. We define an attack as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation'.[43] Twitter, on the other hand, has invented the altogether new concept of 'hateful conduct' which is considered to occur where a user promotes 'violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.'[44]

While making this publicly available is a laudable effort towards transparency, it also underlines that corporate definitions of hate speech may diverge not only between providers but also from established legal concepts. It has indeed been stressed that 'intermediaries, as private entities, are not best placed to make the determination of whether particular content is illegal, which requires careful balancing of competing interests and consideration of defences'.[45] Despite these concerns, the code of conduct adopted by the European Commission together with Microsoft, Facebook, Twitter and YouTube in 2016 that includes a series of commitments to combat the spread of illegal hate speech online in Europe delegates enforcement to these actors, leaving them with difficult interpretational choices.[46]

The UN special rapporteur on freedom of expression indeed concluded that 'the private sector has gained unprecedented influence over individuals' right to freedom of expression and access to information'.[47] To avoid infringements to the right to freedom of expression and the right to privacy of Internet users, his report recommended that restrictions should only be implemented after judicial intervention.[48] This, however, seems difficult to implement in the hate speech context. Vast amounts of content are generated each day, some of which amounts to hate speech. Waiting for a judicial intervention would be extremely time-consuming meaning that hate speech victims have to endure related consequences as content remains online and such solutions are also hard to operatise at scale.

It thus appears that alternative governance and decision-making processes are needed that are capable of combining the speed and efficiency of private and partly-algorithmic enforcement with the importance of transparency, human rights protection and public oversight. Some are experimenting with solutions to these issues. The Online Hate Index developed at the University of Berkeley is an effort to incorporate users' views into the definition of what amounts to hate

[43] https://transparency.facebook.com/community-standards-enforcement#hate-speech
[44] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[45] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue (2011), https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf, p.12.
[46] http://europa.eu/rapid/press-release_IP-16-1937_en.htm
[47] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue (2011), https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf, p.13.
[48] Ibid,14.

speech.[49] Future research should focus on possible governance models that enable the combination of these binary objectives. It is for instance worth reflecting on what polycentric governance processes and user involvement could add to present processes. Further, the ongoing sophistication of AI should be borne in mind. Of particular relevance in this context may be the potential of explainable AI ('XAI'), which refers to the design of more transparent algorithmic decisions that expose how they make decisions.[50]

V. THE USE OF AI AS A SOURCE OF FURTHER REGULATORY COMPLEXITY?

The use of AI – to combat online hate speech and beyond – raises a host of additional regulatory questions. Indeed, the mere development of efficient machine learning models with the capacity to detect online hate speech and allow companies to move beyond the currently still human-centred detection mechanisms presupposes that a company has (i) access to the relevant training data, and (ii) the required data analytics expertise (which can be in-house or procured externally).

However, not all companies are created equal in relation to their ability to gather the necessary training data and develop adequate machine learning models on the basis of them. These issues of course relate to a broader ongoing debate about the implications of AI for businesses, regulation and more generally the Digital Single Market.[51] This raises the question of whether incentivising companies to rely on artificial intelligence for online hate speech detection could have undesirable economic effects. It could indeed be speculated that there are important divergences regarding the suitability of AI or human-centred detection models for large and small players respectively. There may be reason to believe that large players are in a better position to develop such adequate models in light of their data treasures and in-house expertise in data analytics. Similarly, new market entrants may be also be in a position of disadvantage in this respect. The relative ease of regulatory compliance for economic actors that are well-positioned in relation to data and data analytics may thus be a source of a competitive advantage for them *vis-à-vis* competitors. On the other hand, however, it may also be that sophisticated AI tools could be of particular benefit for smaller players, which may be unable to hire an army of human-content checkers, but could afford adequate third-party detection tools (or related expertise to build their own).

These issues highlight the close link between the topic of the online detection of hate speech and broader debates regarding, inter alia, access to data in the European Union.[52] Further economic research is needed in this area to gain a better impression of the related market dynamics, which should be considered when devising policy options in this respect. Ultimately each system will require a balancing between various considerations and trade-offs, yet further research could shed light on some of these dynamics, which presently still remain a matter of speculative debate.

In addition, it would also be worth considering the specific legal questions that are raised by the use of automated detection models. For example, it appears that Article 22 GDPR, which governs decisions reached through solely automated means, that produce legal effects or otherwise significantly affect an individual, may also come into play in at least those circumstances where there is no meaningful assessment of the AI's categorisation by a human.

---

[49] https://www.adl.org/resources/reports/the-online-hate-index
[50] Note that this may also be required under the EU's General Data Protection Regulation.
[51] European Commission, Artificial Intelligence for Europe (25 April 2018), https://ec.europa.eu/digital-single.../en/.../communication-artificial-intelligence-europe
[52] See also the Commission's Proposal for a Regulation on the Free Flow of Personal Data : https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-495-F1-EN-MAIN-PART-1.PDF

VI. QUESTIONS FOR DISCUSSION AND FURTHER RESEARCH

There can be no doubt that, in its current form, machine and deep learning techniques are not a panacea for online hate speech. Yet, as Brittan Heller, director of the US Anti-Defamation League's Center for Technology and Society, has argued, 'just because AI does not solve the problem entirely doesn't mean it's useless'.[53] Learning models can indeed fulfil a valuable role in the detection of hate speech where results are subsequently verified by a human. Indeed, assistance through AI may enable humans to dedicate more time to border cases where it isn't clear whether something is or isn't hate speech.

The near to mid-term future is thus likely one of hybrid models that combine the advantages of algorithmic decision-making with human context awareness. For example, the German *Landesanstalt für Medien Nordrhein Westfalen* recommends the use of 'human-machine-filters' – a combination of humans and algorithms whereby algorithms carry out the first stage of processing, and humans thereafter check results.[54] The debates surrounding this issue however also highlight numerous open questions regarding the use of these techniques in relation to hate speech on platforms, as well as uncertainties regarding their future potential.

[53] https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/
[54] Leif Krampf & Stefan Weichert, Hasskommentare in Netz. Steuerungsstrategien für Redaktionen, Landesanstalt für Medien NRW (2018)

Questions for further discussion

SESSION 1

- To what extent are machine learning techniques currently being used to filter online hate speech? What are the costs of those techniques? Do the latter tend to over-intervene?
- In the future, what is the potential for developments in AI such as sophisticated forms of deep learning for hate speech detection? Can AI ever understand context and if yes, in what time frame?
- What is the potential of explainable AI ('XAI') in this domain?

SESSION 2

- Is it at all possible to provide a legal definition of a context-dependent and evolving notion such as hate speech? Is there a need for a common European definition of hate speech?
- Should the difficult balance between human rights be left to algorithms? How can we ensure the best balance between machines and humans in policing hate speech?
- What are the likely economic implications of using automated rather than human-centered detection tools?

SESSION 3

- Do we need new governance models that strike an adequate balance between public and private intervention?
- Is co or self-regulation through AI an appropriate means of enforcing public policy objectives? What has been the practical impact of the Code of Conduct initiated in the EU in 2016?
- What should be the role of the next EU legislature (2019-2024) to ensure to fight online hate speech in a manner which is effective and legitimate? In that regard, should the e-commerce Directive be reviewed?

ABOUT THE AUTHOR

Michèle Finck is a CERRE Research Fellow and a Senior Research Fellow at the Munich-based Max Planck Institute for Innovation and Competition. Since 2013, she is also a lecturer in EU law at Keble College (University of Oxford).

Her research focuses primarily on digital platforms, artificial intelligence and blockchain technology as well as EU law. She studied law at King´s College London, the Sorbonne and the Florence-based European University Institute where she obtained an LLM degree. She also has a doctorate in law from the University of Oxford.

cerre

Centre on Regulation in Europe

Avenue Louise, 475 (box 10)
1050 Brussels, Belgium

+32 2 230 83 60

info@cerre.eu

cerre.eu

@CERRE_ThinkTank