Implications of Artificial Intelligence Content Moderation on Free Speech: Regulating Automated Content Moderation Under International Human Rights Law Through A Comparative Lens

By Mostafa Abdelaziz Ashhab Elkadi

Submitted to the Department of Legal Studies, Central European University

In partial fulfillment of the requirements for the degree of Masters of Law in Human Rights

Thesis supervisor: Professor Sejal Parmar

Vienna, Austria

2021

CEU eTD Collection

Abstract

This thesis is aiming to answer the two following questions of 1) does the usage of Artificial Intelligence and machine learning algorithms for removing online speech comply with Freedom of Expression under International Human Rights Law? The presumed answer for this is no, so through checking the recent developments through mostly soft law, the thesis is aiming to answer the second question of 2) how can Artificial Intelligence filters comply with Freedom of Expression under International Human Rights Law?

The methodology used in the research is relational between law and technology; it takes both a normative as well as a comparative approach. The main focus is based on defining Artificial Intelligence filters, while also looking to conduct content moderation in a more acceptable manner under International Human Rights Law while allowing self-regulation for social media platforms.

Liability regimes of Internet Service Providers are also tackled and analyzed, and through different liability regimes one can witness how it can incentivize social media platforms in using Artificial Intelligence filters.

CEU eTD Collection

Acknowledgments

I would like to start off by thanking my supervisor Prof. Sejal Parmar for being supportive and understanding throughout the entire process of my thesis.

I would also like to seize this opportunity to thank my friends and family back in my home country, especially my dad and Khalid Abdelfattah for always supporting me, and believing that I can achieve anything no matter how difficult it is, but I have to say that without their support I would not have been able to finish this LLM degree.

I could go on and on with thanking people for their support, but at this point I acknowledge and thank my friend and my former team mate, co-worker, class mate, and at some point was my former head as well Ibrahim Sabra for his pragmatic assistance and pointers.

Yet the most dedicated thank you goes to Sophia Fehrenbach for all her kind gestures and her emotional support throughout the entire year, thank you very much Sophia, you really made a difference.

CEU eTD Collection

Table of contents

CEU eTD Collection

CEU eTD Collection

Preliminary Chapter

1) Thesis objective

The thesis will examine measures taken by social media platforms when it comes to content moderation, more specifically automated filtering of content, and identify the best practices. The main focus will be trying to balance between how can social media platforms use Artificial Intelligence filters while still remaining within the ambit of Freedom of Expression.

For this reason, the focus lies more on Freedom of Expression, and censorship through Artificial Intelligence filters, rather than just tackling all of the issues caused by Artificial Intelligence over social media platforms.

Therefore, the thesis will not be tackling other rights or phenomena that arise due to Artificial Intelligence systems on social media platforms, thus issues relating to the right to non-discrimination, the right to privacy, freedom of assembly, and any other right will not be tackled in this paper.

The main aim is to reach the appropriate limitations that should be set for the usage of Artificial Intelligence and machine learning algorithms for filtering content under International Human Rights Law, and how regional systems are tackling the same issue, while also tackling the shortcomings from the law by providing recommendations.

2) Legal issue

The main question that is intended to be answered is: does the usage of Artificial Intelligence and machine learning algorithms for removing online speech comply with Freedom of Expression under International Human Rights Law?

CEU eTD Collection

From this one question, there is the assumption that the answer is no, it does not comply with International Human Rights Law. Thus, one would have to tackle the follow-up question, and it would be: how can Artificial Intelligence filters comply with Freedom of Expression under International Human Rights Law?

There is also the scenario of the typical answer of "It depends", which is also a valid answer since it is still not clear what Artificial Intelligence systems are capable of. Yet for the purposes of this research paper, the answer to the initial question is no, thus the follow-up question is going to be tackled in detail.

3) Research methodology

This will be a relational research paper, between law and technology. The research is going to follow a normative method of research, thus tackling the evolution of the concept of Freedom of Expression from the offline sense to the online regulations. The research will then focus on how differently two regional systems on Human Rights tackled the topic, thus it will follow a comparative approach as well.

Further, a doctrinal method will be initiated, as this method requires and analysis of legal provisions by tackling domestic and international case law, statutes, conventions, soft law, etc. However, the topic has not been tackled explicitly under legal provisions. Thus, the analysis will be drawn from mostly secondary sources and domestic decisions.

Further, due to the lack of resources, the thesis will only be tackling the issue from the perspective of the most common practices or International law, more specifically Article 19 of the International Covenant on Civil and Political Rights, Article 13 of the American Convention on Human Rights, and Article 10 of European Convention on Human Rights. The African

CEU eTD Collection

system unfortunately will not be tackled, as there are only two documents that can be tackled within the context of AI and Freedom of Expression.[1]

4) Thesis Structure

The thesis will comprise of 3 chapters. The first chapter will be more of a normative chapter, displaying the framework for Freedom of Expression, filters and Artificial Intelligence from the International Law perspective, if not from an International Law perspective, then from the perspective of the most followed practices. The first chapter will also include the technology of how Artificial Intelligence filters operate, and what is the driving factor behind the usage of those filters.

The second chapter will tackle the current framework set by the Inter-American System on Human Rights when it comes to Freedom of Expression interplay with Artificial Intelligence filters, more specifically the focus will lie on content regulation as well as intermediary liability, and how those two could influence or steer away from social media platforms from using Artificial Intelligence filters, all through different forms of either soft law or hard law.

Yet it is to be noted that in the Inter American System some citations will refer to joint declarations or common documents between different Human Rights systems, and the researcher is aware that those documents are common between different systems, yet they are an integral part of shaping the law, and they are part of the Inter American System framework, and given the

[1] '473 Resolution On The Need To Undertake A Study On Human And Peoples' Rights And Artificial Intelligence (AI), Robotics And Other New And Emerging Technologies In Africa - ACHPR/Res. 473 (EXT.OS/ XXXI) 2021' (Achpr.org, 2021) <https://www.achpr.org/sessions/resolutions?id=504> accessed 10 June 2021; DECLARATION OF PRINCIPLES ON FREEDOM OF EXPRESSION AND ACCESS TO INFORMATION IN AFRICA Adopted by the African Commission on Human and Peoples' Rights at its 65th Ordinary Session held from 21 October to 10 November 2019 in Banjul, The Gambia.

CEU eTD Collection

lack of resources on this topic from the Inter American System, those common documents are essential for the Inter American System.

The third chapter will tackle the Framework set by the European system, through both the European Union and the Council of Europe, as they both complement each other in this area. The research will go through soft law and hard law documents as well as through jurisprudence, for how regulating Artificial Intelligence filters ought to be. It is noteworthy to mention that the European system has its advancements, not just within the area of filters, but in general when it comes to Artificial Intelligence. However, the focus will also be on content moderation, intermediary liability, and how the system is pushing for the usage of Artificial Intelligence filters over social media platforms.

The final part labeled as conclusion and recommendations will be divided into two parts. Firstly a conclusion that draws analysis between international law, the two systems, and analysis between the two systems themselves, as well as domestic decisions or laws that one can argue that it would be better than the currently adopted methods. Secondly, the recommendations part will mostly be based on the conclusions reached throughout the thesis, yet some of the recommendations will be based on outside of the box solutions that were not necessarily tackled throughout the thesis.

In sum, the last section will be an analysis of the best practices that could be followed in order to fully protect the right to freedom of expression.

CEU eTD Collection

Chapter 1: Freedom of expression integration with technology

1) Freedom of expression under International Human Rights Law

Freedom of expression ("FoE") is a fundamental right that is guaranteed by Article 19 of both the Universal Declaration of Human Rights ("UDHR")[2] and the International Covenant on Civil and Political Rights[3] alongside with other regional instruments ("ICCPR").[4] There are also specific guidelines that are ought to be followed while interpreting Article 19 of the ICCPR, this interpretation can be found in General Comment number 34.[5]

FoE is of utmost importance[6]. It does not only encompass the right of the individual to impart information and ideas, but it also has a societal aspect whereby the public is entitled to receive them[7] and includes the ideas "that offend, shock or disturb the State or any sector of the

[2] Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III), Art. 19.

[3] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976), Art.19.

[4] European Convention on Human Rights (adopted 4 November 1950, entered into force 3 September 1953) art. 10; ACHR (adopted 22 November 1969, entered into force 18 July 1978) Art.13; African Charter ON HUMAN AND PEOPLES RIGHTS (ADOPTED IN NAIROBI JUNE 27, 1981, ENTERED INTO FORCE OCTOBER 21, 1986), art.9.

[5] UNCHR 'General Comment 34' In 'Article 19 (Freedom Of Opinion And Expression)' (2011) UN Doc CCPR/C/GC/34

[6] Handyside v. The United Kingdom App no 5493/72 (ECtHR, 7 Dec 1976), para. 49; Venice Commission 'Report on the relationship between FoE and freedom of religion' 76th Plenary Session(2008), Doc No (CDL-AD(2008)026), para .43; UNGA, Calling of an International Conference on Freedom of Information (1946), UN Doc A/RES/59; Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism, Advisory Opinion OC-5/85, IACtHR Series A No 5 (13 November 1985) para.70; Media Rights Agenda and Others v Nigeria, Communications 105/93, 128/94, 130/94 and 152/96, African Commission on Human and Peoples' Rights 12th Annual Activity Report (31 October 1998); Ekmekgjian v. Sofovich, Supreme Court of Argentina (7 July 1992) as cited in The ARTICLE 19 FoE Handbook, August 1993 page 66.

[7] Herrera-Ulloa v. Costa Rica Judgment, Series No.107 (IACtHR 2 July 2004); Ivcher-Bronstein v. Peru Judgment (IACtHR 6 February 2001) para.148; UNCHR 'General Comment 34' In 'Article 19 (Freedom Of Opinion And Expression)' (2011) UN Doc CCPR/C/GC/34, para. 20; Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism, Advisory Opinion OC-5/85, IACtHR Series A No 5 (13 November 1985),

CEU eTD Collection

population"[8] as it would be "unreasonable to restrict freedom of expression only to generally accepted ideas"[9]. This is important since FoE best serves its purpose when it "induces a condition of unrest, creates dissatisfaction with conditions as they are, or even stirs people to anger".[10]

Included within the ambit of freedom of expression is the freedom of political expression which lies at "the very core of the concept of a democratic society"[11]. In particular, dissenting opinions make for a healthy political climate[12] and thus restrictions that stifle political debate call for a closer scrutiny on the part of the states.[13]

Further, with the rise of the digital age, it has been noted that FoE is protected both online and offline.[14] And when it comes to corporate responsibility, some companies do apply the principles of human rights in their operations.[15]

---

para.30; Supreme Court of Sri Lanka, M Joseph Perea and Ors vs. Attorney General App. No. 107-109/86, 25 May 1987; Madanhire and another v. Attorney General, Judgment No. CCZ 2/14, Zimbabwean Constitutional Court, 12 June 2014, para. 7; Lingens v. Austria App no 9815/82 (ECtHR, 8 July 1986), para.41.

[8] Bladet Tromso and Stensaas v. Norway, App no 21980/93 (ECtHR, 20 May 1999), para.62.

[9] Hertel v. Switzerland App. no 59/1997/843/1049 (ECtHR, 25 August 1998), para.50.

[10] Terminiello v Chicago 337 US 1,4 (1949).

[11] Lingens v. Austria App no 9815/82 (ECtHR, 8 July 1986), par.42; Spanish Constitutional Court, Voz de España case, STC of June 81, Boletín de Jurisprudencia Constitucional 2, 128, para. 3. As cited in IPI & Media Legal Defence Inciative, FoE, Media Law and Defamation, (February 2015) <http://www.mediadefence.org/sites/default/files/resources/files/MLDI.IPI%20defamation%20manual.English.pdf> Accessed 28 January 2021.

[12] Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism, para. 69.; Binod Rao v M R Masani, 78 Bom.LR 125, Bombay High Court (1976).

[13] Feldek vs Slovakia App no 29032/95 (ECtHR, 12 July 2001), para.83.

[14] Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a Guide to human rights for Internet users (Adopted by the Committee of Ministers on 16 April 2014 at the 1197th meeting of the Ministers' Deputies), para 5.

CEU eTD Collection

However, despite the fact that some companies do apply those principles, the UN guiding principles on business and human rights did set a "global standard of expected conduct" that should be applicable as to all operations ran by any company.[16]

2) Companies' responsibilities for human rights

The minimum standards set for all companies by the guiding principles are many, they should in general "Avoid causing or contributing to adverse human rights impacts" and "seek to prevent or mitigate such impacts directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts";[17] companies should also make "A policy commitment to meet their responsibility to respect human rights",[18] this also includes FoE.

Due diligence should be carried out by businesses, this due diligence should include assessing "potential human rights impacts", while also responding properly to the findings concerning the impacts on human rights. [19] After the completion of this process, the criteria for verification of properly addressing human rights impacts is to "Draw on feedback from both internal and external sources, including affected stakeholders."[20]

<https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804d5b31> Accessed 28 January 2021

[15] Danish Institute for Human Rights submission. Cf. Yahoo/Oath submission, 2016 <https://www.ohchr.org/Documents/Issues/Expression/Telecommunications/Yahoo.pdf> Accessed 28 January 2021

[16] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 11.

[17] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 13.

[18] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 15.

[19] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 17.

[20] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 20.

CEU eTD Collection

The measure that includes efficiency for those human rights' impacts is to include "complaint mechanisms" that would be able to achieve "suitable remediation".[21]

3) Intermediary liability

Aside from the general company standards, specifically when it comes to Internet service providers or Social media platforms, several states had already regulated the rules in order to protect those intermediaries from liability of third party content.[22]

For example, the European Union followed the safe harbor approach through the E-commerce directive, such directive exempt the intermediaries from liability when it comes to third party content. However, it is provided that intermediaries should only play their roles as "cache", "mere conduit", or "host". If they exceeded those roles, or did not fulfill them properly as stated in the directive, then they could be held liable for third party content. [23]

Another example is Section 230 of the Communications Decency Act in the United States; it followed the broad immunity approach, as it exempted intermediaries of "interactive computer services" that publish or host data about others, as it specifically stated that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[24]

CEU eTD Collection

[21] UN' Guiding Principles on Business and Human Rights meeting (16 June 2011), principle 22, 29, & 31.

[22] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 14.

[23] Council Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1 of 8 June 2000, Articles 12-14.

[24] Communication Decency Act of 2014 (US) section 230 (C) (1).

In Brazil, a court order is deemed necessary for the removal of third party content,[25] as the Brazilian law stated "the provider of internet applications can only be subject to civil liability for damages resulting from content generated by third parties if, after a specific court order, it does not take any steps to, within the framework of their service and within the time stated in the order, make unavailable the content that was identified as being unlawful, unless otherwise provided by law." [26]

While other countries such as India establishes a "notice and takedown" course of action which includes having a court order to remove the illegal content.[27] Further, the notice and takedown system is widely adopted in relation to online illegal content.[28] As it could be imposed by a court order[29] and intermediaries shall obey such orders[30] so as not to be held liable.

Thus, there are different systems when it comes to the regulation of FoE especially when such expression occurs online.

4) Artificial intelligence significance to FoE

[25] Marco Civil da Internet, federal law 12.965, arts. 18−19.

[26] Marco Civil da Internet, federal law 12.965, Art. 19.

[27] Supreme Court of India, Shreya Singhal v. Union of India, decision of 24 March 2015.

[28] Council Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1 of 8 June 2000; Digital Millennium Copyright Act 1998 (US) section 512 ; Electronic Communications and Transactions Act 25 of 2002 (South Africa); Communication Decency Act of 2014 (US) section 230.

[29] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (16 May 2011) A/HRC/17/27, p.48; 'Manila Principles on Intermediary Liability' (2015) page 2.

[30] UN Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on FoE and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on FoE and Access to Information, Joint Declaration on FoE and the Internet (2011), Para 2(a)

CEU eTD Collection

When using those AI technologies with filters over social media platforms, they can do –among other actions- keyword filtering, or hash matching. Keyword filtering is the basic filter of blocking certain words from being using, in other words, "blacklisted" words cannot be used. "Hash matching" on the other hand is a bit more complex, where this form of technology operates by granting a certain speech a digital fingerprint which is used in blocking future publications that have a similar type of speech by comparing this digital fingerprint to this speech.[31]

Hash matching algorithms is used by YouTube for example in cases of potential copyright infringement, or in the context of Microsoft, it is used for sexual abuse of children content. This may result in either the automated block of the content, or sending it for human review. There is another –more advanced- kind of filters, which is based on machine learning algorithms uses natural language processing to identify different types of content that is essentially prohibited by the social media platform's policy. Those kinds of filters do circumvent the online content in a sense of prior restraints.[32]

Given that the legal rules regulating speech vary from a jurisdiction to another, it was recommended by the European Commission that States would obligate service providers to actively monitor and remove illegal content.[33] This was not an isolated incident in the European Union alone, as in 2017 Kenya adopted Guidelines for Prevention of Dissemination of

[31] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020) P. 2

[32] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020) P. 2

[33] COMMISSION RECOMMENDATION (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, para. (6), (39), (41) & 42.

CEU eTD Collection

Undesirable Bulk Political SMS and Social Media Content via Electronic Communications Networks, which required platforms to ""pull down accounts used in disseminating undesirable political contents on their platforms" within 24 hours." [34]

The Special Rapporteur on the Promotion on FoE recognizes that "In recent years, States have pushed companies towards a nearly immediate takedown of content, demanding that they develop filters that would disable the upload of content deemed harmful."[35]

Such rules of monitoring and filtering speech are understandable, given that States may have legitimate grounds such as national security measures. However, this risks the individual's enjoyment of their right to free speech, as intermediaries in this case would be heavily burdened with removing any content that may seem to violating the law, thus avoiding any risk as to being held liable.[36]

There has been a push towards the automation of filtering content, as it was stated that "Fully automated deletion or suspension of content can be particularly effective and should be applied where the circumstances leave little doubt about the illegality of the material".[37]

Such automation could only be achieved using Artificial Intelligence ("AI") filters, which the Special Rapporteur on FoE defined as "AI is often used as shorthand for the increasing

[34] Communication No. OL KEN 10/2017 <https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL_KEN_10_2017.pdf> Accessed 28 January 2021

[35] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (9 October 2019) A/74/486, para. 34

[36] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 28.

[37] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling Illegal Content Online Towards an enhanced responsibility of online platforms, Brussels, 28.9.2017, COM(2017) 555 final, p. 14.

CEU eTD Collection

independence, speed and scale connected to automated, computational decision-making. It is not one thing only, but rather refers to a "constellation" of processes and technologies enabling computers to complement or replace specific tasks otherwise performed by humans, such as making decisions and solving problems."[38] Still, AI can be potentially problematic as machines will not operate based on the same concepts of human intelligence.[39]

Taking into consideration that such technologies is used by online platforms "to help moderate content on their platforms, often acting as the first line of defense against content that may violate their rules".[40] The online platforms themselves put very broad terms when it comes to defining what goes against their rules[41] as they were criticized by the Special Rapporteur on the promotion of FoE in which the report stated "Company policies on hate, harassment and abuse also do not clearly indicate what constitutes an offence. Twitter's prohibition of "behavior that incites fear about a protected group" and Facebook's distinction between "direct attacks" on protected characteristics and merely "distasteful or offensive content" are subjective and unstable bases for content moderation."[42]

[38] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 3 & 13; AI Now 'The AI now report: the social and economic implications of AI technologies in the near term' 2016. < https://ainowinstitute.org/AI_Now_2016_ Report.pdf> Accessed 28 January 2021

[39] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 3

[40] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para. 1.

[41] Twitter Violent organizations policy (October 2020) <https://help.twitter.com/en/rules-and-policies/violent-groups> Accessed 28 January 2021; Facebook community standards (dangerous organizations) <https://www.facebook.com/communitystandards/#dangerous-organizations> Accessed 28 January 2021; YouTube policies 'violent or graphic content policies' <https://support.google.com/youtube/answer/2802008?hl=en> Accessed 28 January 2021

[42] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 26.

CEU eTD Collection

When using those AI technologies with filters over social media platforms, they can do –among other actions- keyword filtering, or hash matching. Keyword filtering is the basic filter of blocking certain words from being using, in other words, "blacklisted" words cannot be used. "Hash matching" on the other hand is a bit more complex, where this form of technology operates by granting a certain speech a digital fingerprint which is used in blocking future publications that have a similar type of speech by comparing this digital fingerprint to this speech.[43]

Hash matching algorithms is used by YouTube for example in cases of potential copyright infringement, or in the context of Microsoft, it is used for sexual abuse of children content. This may result in either the automated block of the content, or sending it for human review. There is another –more advanced- kind of filters, which is based on machine learning algorithms uses natural language processing to identify different types of content that is essentially prohibited by the social media platform's policy. Those kinds of filters do circumvent the online content in a sense of prior restraints.[44]

This automation of removal "removes human intervention from parts of a decision - making process, completing specific tasks with computational tools. This can have positive implications from a human rights perspective if a design limits human bias… Automation also enables the processing of vast amounts of data at a speed and scale not achievable by humans, potentially

[43] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020) P. 2

[44] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020)

CEU eTD Collection

serving public safety, health and national security."[45] However, there are some drawbacks to this automation as "automated systems rely on datasets that, in their design or implementation, may allow for bias and thus produce discriminatory effects"[46]

As it happens, an AI facial recognition application that was used in the United States of America in order to track down perpetrators, the application mistakenly classified 28 members of the Congress as people that should be arrested, and that was based on the 25,000 photos database fed to the system by the police. The application showed a higher error rate when it dealt with congressmen with darker skin, thus it showed automatic bias.[47]

Moreover, in automated decision making when an automated decision making system was installed to determine the care time needed for disabled people based on their situation, a patient was originally assigned 56 hours per week by a highly qualified nurse, however the automated system reduced this amount of time to 32 hours per week which resulted in reducing the person's quality of life, and that was done without any reasoning or a chance to comment or intervene. Thus, it left the patient without any reasoning and without any right to appeal. [48]

This was a raised concern by the Special Rapporteur as it was stated "Users and civil society experts commonly express concern about the limited information available to those subject to content removal or account suspension or deactivation, or those reporting abuse such as

[45] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 6

[46] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 6

[47] AI Now Institute, AI Now Report 2018, New York University, December 2018, p.15 – 17: <https://ainowinstitute.org/AI_Now_2018_Report.pdf> Accessed 28 January 2021

[48] AI Now Institute, AI Now Report 2018, New York University, December 2018, p.18 – 22: <https://ainowinstitute.org/AI_Now_2018_Report.pdf> Accessed 28 January 2021

CEU eTD Collection

misogynistic harassment and doxing. The lack of information creates an environment of secretive norms, inconsistent with the standards of clarity, specificity and predictability. This interferes with the individual's ability to challenge content actions or follow up on content-related complaints; in practice, however, the lack of robust appeal mechanisms for content removals favors users who flag over those who post."[49]

Further, this automation may in fact be valuable for the companies in terms of assessing the enormous amount of user generated content. However, when used to remove content, it raises concerns of over-blocking, pre-publication censorship.[50] An example could be provided when Mr. Frédéric Durand-Baïssas had uploaded a nude painting drawn by the French painter Gustave Courbet, and despite the painting not going against the community standards of Facebook, the painting was censored and the account that was used to post it was suspended.[51] This autonomous process is what unjustifiably suppresses FoE.

Moreover, more censorship and more unpredictability could be the result of machine-learning AI systems, as they are adaptable; they are capable of progressively identifying new problems and developing new answers. Those new answers may not be foreseen by humans.[52]

[49] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 58.

[50] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 32-33.

[51] Hakim Bishara, 'Facebook Settles 8-Year Case With Teacher Who Posted Courbet'S "Origin Of The World"' (Hyperallergic, 2019) <https://hyperallergic.com/512428/facebook-settles-8-year-case-with-teacher-who-posted-courbets-origin-of-the-world/> accessed 22 January 2021.

[52] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 8.

CEU eTD Collection

When we tackle this issue from the user' perspective, when it comes to the algorithms, despite the fact the users may be informed about the existence of such algorithms, the users do not have a manual as to what may be deemed offensive or not, thus they lack clarity as to what may be restricted and what may be deemed appropriate, "this means that individuals will often have their expression rights adversely affected without being able to investigate or understand why, how or on what basis."[53] While also taking into consideration that machine learning algorithms "may change their own rules and algorithms over time."[54]

CEU eTD Collection

[53] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 31-32.

[54] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 40

Chapter 2: The Inter American System on Human Rights framework on AI filtering speech

1) Existing framework for FoE in the context of AI

The right to FoE is protected under Article 13 of the American Convention on Human Rights ("ACHR"), the Article guarantees the exercise of FoE through any medium, and expressly – unlike other Human Rights instruments- prohibits "prior censorship" in paragraph 2, yet grants an exception in paragraph 4 by subjecting "public entertainment" to prior censorship when it is provided by law, for the sake of "moral protection of childhood and adolescence".

This leads us to the understanding that the "freedom to seek, receive, and impart information" as guaranteed in Article 13 can only be subject to subsequent sanctions if the right is abused, but it should never be subject to measures that would be prior to expressions.[55]

In the light of the aforementioned technologies, in 2010 the special rapporteurs identified ten key challenges to FoE in the current decade, one of those 10 points was relating to FoE over the internet, the specific challenge was phrased in a sense that shows that the fear of the governments control over the free flow of information. The more specific point that is extremely interlinked to this paper's topic is the fear of imposing filters.[56] Thus, it is obvious that the issue of AI filters in relation with FoE was foreseeable back in 2010; however it was not foreseeable how this filtering was going to happen.

[55] 'OAS :: Chapter II - Freedom Of Expression In The Inter-American System' (Oas.org) <https://www.oas.org/en/iachr/expression/showarticle.asp?artID=630&lID=1> accessed 5 March 2021.

[56] 'OAS :: TENTH ANNIVERSARY JOINT DECLARATION: TEN KEY CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE' (Oas.org, 2010) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=784&lID=1> accessed 5 March 2021.

CEU eTD Collection

Building upon that, in 2019, the Special Rapporteurs went into more details as they were able to identify that private control would be an invasive to FoE. Given the nature that we currently live by, where corporations have control over the social media, and they have the "enormous power" to regulate the flow of information.[57]

For this reason, States were urged to develop -among other actions- policies for private content regulation, thus making it consistent with International Human Rights Law, to hold companies responsible for human rights violations in compliance with the Guiding Principles on Business and human rights, and finding legal and technological solutions to allow the algorithmic feeding of data to the AI systems.[58] Thus, in a sense, it can be demonstrated that the IAS did not expressly ban using AI filters. Yet, in another sense, it banned prior censorship, which AI filters result in.

Given the role that the internet plays a "unique transformational tool"[59] in ensuring the right to FoE, it is of importance to determine the legitimacy of every restriction that is imposed over the internet,[60] regardless of who imposed it.

While assessing those restrictions, whether they are related to AI or in general, one has to take into account the interests of the individuals involved, and the consequences of the impact of the

[57] 'OAS :: TWENTIETH ANNIVERSARY OF THE JOINT DECLARATION: CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE' (Oas.org, 2019) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1146&lID=1> accessed 5 March 2021.

[58] 'OAS :: TWENTIETH ANNIVERSARY OF THE JOINT DECLARATION: CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE' (Oas.org, 2019) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1146&lID=1> accessed 5 March 2021.

[59] 'OAS :: JOINT DECLARATION by the UN Special Rapporteur for Freedom of Opinion and Expression and the IACHR-OAS Special Rapporteur on Freedom of Expression' (Oas.org, 2012) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=888&lID=1> Accessed 5 March 2021

[60] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 25.

CEU eTD Collection

restriction on both the "functioning of the internet" and the "Internet's capacity to guarantee and promote freedom of expression against the benefits that the restriction would have in protecting other interests".[61]

1.1) Content Regulation

The idea of content regulation is of vital necessity, as filtering and blocking content is used in different countries in order to disallow their populations from accessing certain types of information that would be in the public's interest to know, but would contradict the government's interest if the public were to know.[62]

On one hand, It can be argued that correcting the flawed information that has already been disseminated would be the least costly measure[63] as "Only when this is insufficient to repair the harm that has been inflicted may recourse be made to the imposition of legal liabilities more costly for those who have abused their right to freedom of expression, and – while doing so- have produced an actual and serious damage to the rights of others"[64]. On the other hand in the context of online FoE "Self-regulation can be an effective tool in redressing harmful speech".[65]

[61] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 25-26

[62] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 39

[63] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 32

[64] Report of the Special Rapporteur for Freedom of Expression Annual report of the IACHR (30 December 2009) OEA/Ser.L/V/II Para 109

[65] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 1 (E) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021.

CEU eTD Collection

When applying the aforementioned filtering system of AI, on the one hand, to this kind of flawed information, the information may be blocked from being disseminated in the first place, thus it will not be susceptible for rectification, nor will the speaker will be susceptible for legal consequences if their speech abused the limits of FoE. On the other hand, the Joint Declaration promoted self-regulation without setting limits to this self-regulation, which could potentially open the door for intermediaries to use those kinds of AI filters.

However, one could make more sense of this framework, if self-regulation was to only be allowed in the context where it would not allow for prior restraints over speech. For those reasons, one is inclined to agree that rectification would be the more viable solution, rather than imposing prior restraints by AI filters, except in scenarios where enough human review is involved, rather than just hash matching and keyword blocking.

What further affirms the previous conclusion is "Content filtering systems which are imposed by a government or commercial service provider and which are not end-user controlled are a form of prior censorship and are not justifiable as a restriction".[66]

However, the contradiction in this argument occurs when inciting to commit illegal actions where prior censorship was allowed. Yet still "filtration or blocking should be designed and applied so as to exclusively impact the illegal content without affecting other content".[67] If this was to be applied then it would directly and bluntly go against the wording or Article 13. But this

[66] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 1 (B) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021.

[67] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 37

CEU eTD Collection

ought to be applied "after the illegal content to be blocked has been fully and clearly identified".[68] In this sense it would be justified under Article 13.

This model would be inapplicable under the AI filtering system, as it would be justifiable to filter illegal content, yet the content needs to be published beforehand on social media platforms and evaluated by courts in order to clearly identify the illegality of the content, but AI filters would not allow that, the solution in this sense is to suspend controversial content until it has been reviewed –in a reasonable time- by the competent court, and then resume its publication later on if the court lands on a positive decision.

Further, intermediaries should not be obligated to monitor the content that is generated from the users, yet they should abide by the regulations provided by the government in order not to be held accountable for such content.[69] Which goes even further in asking the question of why do intermediaries employ such AI filter is they are not obligated to monitor the content.

In one instance the Inter-American Commission on Human Rights ("IACHR") declared that deciding to ban or confiscate hard copied materials would be inconsistent with Article 13 of the American Convention on Human Rights ("ACHR"), for example, Chile's decision to ban the distribution of a book violated the author's right to impart information, as this was bluntly prior censorship.[70]

CEU eTD Collection

[68] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/1, P. 37

[69] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 2 (b) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021.

[70] Francisco Martorell v. Chile Case 11.230 Report No. 11/96 Inter-Am.C.H.R.,OEA/Ser.L/V/II.95 Doc. 7 rev. at 234 (IACHR 1997) para. 58&59.

In another instance, the IACHR decided that confiscation and the ban of published books constitutes prior censorship, and is not consistent with Article 13 and the IACHR stated that "It is equally true that the right to impart information and ideas cannot be invoked to justify the establishment of private or public monopolies of the communications media designed to mold public opinion by giving expression to only one point of view."[71]

Putting the cases in conjunction to the idea of suspending controversial content, until it has been reviewed, then allow for its publication. The facts of the Grenada Case do not necessarily contradict this solution, as online publications could take place, and if proven controversial, that is what temporary injunctions are for. Thus, by analogy, this solution would be consistent with Article 13.

1.2) Intermediary Liability

"in most cases, intermediaries do not have—and are not required to have—the operational/technical capacity to review content for which they are not responsible. Nor do they have—and nor are they required to have—the legal knowledge necessary to identify the cases in which specific content could effectively produce an unlawful harm that must be prevented. Even if they had the requisite number of operators and attorneys to perform such an undertaking, as private actors, intermediaries are not necessarily going to consider the value of freedom of expression when making decisions about third-party produced content for which they might be held liable."[72]

[71] Steve Clark v. Grenada Case 10.325 Report No. 2/96 Inter-Am.C.H.R., OEA/Ser.L/V/II.91 Doc. 7 at 113 (IACHR 1996) para 8

[72] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 44

CEU eTD Collection

However, it is easier for authorities to hold intermediaries accountable for illegal speech rather than holding people accountable for their own speech.[73] This could be –among other reasons- a reason where intermediaries use AI filters, just in order to avoid being held accountable for third party content.

There are two acceptable approaches under the IAS when it comes to holding intermediaries accountable for such content.[74]

The first approach is ideal for intermediaries (Broad immunity), as it completely negates holding intermediaries liable for any illegal content that is essentially disseminated by third party users. This is evident through the following text "[n]o one who simply provides technical Internet services such as providing access, or searching for, or transmission or caching of information, should be liable for content generated by others, which is disseminated using those services, as long as they do not specifically intervene in that content or refuse to obey a court order to remove that content, where they have the capacity to do so ('mere conduit principle')."[75]

Under this broad immunity model intermediaries will not have to fear being prosecuted for leaving third party illegal content online, unless they neglect court orders which would be

[73] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 40

[74] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 39-53

[75] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 2 (a) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021.

CEU eTD Collection

compelling them to remove this content. Thus it was actually recommended to just hold the authors of the speech accountable for their own speech.[76]

The second approach is based on efforts spent by the intermediary to avoid liability; this model's name is named the safe harbor model. An example for this mechanism is the "notice and takedown" whereas the intermediary a notice to remove the illegal content, and then the intermediary takes down the content; otherwise they would be held liable.[77] Also, the "notice and notice" is another mechanism where the intermediary has to notify the user of the alleged illegality of their speech.[78]

In some cases, this fault based liability model puts private intermediaries in a position where they would have to decide on the legality or rather illegality of the content; subsequently this would lead censorship, as this kind of notice –depending on the jurisdiction- could be judicial or extrajudicial.[79] For this reason, this model needs corrections, as it is not efficient to leave private actors to determine the illegality of content. The kind of correction that is needed for this model is to have the notice as a judicial court order, then taking down the content, and if the intermediary disobeys, that is when the intermediary would be held liable.

CEU eTD Collection

[76] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 2 (a) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021.

[77] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 46

[78] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 49

[79] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 47-48

The safe harbor model, despite granting immunity in a conditional manner, its results can drive intermediaries to censor legal content, just for the sake of not being held liable.[80] In the sense of AI, this would be a sufficient reason for an intermediary that does not have the capacity to review all of its content, to employ AI filtering mechanism, where prior censorship would prevail.

2) Conclusion and analysis

It is obvious that AI filters play a role in censoring FoE in the current context of online speech, whether it is hash matching, keyword blocking, or even machine learning, they do constitute different forms of prior censorship. They could be justified if there is a sufficient human reviewer that would provide legal aid in a timely manner. Yet legal aid is not always sufficient to determine the legality or illegality of every action.

The more proper mean would be to have a collaboration between the judicial system, as well as the social media platforms, where social media platforms could refer controversial content to the judiciary, just to determine their legality.

In terms of suspending content until it has been reviewed, a time limit should be set, so that the speech should not lose its value, as it may be contributing to a public debate that is relevant at a time, but not relevant in another.

With regards to intermediary liability, in case the Safe harbor model is to be amended for intermediaries to only comply with court orders, except in cases where keyword filtering would take place, not sophisticated filtering, then it would be the more favorable model in terms of

[80] Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 43.

CEU eTD Collection

limiting the usage of AI filters. However, in case the two models are put into perspective the way they are, then the broad immunity model would be the better outcome, due to the fact that intermediaries will not need to use AI filters.

Thus, governments as well as private actors need to pay more attention towards finding solutions for inherent risks of deploying AI filters and subjecting legitimate speech to prior censorship.[81] On the one hand intermediaries need to adopt the UN guiding principles on Human Rights,[82] and on the other hand, States need to implement in their laws, that private actors should adopt the aforementioned principles.

In this sense, the IAS one is inclined to conclude that the IAS needs to further develop its legal framework, as well as its mechanisms in order to properly deal with the issue of AI filtering systems.

CEU eTD Collection

[81] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020), P. 5

[82] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND ELECTIONS IN THE DIGITAL AGE' (Oas.org, 2020) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1174&lID=1> accessed 7 March 2021.

Chapter 3: The European System on Human Rights framework on AI filtering speech

1) Existing legal framework

FoE is regulated under Article 10 of the European Convention on Human Rights ("ECHR"), yet the different aspects of technology when it comes to filtering systems, and online FoE has more detailed documents aside from the big umbrella of Article 10, and throughout this chapter those documents will be tackled, yet it is noteworthy to mention that the ECHR does not directly prohibit prior censorship much like Article 13 of the ACHR.

Unlike any other regional system, a general definition of AI was provided by the European Commission whereas AI" refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."[83] Further, "AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."[84]

The methodology of how AI works is consistent with this definition, as AI operates through the collection of and interpretation of data, while also reasoning and deciding the best action to do

[83] European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018) P. 1

[84] European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018) P. 1

CEU eTD Collection

with this data, then acting according to the best possible outcome, which could even be modifying the environment around this data.[85]

The idea of reasoning or in other terms processing the information and deciding how to act accordingly is a common denominator between AI systems.[86] While applying this to AI filters, the filter processes the information, reasons the best decision as to what to make with this information, which can include the removal of certain content. This process is complex enough on human, to determine the legality or illegality of the speech, yet when it comes to a machine making those kinds of decisions; it is even more complicated as the machine deals with it as a sequence of 1s and zeros.[87]

1.1) Content regulation

In even more compliance with the definition provided above, social media platforms use AI systems to prioritize content, this kind of data processing moderates content, and put people in filter bubbles without allowing for human intervention, which negatively impacts the right to FoE.[88]

[85] A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019) P. 1

[86] A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019) P. 2

[87] A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019) P. 2; 1s and Zeros is a reference to the binary language of computers

[88] European Commission WHITE PAPER on Artificial Intelligence – A European approach to excellence and trust COM(2020) 65 final (19 Feb 2020) P. 11

CEU eTD Collection

To reiterate what was mentioned in the previous chapter from the Joint declaration on FoE "Self-regulation can be an effective tool in redressing harmful speech".[89] However, the Court of Justice of European Union ("CJEU") took the position that intermediaries are not obligated in any manner to install filtering systems.[90] This approach is stemmed from Article 15 of Electronic Commerce Directive ("ECD"), as the Article establishes that States cannot obligate intermediaries to monitor their content.[91] Thus in this sense, the approach of self-regulation is not strictly followed.

However, the European Court on Human Rights ("ECtHR") on the other hand took this to idea of self-regulation to a different standard, as in the case of Delfi the court found that the applicant (a news portal) did not act in a negligent manner due to the fact that they had installed a filtering system thus it undertook its "duty to avoid causing harm to third parties", yet the filters were not advanced enough to filter the illegal speech which allowed the illegal comments to stay online for 6 weeks, which resulted in holding the company liable for those comments.[92] This approach is contradictory to the approach followed by the CJEU as well as the ECD, and it works as an

[89] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 1 (E) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021; see also COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 2

[90] Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10 (CJEU 24 November 2011) Para 53; Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV Case C‑360/10 (CJEU 16 February 2012) Para 52

[91] Council Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1 of 8 June 2000, Article 15; COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 6

[92] Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015) para 156

CEU eTD Collection

incentive for intermediaries to install more sophisticated filtering systems, which would more likely use AI, so as to avoid liability.

It is to go without saying that the removal of AI systems conducted by private actors is supposed to conform to the "legality, legitimacy and proportionality" criteria, as laid down by Article 10 of the ECHR.[93]

The Committee of Ministers ("CoM") saw the threat of those AI systems, and it pointed out that the adoption of "appropriate legislative, regulatory and supervisory frameworks related to algorithmic systems" is of utmost necessity, and when such algorithmic systems are deployed by private actors, they should "comply with the applicable laws and fulfil their responsibilities to respect human rights in line with the UN Guiding Principles on Business and Human Rights and relevant regional and international standards".[94] Thus, the system in Europe makes a reference to international law, much like in the Inter-American system when it comes to the joint declarations, which factors into shaping the system. However, the CoM paid specific regards to the UN guiding principles on Business and Human Rights leads for better uniformity when it comes to regulating AI.

The CoM also paid specific regard to vulnerable groups as well as other categories of stakeholders when it comes to algorithmic systems "with a view to ensuring that human rights impacts stemming from the design, development and ongoing deployment of algorithmic

[93] Recommendation CM/Rec(2016)5[1] of the Committee of Ministers to member States on Internet freedom Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies, para 2.2.2

[94] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, para 3

CEU eTD Collection

systems are comprehensively monitored, debated and addressed",[95] this stems from the issues of algorithmic bias,[96] thus the algorithms would not favor one kind of content over another, for this one can argue that the legal framework being developed by the European system is improving in the area of AI.

Yet it was recommended by the CoM to integrate algorithmic systems in different societal aspects "with a view to effectively protecting human rights",[97] one can argue that this is to be criticized as the decisions made by those AI filters is not predictable, the legal framework may be advanced, yet from a pragmatic approach, the algorithms may not be in line with the legal framework set. Thus, this could be achieved later in the future when technology is further developed.

When it comes to having filter bubbles, the AI algorithms decides on the kind of content the user would be receiving, this can negatively impact the free flow of information,[98] which stems from the right to access information, or in other words the right to know, which is part of Article 10 of the ECHR as it entailed that within the ambit of the right to FoE is "to receive and impart

[95] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, para 5

[96] Entering The New Paradigm Of Artificial Intelligence And Series (Council of Europe and Eurimages 2019) P. 17 <https://rm.coe.int/eurimages-entering-the-new-paradigm-051219/1680995331> accessed 8 June 2021.

[97] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, para 6

[98] Unboxing artificial intelligence: 10 steps to protect human rights (Council of Europe Commissioner for Human Rights 2019) P. 12-13; Entering The New Paradigm Of Artificial Intelligence And Series (Council of Europe and Eurimages 2019) P. 22 <https://rm.coe.int/eurimages-entering-the-new-paradigm-051219/1680995331> accessed 8 June 2021.

CEU eTD Collection

information and ideas", thus in consequence having those filter bubbles would be problematic when it comes to receiving information.

In a general sense, due to the lack of predictability of AI, there should be meaningful human intervention, or in other words "human-in-command" where the human would be able to retain control over the system at any given point,[99] which in other words would allow humans to override the decisions made by the AI systems.

Any AI system should respect the basic standards set by international law, whether those standards are set under International Humanitarian Law or International Human Rights law, this is specifically applicable for FoE,[100] as prior censorship should not be allowed in those cases.

1.2) Intermediary Liability

Affirming the aforementioned Delfi judgment,[101] the Declaration on freedom of communication on the Internet affirmed that authorities should not install filters except when 1) "for the protection of minors, in particular in places accessible to them"; and/or 2) the filter should be in compliance with Article 10 of the ECHR and "measures may be taken to enforce the removal of

[99] European Economic and Social Committee ("EESC") 526th EESC plenary session of 31 May and 1 June 2017 'Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society'' 2017/C 288/01 31 August 2017 para 1.6

[100] Council of Europe Parliamentary Assembly 'Technological convergence, artificial intelligence and human rights' Recommendation 2102 (2017) para 10

[101] Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015) para 156

CEU eTD Collection

clearly identifiable Internet content or, alternatively, the blockage of access to it, if the competent national authorities have taken a provisional or final decision on its illegality."[102]

However, the Council of Europe ("CoE") Declaration on freedom of communication on the Internet specifically stated that "Member states should ensure that service providers are not held liable for content on the Internet when their function is limited, as defined by national law, to transmitting information or providing access to the Internet."[103]

Yet if their functions goes beyond "transmitting information or providing access" as in storing content from third parties, they can be held "co-responsible" for the illegal content if they fail to act expeditiously to remove the content "as soon as they become aware, as defined by national law, of their illegal nature"[104]

Further, when it comes to the notice and take-down system, intermediaries should not design the aforementioned systems to pull down legal content "Notices should contain sufficient information for intermediaries to take appropriate measures" that applies when the notice is from the State, as the State should assess the illegality of the content in compliance with international

[102] COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 3

[103] COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 6

[104] COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 6; see also Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies) para 1.3.7

CEU eTD Collection

law, and the intermediaries should provide the same notice to the owner/publisher of this content, and users trying to access this content.[105]

Despite all of those specific details, the general system in holding intermediaries liable is codified within the ECD Articles 12-14, whereas the followed system is the Safe Harbor system, thus it is a fault based approach, if social media platforms manage to act expeditiously to remove the illegal content, and they will not be held liable.

The system comprises of the aforementioned notice and take-down, whereas the intermediary receives a notice, which makes the social media platform aware of the illegal content, thus not necessarily, but could be a court order, and if the social media platform fails to take-down the content expeditiously then it would be held liable for the illegal content.[106]

2) Conclusion and analysis

It is to be acknowledged that the European system has more documents when it comes to regulating AI when it comes to filtering; it introduces the concept of filter bubbles, which is not mentioned in different systems for example.

Yet one can argue that, filter bubbles can be allowed with an exception to general news or specific news chosen by the user, as the algorithms helps users reach more content that relates to their interests.

[105] COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 6; see also Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies) para 1.3.7

[106] Council Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1 of 8 June 2000, Artilces 12-14

CEU eTD Collection

It has been demonstrated that AI learns from the surrounding environment or from previous encounters, which is why a different definition to AI was proposed which provides that they are *"systems are software (and possibly also hardware) systems designed by humans*[107] *that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."*[108]

While the definition is much more accurate with the technology, and fits well into the protection of FoE, yet the idea that it is limited to only "Humans" can pose some problems in the future, despite the fact that humans may "use AI techniques to optimize their design". The unpredictability of the system could result in AI systems creating other AI systems, and that would not be governed by any law since they would fall out of this definition. Thus, one would agree to the adoption of this proposed definition if the phrase "*designed by humans*[109]*"* was to be

[107] "Humans design AI systems directly, but they may also use AI techniques to optimise their design."

[108] A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019) P. 6

[109] "Humans design AI systems directly, but they may also use AI techniques to optimise their design."

CEU eTD Collection

removed from the text. Even more preferably, the aforementioned definition provided by the European Commission[110] is better to some extent as it is broader and does not put the same limitations as the one proposed by the experts meeting.

Moreover, when it comes to content moderation one can witness the contradictory opinions, as there is no obligation to monitor based on Article 15 of ECD, and the CJEU affirmed that there is no obligation for social media platforms to install filtering systems,[111] yet the Delfi judgment held Delfi responsible for the mere fact that the filter installed was not sophisticated enough to act as an upload filter for specific keywords, as the phrasing of the sentences was not complex,[112] this judgment would put an incentive for different platforms to install more sophisticated AI driven filters, which does not contradict the idea of self-regulation,[113] but it puts more pressure towards self-regulation.

It is apparent that the European approach is leaning more towards self-regulation since the speculations from the Delfi judgment can be affirmed through the European Commission

[110] European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018) P. 1

[111] Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10 (CJEU 24 November 2011) Para 53; Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV Case C‑360/10 (CJEU 16 February 2012) Para 52

[112] Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015) para 156

[113] 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) Para 1 (E) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1> accessed 6 March 2021; see also COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 2

CEU eTD Collection

recommendation that States would obligate service providers to actively monitor and remove illegal content.[114] This bluntly goes against Article 15 of the ECD.

The issue with self-regulation is that the removal of AI systems conducted by private actors is supposed to conform to the "legality, legitimacy and proportionality" criteria, as laid down by Article 10 of the ECHR.[115] Yet this Article is addressed to Member States of the ECHR, not private actors, thus it is a State obligation to conduct the aforementioned criteria, and it is not something that should be left to be conducted by private actors.

More problems arise from self-regulation when it comes to filter bubbles, as it would impact the users' right to be informed, which relates to their FoE under Article 10 of the ECHR, the algorithmic process with regards to the creation of those filter bubbles should allow for the users to opt out, and whether the decide to opt out or to stay within the filter bubble, limitations should be drawn where necessary general news would reach everyone, regardless of how vulgar it is, or how contradictory to the stances of the social media platform.

Finally when it comes to intermediary liability as mentioned in the previous chapter, intermediary liability can be problematic while applying the Safe Harbor approach as it incentivizes using sophisticated AI filters, the European system only applies the Safe Harbor approach while dealing with social media platforms.

This lead can lead one to a similar conclusion to the last Chapter when it comes to intermediary liability, thus amending the system to only comply with court orders instead of general notices,

[114] COMMISSION RECOMMENDATION (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, para. (6), (39), (41) & 42.

[115] Recommendation CM/Rec(2016)5[1] of the Committee of Ministers to member States on Internet freedom Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies, para 2.2.2

CEU eTD Collection

except in cases where keyword filtering would take place, yet not sophisticated filtering, the other option would be to just follow the broad immunity approach.

Yet one should point out and emphasize the positive side of how the European approach adopts the Safe Harbor liability system, whereas they hold the intermediary "Co-responsible", thus the intermediary should not be liable for the illegal content all by itself. [116]

CEU eTD Collection

[116] COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies) Principle 6; see also Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies) para 1.3.7

Conclusion and recommendations

1) Conclusion

For the beginning of the conclusion, one can observe that there have been several definitions of AI that have been provided, one was provided by the UN special rapporteur on FoE,[117] the second was provided by the European Commission,[118] and the third was proposed during an experts meeting conducted by the European Commission.[119] This goes without saying but neither the IAS nor the African System on Human Rights provided definitions for AI.

Putting the three definitions into perspective, the Special rapporteur's definition is more generic than which leaves the door open for any possible outcomes those technologies can pose, yet it can be criticized for saying "AI is often used as shorthand for the increasing independence"[120] because despite the fact that the statement is true using the word "often" leaves a wide door open for interpretations, and the usage of AI filters can occur without necessarily being shorthanded.

The definition provided by the experts meeting has been criticized in the concluding part of Chapter 3, and that leaves us with the definition provided by the European Commission, which is

[117] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 3 & 13; AI Now 'The AI now report: the social and economic implications of AI technologies in the near term' 2016. < https://ainowinstitute.org/AI_Now_2016_ Report.pdf> Accessed 28 January 2021

[118] European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018) P. 1

[119] A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019) P. 6

[120] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 3 & 13; AI Now 'The AI now report: the social and economic implications of AI technologies in the near term' 2016. < https://ainowinstitute.org/AI_Now_2016_ Report.pdf> Accessed 28 January 2021

CEU eTD Collection

generic much like the UN Special rapporteur's definition without detailing a reason for the usage, such as being shorthanded.

For that reason, the researcher is inclined to concede to the definition provided by the European Commission. Yet if there is an opportunity to redefine AI, the researcher would propose something that would mix between the UN Special Rapporteur's definition, as well as the European Commission's definition which would be: AI refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. It is not one thing only, but rather refers to a constellation of processes and technologies enabling computers to complement or replace specific tasks otherwise performed by humans, such as making decisions and solving problems. Further, AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).[121]

When it comes to content moderation, one can observe that the European System made a specific reference to the International standards set by the UN guiding principles of Business and Human Rights.[122] In the area of AI this idea of regional systems integrating international documents into their framework is not yet common due to the lack of materials that meets consensus from all

[121] The definition is merely merging and cutting without rephrasing the words of the following citations: Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 3 & 13; AI Now 'The AI now report: the social and economic implications of AI technologies in the near term' 2016. < https://ainowinstitute.org/AI_Now_2016_ Report.pdf> Accessed 28 January 2021; European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018) P. 1

[122] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, para 3

CEU eTD Collection

States, yet the IAS took part in several joint declarations joint Special Rapporteur reports as provided in Chapter two, and those joint documents led to the shaping of their system.

Further, Article 13 of the ACHR clearly prohibits censorship, which is not the case with either Article 19 of the ICCPR nor is it the case with Article 10 of the ECHR, yet the joint declarations promoted self-regulation, the same approach was adopted by the European system, that is despite the 'no obligation to monitor' of Article 15 of the ECD. The issue of self-regulation as already mentioned in the conclusion of Chapter 3 is that any restriction on FoE needs to abide with the proportionality test as provided in Article 19 (3) of the ICCPR, yet this is a State obligation, thus States and only States are in a position to assess those kinds of restrictions.

Moreover, one has to point out that when it comes to AI and content moderation that the European system is rather advanced than both international instruments and other regional systems, as it has identified the issues of filter bubbles that are problematic to the right to receive information. While this issue has been addressed, yet it remains unresolved, as those bubbles are built based on personal data and personal interests.

While all systems have tackled filtering systems, yet there remains no solution aside from having a human element involved in the process. There has been very little case law concerning the operation of AI filters one of which is the case of Muthukumar v. Telecom Regulatory Authority of India & Ors. before the Madras High Court in India,[123] where the case concerned the ban of downloading the Tik Tok application due to hosting content that is described as disturbing and explicit, that degrades culture and encourage in engaging with pornography.

[123] WP(MD) No. 7855 of 2019 (24 April 2019)

CEU eTD Collection

In paragraph 4 and 10 of the case the parent company of Tik Tok, Byte Dance argued otherwise, as the content on Tik Tok is subject to AI moderation on a first level, then to human moderation on the next three levels, and still, complaints can be sent to human moderators over what the AI systems take action against, then the human reviewer takes action accordingly. In paragraph 12 the court concluded that this alongside other measures that include the number of human reviewers and the languages they conduct the reviews in was considered sufficient human involvement in the AI system by the Madras High Court of India.

While taking this into consideration, one can observe that there is more involvement of humans than there is when it comes to acceptable AI systems when it comes to the previous case, this can be a proper approach to follow when it comes using AI filters if it is pragmatically as described in the aforementioned case. That is while also taking into consideration that intermediaries have no obligation to monitor their content, this can incentivize governments to oblige intermediaries to monitor their content as they already have.

For this reason, one can argue that no amount of human intervention would alleviate the impact of AI,[124] yet one can confidently say that AI is here stay since the amount of content that should be reviewed cannot be reviewed by only humans, thus the approach followed by the Madras High Court can be the more proper approach to adopt, alongside with the recommendations that will be provided in the upcoming section.

When it comes to intermediary liability, while international law and the IAS has no preference between the Safe Harbor approach as well as the Broad Immunity approach, the European system clearly went for the Safe Harbor approach, this issue was tackled in the conclusion and

[124] Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020) P. 1-6

CEU eTD Collection

analysis of chapter two, and one may lean towards the idea that with the current models of liability, the more favorable one when it comes to AI systems would be the Broad Immunity approach unless the Safe Harbor model was amended as discussed in the conclusion and analysis of Chapter two.

2) Recommendations

In addition to the conclusion and analysis, several other points can be taken into consideration, those are:

1. Social media platforms should allow users to select a category of what they are posting, thus if it is general news, then it would not be subject to a filter bubble.

2. The usage of AI filters should not be allowed when they are deployed in a manner that would not allow for human reviewers,[125] or in other words having a deep level of human oversight.

3. Governments alongside other stakeholders should take initiatives to allow individuals to develop their understanding of how AI systems function, and how those automated decisions are made, thus they can make an informed decision as to whether or not they would like to engage with such systems.[126]

[125] Unboxing artificial intelligence: 10 steps to protect human rights (Council of Europe Commissioner for Human Rights 2019) P.19

[126] Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, Para 7(a)&(b)

CEU eTD Collection

4. Any AI system needs to be trustworthy, this would entail it fulfilling 3 requirements:[127]
    a. Complies with applicable regulations and laws;
    b. Ethical, as in it would abide by ethical values; and
    c. Robust with good intention, thus it would not constitute bias in any manner.
5. The human-in-command approach has to adopted for safe and responsible AI deployment.[128]
6. A monitoring system for AI systems can be initiated; it would assess their transparency with the users, safety, accountability, comprehensibility, and ethical values.[129]
7. States should create judicial bodies that would collaborate with corporations in order to assess the legality of speech, and not leave this job for private actors, yet for this to be feasible, at least an international declaration needs to be issued to encourage States to take such action, and as a last resort measure, an international quasi-judicial body should be established to work in swift cooperation with internet intermediaries in order to determine the legality of online speech.

[127] European Commission independent High-Level Expert Group on Artificial Intelligence 'Ethics Gudelines for Trustworthy AI' (2019) P. 5

[128] European Economic and Social Committee ("EESC") 526th EESC plenary session of 31 May and 1 June 2017 'Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society'' 2017/C 288/01 31 August 2017 para 1.6

[129] European Economic and Social Committee ("EESC") 526th EESC plenary session of 31 May and 1 June 2017 'Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society'' 2017/C 288/01 31 August 2017 para 1.8

CEU eTD Collection

Bibliography

1) Case law

1. Animal Defenders International v. the UK App no. 48876/08 (ECtHR 22 April 2013)
2. Binod Rao v M R Masani, 78 Bom.LR 125, Bombay High Court (1976)
3. Bladet Tromso and Stensaas v. Norway, App no 21980/93 (ECtHR 20 May 1999)
4. Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism, Serie A No. 05 (IACtHR 13 November 1985)
5. Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015)
6. Francisco Martorell v. Chile Case 11.230 Report No. 11/96 Inter-Am.C.H.R.,OEA/Ser.L/V/II.95 Doc. 7 rev. at 234 (IACHR 1997) para. 58&59.
7. Hertel v. Switzerland App. no 59/1997/843/1049 (ECtHR 25 August 1998)
8. Muthukumar v. Telecom Regulatory Authority of India & Ors. WP(MD) No. 7855 of 2019 (Madras High Court 24 Apr. 2019).
9. Rolf Anders Daniel PIHL v Sweden, Application no 74742/14 (ECtHR 9 March 2017)
10. Steve Clark v. Grenada Case 10.325 Report No. 2/96 Inter-Am.C.H.R., OEA/Ser.L/V/II.91 Doc. 7 at 113 (IACHR 1996)
11. Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV Case C‑360/10 (CJEU 16 February 2012)
12. Muthukumar v. Telecom Regulatory Authority of India & Ors WP(MD) No. 7855 of 2019 (Madras High Court 24 April 2019)
13. Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10 (CJEU 24 November 2011)

2) Soft law documents

CEU eTD Collection

14. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35
15. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348
16. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (9 October 2019) A/74/486
17. Report of the Special Rapporteur for Freedom of Expression Annual report of the IACHR (30 December 2009) OEA/Ser.L/V/II
18. Guiding Principles on Business and Human Rights, Implementing the United Nations 'Protect, Respect and Remedy' Framework, HR/PUB/11/04 (2011)
19. Manila Principles on Intermediary Liability 'Best Practices Guidelines for Limiting Intermediary Liability for Content to Promote Freedom of Expression and Innovation' (Version 1.0, 24 March 2015)

3) Articles

20. Article 19, 'Internet Intermediaries: Dilemma of Liability' (2013), available at: http://www.article19.org/data/files/Intermediaries_ENGLISH.pdf
21. Article 19, Privacy International, 'Privacy and Freedom of Expression In the Age of Artificial Intelligence' (2018), available at: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf
22. Emma J Llans 'No amount of "AI" in content moderation will solve filtering's prior restraint problem' Big Data & Society January–June (2020)

CEU eTD Collection

23. Heather Brown, Emily Guskin, and Amy Mitchell 'The Role of Social Media in the Arab Uprisings' (Pew Research Center Journalism & Media, 2012) available at: https://www.journalism.org/2012/11/28/role-social-media-arab-uprisings/
24. Toni M. Massaro & Helen Norton 'Siri-Ously? Free Speech Rights And Artificial Intelligence' Northwest University Law Review Vol. 110, No. 5 (2016), available at: https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1253&context=nulr
25. Yavar Bathaee 'Artificial Intelligence Opinion Liability' 35 BERKELEY TECH. L.J. 113 (2020).

4) Books

26. Catalina Botero Marino 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13
27. Wolfgang Benedek and Matthias C. Kettemann, Freedom Of Expression And The Internet (Council of Europe 2013)

5) National Law

28. Electronic Communications and Transactions Act 25 of 2002 (South Africa)

6) European Union

29. A Definition of AI: Main Capabilities and Disciplines (European Commission Independent High-Level Expert Group on Artificial Intelligence 2019)
30. COMMISSION RECOMMENDATION (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online
31. Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions

CEU eTD Collection

Tackling Illegal Content Online Towards an enhanced responsibility of online platforms, Brussels (28 September2017) COM(2017) 555 final

32. European Commission independent High-Level Expert Group on Artificial Intelligence 'Ethics Gudelines for Trustworthy AI' (2019)
33. European Commission WHITE PAPER on Artificial Intelligence – A European approach to excellence and trust COM(2020) 65 final (19 Feb 2020)
34. European Commission, Communication from the Commission 'Artificial Intelligence for Europe' {SWD(2018) 137 final} COM(2018) 237 final (2018)
35. European Economic and Social Committee ("EESC") 526th EESC plenary session of 31 May and 1 June 2017 'Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society'' 2017/C 288/01 31 August 2017

7) Council of Europe

36. Council Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1 of 8 June 2000
37. COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers' Deputies)
38. Council of Europe Parliamentary Assembly 'Technological convergence, artificial intelligence and human rights' Recommendation 2102 (2017)

CEU eTD Collection

39. Entering The New Paradigm Of Artificial Intelligence And Series (Council of Europe and Eurimages 2019) <https://rm.coe.int/eurimages-entering-the-new-paradigm-051219/1680995331> accessed 8 June 2021.
40. Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a Guide to human rights for Internet users (Adopted by the Committee of Ministers on 16 April 2014 at the 1197th meeting of the Ministers' Deputies), available at: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804d5b31
41. Recommendation CM/Rec(2016)5[1] of the Committee of Ministers to member States on Internet freedom Adopted by the Committee of Ministers on 13 April 2016 at the 1253rd meeting of the Ministers' Deputies
42. Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies)
43. Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies
44. Unboxing artificial intelligence: 10 steps to protect human rights (Council of Europe Commissioner for Human Rights 2019)
45. Venice Commission Guidelines on Freedom of Peaceful Assembly (2nd Edn) Adopted by the Venice Commission at its 83rd Plenary Session (Venice, 4 June 2010) Study no. 581/2010, CDL-AD(2010)020

8) Declarations, resolutions & Treaties

CEU eTD Collection

46. ‘OAS :: JOINT DECLARATION by the UN Special Rapporteur for Freedom of Opinion and Expression and the IACHR-OAS Special Rapporteur on Freedom of Expression’ (Oas.org, 2012) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=888&lID=1>
47. '473 Resolution On The Need To Undertake A Study On Human And Peoples’ Rights And Artificial Intelligence (AI), Robotics And Other New And Emerging Technologies In Africa - ACHPR/Res. 473 (EXT.OS/ XXXI) 2021' (Achpr.org, 2021) <https://www.achpr.org/sessions/resolutions?id=504> accessed 10 June 2021
48. African Charter on Human and Peoples’ Rights (adopted 27 June 1981, entered into force 21 October 1986) (1982) 21 ILM
49. DECLARATION OF PRINCIPLES ON FREEDOM OF EXPRESSION AND ACCESS TO INFORMATION IN AFRICA Adopted by the African Commission on Human and Peoples’ Rights at its 65th Ordinary Session held from 21 October to 10 November 2019 in Banjul, The Gambia.
50. European Convention on Human Rights (adopted 4 November 1950, entered into force 3 September 1953) 213 UNTS 1932
51. International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171
52. 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND ELECTIONS IN THE DIGITAL AGE' (Oas.org, 2020) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1174&lID=1>

CEU eTD Collection

53. 'OAS :: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET' (Oas.org, 2011) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1>
54. 'OAS :: TENTH ANNIVERSARY JOINT DECLARATION: TEN KEY CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE' (Oas.org, 2010) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=784&lID=1>
55. 'OAS :: TWENTIETH ANNIVERSARY OF THE JOINT DECLARATION: CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE' (Oas.org, 2019) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1146&lID=1>
56. Organization of American States (OAS), American Convention on Human Rights, "Pact of San Jose", Costa Rica, 22 November 1969
57. UN Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, Joint Declaration on Freedom of Expression and the Internet (2011)
58. Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III)

CEU eTD Collection