SPOTLIGHT ON

Artificial Intelligence & Freedom of Expression

#SAIFE

OSCE
The Representative on
Freedom of the Media

OSCE RFoM Non-paper on the Impact of Artificial Intelligence on Freedom of Expression

I. Introduction

A. Key Challenges to Freedom of Expression in the Age of Algorithms and AI

The online ecosystem has become the most participated in forum on the global level. Sometimes described as a "grand public forum"[1], it has fostered the flow of information, and transformed the ways in which journalists perform their work and how audiences consume and engage with media content. In that sense, freedom of expression and media freedom are increasingly exercised online. The international community has recognized and emphasized the importance of online spaces for societies, public discourse and democracy at large, and that the same rights that people enjoy offline must also be protected online, in particular freedom of expression. Human rights and fundamental freedoms apply both online and offline.[2]

Today, content is no longer created and disseminated solely by a limited number of media workers alone, who are bound by professional and ethical standards, but also by citizens. As a result, there is not necessarily any editorial control over a vast amount of published content. These processes have had a tremendous impact on audience behavior and information consumption. At the same time, internet intermediaries, especially social media platforms, have gained a dominant position. They are pivotal actors that undertake many functions of information management previously carried out by traditional gatekeepers, such as editors and publishers. This shift has particularly increased with the exponential growth of content shared and re-shared by internet users. The numbers speak for themselves: every single hour, 500 hours of videos are uploaded onto YouTube and 14.58 million photos on Facebook.[3]

Since the early stage of the internet, various technology solutions have been deployed to facilitate "many-to-many" online communication. These emerging technologies have been used to support the distribution of content to audiences (content curation), as well as to filter and take down illegal or otherwise unwanted content (content removal). These processes provide the basis for how society interacts online today.[4] Machine-learning technologies, automated algorithmic decision-making and other forms of artificial intelligence (AI) applied as automated tools and measures, are increasingly used to shape and arbitrate content online.[5] These practices have also recently gained more support, as states put increasing pressure on intermediaries to automate content moderation processes.[6] There is a trend by states to push platforms to remove content within a very strict time period, which can be as short as 24 hours or even one hour in some proposals.[7] On the other hand, some states push towards more transparency and regular auditing of these tools.[8]

[1] For instance, David Goldstone, "The Public Forum Doctrine in the Age of the Information Superhighway", 1995.
[2] United Nations Human Rights Council Resolution 20/8 "The promotion, protection and enjoyment of human rights on the Internet."
[3] More information available at: Omnicore statistics <https://www.omnicoreagency.com/facebook-statistics/> [last visited on 3 February 2020].
[4] For instance, a Facebook user may simply browse through his/her profile (cognition) or also re-publish (communicate) and re-work (co-operate) content. Social media platforms have led to the creation of unmediated social spaces with blurred lined between private and professional roles.
[5] See for example Google Jigsaw AI project: Perspective, available at: <https://www.perspectiveapi.com/#/home> [last visited on 3 February 2020].
[6] See concise list of AI regulatory initiatives: European Agency for Fundamental Rights, AI policy initiatives (2016-2019), available at: <https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives> [last visited on 23 February 2020].
[7] See, for example: Network Enforcement Act (Netzwerkdurchsetzungsgesetz or - NetzDG) adopted in Germany, 17 June 2017; Directive on Copyright and related rights in the Digital Single Market, (EU) 2019/790, European Parliament, 17 April 2019l EU Code of conduct on countering illegal hate speech online, European Commission, Twitter, Facebook, Microsoft and YouTube, 30 June 2016. https://www.europarl.europa.eu/news/en/press-room/20190408IPR35436/terrorist-content-online-companies-to-be-given-just-one-hour-to-remove-it.
[8] See, for example, recent proposal referred to as "Avia law" approved in July 2019 by the French National Assembly; White paper on Artificial Intelligence -A European approach to excellence and trust, European Commission, COM(2020) 65 final,19.2.2020; Algorithmic Accountability Act proposed in USA, 2019.

Today, algorithms and AI are used for a wide range of interventions, such as spam filters, detection of copyright infringements, chatbots, (editorial) data-analysis, or content ranking and distribution. Additionally, they have been deployed in policing not only online speech but also offline public spaces, for example with the help of smart video surveillance systems using facial recognition technology.[9] However, their impact on freedom of expression, both positive and negative, is still severely under-explored. While responsible implementation can benefit society, there is a genuine risk that commercial, political or state interests could have a deteriorating effect on human rights, in particular freedom of expression and media freedom.[10] Therefore, it is crucial to understand better the human rights implications of their use, and to ensure that algorithms and AI do not censor or have a chilling effect on free speech.

B. Key terminology and concepts

1. Internet Intermediaries, Platforms and Information Gatekeepers

The internet's uniquely layered structure creates three separate relevant categories of actors: those who create or publish information; those who are targeted by this information; and those who provide the platform for its distribution, internet intermediaries. Intermediaries play an essential role in enabling the flow of information between the two other actors without contributing to the content itself. However, they are in a unique position to prevent or mitigate risks that may be inflicted by the other two categories' illegal activity.[11] As such, they may, under certain circumstances, be liable as contributors, and are inevitably put under more pressure by both the potential claimants and law enforcement.[12] Intermediaries, as service providers, enable and manage interactions online, such as connecting users to the internet, hosting content online, and information-management, such as search engines and news aggregators, among others. Intermediaries may carry out multiple roles and, as different regulatory frameworks can apply depending on their function and services, this converging process is tied closely to a number of risks to freedom of expression.[13]

The concentration of roles and functions of intermediaries is often described as "platformization", emphasizing the dominant position of online platforms.[14] Online platforms are software-based facilities offering two- or even multi-sided markets where providers and users of content, goods and services can meet.[15] These platforms play a central role in the digital ecosystem, as an important means by which consumers find online information and online information finds consumers. This intermediary role gives platforms economic power to introduce a "new communication order", to shape the online experience of its users on a personalized basis and to filter what the user sees. For example, "like" and "share" buttons are now an inseparable element of almost any website, not only social media platforms.

In that sense, intermediaries, and especially social media platforms, are in a position of "information gatekeepers" as they engage in the selection of information to be published, in the ranking and editorial control over content, as well as in its removal.[16] As a result, they manage processes that could have a great impact on human rights and democracy at large.

[9] Among other work see: FRA paper on facial recognition technology, available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-facial-recognition-technology-focus-paper-1_en.pdf [last visited on 28 February 2020].
[10] Among other works see: Rikke Frank Jørgensen, Human rights in the age of platforms, MIT Press, 2019.
[11] Andrej Savin, EU Internet Law (second edition), Elgar European Law series, Edward Elgar Publishing, 2017, p. 143.
[12] Ibid.
[13] Council of Europe, Role and responsibilities of internet intermediaries, available at: <https://rm.coe.int/leaflet-internet-intermediaries-en/168089e572> [last visited on 23 February 2020].
[14] Examples of types of platforms include: communications and social media platforms; operating systems and app stores; audiovisual and music platforms; e-commerce platforms; content platforms, which may include content aggregators as well as software/hardware solutions; and search engines.
[15] A Digital Single Market Strategy for Europe - Analysis and Evidence, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015SC0100&from=HU.
[16] As described by E.B.Laidlaw: "The mechanisms include, for example, channeling (i.e. search engines, hyperlinks), censorship (i.e. filtering, blocking, zoning), value-added (i.e. customization tools), infrastructure (i.e. network access), user interaction (i.e. default homepages, hypertext links), and editorial mechanisms (i.e. technical controls, information content)." A framework for identifying Internet information gatekeepers, International Review of Law, Computers & Technology, 2010, p.16.

2. Content Moderation: Types of Content and Role of Intermediaries

Today, online platforms are called to play a more active role in monitoring content online and making decisions on the content's permissibility. However, while in many cases there is a justified reason to remove content that is manifestly illegal irrespective of its context,[17] such as child abuse material, the situation is more complex with regard to content that is considered as "harmful". Whether certain content reaches the level of "illegality" typically depends on the context in which it is presented. This is particularly the case for "hate speech" or "extremist" content. Thus, some forms of content might have a harmful effect, but they can still be protected under international human rights standards and should remain accessible online.

In practice, there are many strategies to manage and counter illegal and unwanted content. Different intermediaries perform various forms of content moderation, such as prioritization, deprioritization, promotion and demotion, monetization and demonetization of online content. This moderation typically takes place on three different levels. This non-paper will focus mainly on content removal and content curation as the most visible techniques of content moderation.

1. In many instances, various types of automated measures, which include algorithms and AI, are deployed as a first level of moderation, to check content through so-called "upload" filters. Such upload assessments vary across platforms, depending on the technology used, and internal policies. If content has characteristics of predefined categories of "unwanted" material, algorithms and AI are supposed to automatically block such content from being published.
2. Due to content overload and attention scarcity, platforms regularly deploy automated tools to moderate content on the second level, to assess which piece of content will be "visible" to which particular user for how long. In this process, AI "ranks" content based on multiple criteria, such as who posted the information, previous interaction with the content, or a similar type of the content, or previous interaction by a "similar user".[18] It is usually not made public which criteria are mixed in the algorithmic decision-making. This means that "black boxes"[19] employing machine-learning technologies decide which content is available to whom.
3. On the third level, and mostly with human intervention, content moderation is based on reporting mechanisms. These are often established under internal policies of companies, also referred to as Notice-and-Take-Down procedures (NTD). In these cases, any user may report "inappropriate" content (based on the platform's internal rules), which triggers a reviewing procedure. Based on such reports, resolved by human moderators and/or AI, problematic content might be removed and the accounts of the poster might be temporarily or permanently blocked.

Insufficient transparency of each of these levels and their processes, both in terms of criteria involved in the decision-making process and the due process itself, are often seen as one of the key challenges of the use of algorithms and AI in content moderation.

To assess online content and decide on its accessibility, intermediaries have adopted a number of internal rules and procedures (e.g. Community Standards and Terms of Services), which serve as a set of guidelines to "judge" content. These rules define which content is considered to be harmful or unwanted, which does not necessarily equate to "illegal" according to national legislation or international frameworks. Therefore, online content regulatory models are governed by the rules set forth by private profit-oriented entities rather than international human rights standards,[20] which set out the criteria that can justify limitations to speech.[21] The consequential lack of consistency and clarity, as well as the pressure on platforms to make swift decisions, in terms of whether certain content can be categorized

[17] Council of Europe, Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries CM/Rec(2018)2, 7 March 2018, para. 1.3.2.

[18] When deciding which content to show to individual users, the following factors are important (not exclusive): character of the person who wants to distribute the content (user, page, group, business, etc.); form of content (text, video, audio, photo, etc.); interest in content from other network users; automatically generated user profile; direct user requests (hide, starred, etc.); special relationships between content and users (tagging, etc.); busting - sponsorship of content by distributors.

[19] Frank Pasquale, The black Box society: The secret algorithms that control money and information, Cambridge, MA: Harvard University Press, 2015, p.8.

[20] United Nations, David Kaye, "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression." A/HRC/38/35, April 6 2018.

[21] For example, sometimes there is a legitimate reason, based on the type of platform (e.g. it is permissible for certain social media, such as Mumsnet to only allow discussion related to motherhood or LinkedIn to allow professional networking and prohibit the use of their platform for other purposes).

as unlawful under national criminal laws, is particularly concerning.[22] The removal of illicit content by platforms raises the issue of the absence of judicial oversight. Without judicial review, there is no proper remedy or accountability mechanism in place. This shift of responsibility from states to intermediaries has already created a significant impact on the enjoyment of human rights, especially freedom of expression.

3. Artificial intelligence

While there is no universally agreed definition of AI, and a need for more discussions, many refer to it as systems designed by humans to operate with varying levels of autonomy, which, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.[23] AI systems act by perceiving their environment through data acquisition, interpreting the collected data, reasoning on this knowledge, or processing the information derived from this data and deciding the best action(s) to take to achieve a given goal. Self-learning forms of AI can also adapt their behavior by analyzing how the environment is affected by their previous actions.[24] Various approaches to AI suggest that AI is an umbrella term for processes that essentially delegate decision-making and execution activities, partially or completely, from humans to software systems.

AI is based on algorithms – a set of human-designed instructions with encoded procedures for transforming input data into the desired output, based on specific calculations.[25] Advanced AI techniques include machine learning, which is often defined as the ability of AI systems to adapt or improve performance autonomously over time without being explicitly programmed in that way. The majority of AI technologies today are in fact machine-learning systems automating a variety of sophisticated tasks, previously presumed to require human cognition. The prerequisite for such an advancement of machine learning is access to big data, extremely large datasets characterized by the volume (amount), the velocity (speed) and the variety of data.[26] After the initial human act of creating the “code” and assigning a specific task, the process of machine learning regularly begins with the observation of large datasets and the application of a statistical process to look for patterns in data and make more precise decisions in the future.

AI therefore has the capacity to extracts actionable knowledge from available data via mathematical models and without meaningful human intervention. Without a deeper understanding of data and context, this can be particularly problematic, for example because of underrepresentation in datasets, inaccurate or missing data, or because of inaccurate causation and correlation of datasets. Further complexity can arise due to the lack of transparency and explainability of algorithmic decision-making. For instance, it is hard to trace back and challenge decisions if satirical content on a matter of public interest is taken down, or states pressure platforms to remove “extremist content,” platforms are purging vital evidence of human rights violations, for example in the context of conflicts.[27]

II. Main Characteristics of AI Processes behind Content Removal

Most of the algorithms and AI applications deployed by intermediaries are in some way tied to the question of scale and complexity of “networked publics”.[28] In this relationship, they should “solve” problems of scale and “subjectivity”, that is to say personal-biases.[29] In practical terms, algorithms and AI are often deployed to identify and remove specific content. Thus, to remove the intended content, they would need to analyze different aspects related to this particular content, which is a complex task, especially given its application across regions and languages. For example, detecting bullying online requires an understanding of the relationship between two or more users, their age, the number of

[22] E.g. NetzDG or Loi d'Avia.
[23] The OECD Principles on Artificial Intelligence, adopted in May 2019.
[24] European Commission’s High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019.
[25] T. Gillespie, The relevance of algorithms, Media technologies: Essays on communication, materiality, and society, MIT Press, 2014, p. 167. Defined as "encoded procedures for transforming input data into the desired output, based on specific calculations."
[26] H. Surden “Machine Learning and Law”, 2014, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2417415.9 “Preparing for the Future of Artificial Intelligence”>.
[27] New York Times, YouTube Is Erasing History: Under pressure to remove “extremist content,” platforms are purging vital human rights evidence, 23 October 2019. https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html.
[28] Dana Boyd, “Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications”, in Zizi Papacharissi (ed.), A Networked Self: Identity, Community, and Culture on Social Network Sites (Routledge, New York, 2011), p. 39.
[29] T. Gillespie, Custodians of the Internet, Yale University Press, 2018, p.97.

exchanged messages, the nature of their connection, as well as previous interaction history and shared connections.[30]

When removal relies on algorithms and AI, studies show that automated decisions can fail to understand the contextual nuances behind pieces of content.[31] The identification of context dependent content requires a proper understanding of societal, political, historical and cultural nuances, in order to recognize the harm that such content may potentially carry, and whether it should be removed based on human-designed instructions. There are numerous examples of how automated tools, such as algorithms and AI, struggle to detect illegal content that requires contextual understanding, while filtering and taking down perfectly legitimate content from platforms.[32] At the same time, it is relevant to emphasize that the removal of hateful content does not remove the underlying hate. Thus, the problem could be exacerbated if users are blocked immediately, and thereby pushed out of the open public discussion, which could encourage them to join dubious platforms and conspiracy theories.

Furthermore, cultural and legal differences across the world put into question the application of systems trained on data from one region to work effectively in other areas. Thus, there are also often significant shortcomings in automated tools that emphasize the importance of genuine human involvement, sometimes referred to as "human in the loop", that should guarantee that the efficiency of algorithms and AI will remain amenable to human intervention.[33]

Platforms are operating as "speech police" based on vague "community standards" supported by algorithms and AI. As a result, they regularly fail to ensure that users can understand what has been taken down and why. Instead, they should inform users by being open about their takedown processes and results,[34] and should put in place clear, simple procedures for users to challenge takedowns with the support of human reviewers of automated decisions.

A. Security Threats

AI and algorithms are often deployed to detect content that is – under most laws and platform standards – perceived as threatening to national security. Governments and legislators are increasingly pressuring intermediaries, most notably platforms, to take a more proactive role in policing "terrorist" or "extremist" content, and to develop proactive automated measures to identify content falling under this category in a very short time frame.

However, evidence and researched based justification for swift removal of online content is currently missing. There is a lack of evidence that demonstrates that the successful removal of "terrorist" content online in fact results in reduced security threats. In the same vein, there are also only a few studies[35] on the effectiveness of algorithms and AI specifically designed to identify illegal content. In addition, there is always a certain "grey" area that, due to particular context and nuances, calls for a sophisticated and balanced assessment and there are cases of a "false negative" when a system incorrectly identifies illegal content to be "innocent", or a "false positive" when a system removes "innocent" content.[36]

Another concern is that, in order to address such illegal content sustainably, the engagement of law enforcement is required. For this reason, instant removal of such content can be seen as an extension of concealed state action.[37]

Additionally, some content removal operations are linked to broader security measures, in order to protect the integrity of the platform, integrity of service and management of traffic data. This includes

[30] OFCOM and Cambridge Consultants, Use of AI in Online Content Moderation (2019).
[31] This is the case, for instance, of automated takedowns of political speech and marginalized voices based on copyright upload filters. See Reda, J. (2017). When filters fail: These cases show we can't trust algorithms to clean up the internet, available at: < https://juliareda.eu/2017/09/when-filters-fail/> [last visited 11 February 2020].
[32] For example: Tech Dirt, YouTube Takes Down Ariana Grande's Manchester Benefit Concert On Copyright Grounds, 17 June 2017, available at: <https://www.techdirt.com/articles/20170606/17500637534/youtube-takes-down-ariana-grandes-manchester-benefit-concert-copyright-grounds.shtml>[last visited on 11 February 2020].
[33] https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.
[34] EFF, Platform Censorship: Lessons From the Copyright Wars,
[35] OFCOM and Cambridge Consultants, Use of AI in Online Content Moderation (2019), See also: B.Ganor Artificial or Human: A New Era of Counterterrorism Intelligence?, Studies in Conflict & Terrorism, 2019.
[36] OFCOM and Cambridge Consultants, Use of AI in Online Content Moderation (2019).
[37] Sarah Koslov, Incitement and the geopolitical influence of Facebook Content Moderation, Georgetown Law Technology Review (183) 2019, p.194.

measures against inauthentic behavior, commercial spam, bots or impersonation.[38] The application of algorithms and AI in these operations also raises free speech concerns. However, more transparency and study is needed to understand the possible impact on legitimate content and on freedom of expression.

B. “Hate Speech”

There is no uniform definition of “hate speech” under international human rights law, and the detection of hate speech content is subject to societal, political, historical and cultural nuances. In addition, the wide range of hateful expressions requires different responses based on the severity of the speech in question.[39] Community guidelines of social media companies fail to reflect complex nuances and, therefore, their implementation through automated measures supported by algorithms and AI can lead to the removal of perfectly legitimate content.[40] Besides the problem of over-removals, it is also concerning if all hateful content remains online. This can have a collectively harmful effect, particularly on marginalized and underrepresented groups. In that sense, hate speech can have a silencing effect. Finally, to counter hate speech, which is first and foremost a societal problem, diverse initiatives and policies need to be undertaken by numerous actors. An automated regulation of hate speech can otherwise have a detrimental impact on public discourse and lead to a chilling effect and self-censorship.

As context plays a salient role in the assessment of content, a simple analysis of words and phrases will rarely result in an accurate assessment. AI systems struggle to recognize figurative speech, to discern mockery from illicit hate speech, and offensive language that sometimes follows heated public debate over issues of public importance. Facebook’s 2018 report agreed that technology still does not work that well in terms of detecting contextually complex hate speech, and that it has to be supported by human reviewers.[41] However, Facebook recently claimed that, using machine learning, it has developed a new type of detection technology that can identify and flag hate speech using several different methods,[42] improving its success rate of automated measures.[43]

There is an additional risk when AI is trained on data from different jurisdictions, which can create unwanted consequences in other societies with different cultural communication rules. A recent study of Twitter content, written in standard American English and African American English, has demonstrated evidence of systematic racial bias of tweets written in African American English. The study concluded: “Consequently, these systems may discriminate against the groups who are often the targets of the abuse we are trying to detect.”[44]

III. Main characteristics of AI processes behind content curation

The underlying business models of many online platforms rely heavily on user attention and engagement, which are considered and treated as an economic resource. The time users spend on online platforms is one of the key factors that determines platforms’ economic gain. As a result, most online platforms curate their news feeds and search results in order to increase engagement and time spent on the platform. They aim to increase profit by amplifying sensational or potentially harmful content, so-called “clickbait” content. Against this backdrop, algorithmic and AI solutions that determine

[38] https://ec.europa.eu/digital-single-market/en/news/first-results-eu-code-practice-against-disinformation.

[39] International human rights law distinguishes between a) severe forms of “hate speech” that States are required to prohibit, including through criminal, civil, and administrative measures, under both international criminal law and Article 20(2) of the ICCPR; b) other forms of “hate speech” that States may prohibit to protect the rights of others under Article 19(3) of the ICCPR, such as discriminatory or bias-motivated threats or harassment; and c) “lawful hate speech” which nevertheless raises concerns in terms of intolerance and discrimination, meriting a critical response by States. C.f. ARTICLE 19, ‘Hate Speech’ Explained, 2015.

[40] Shirin Ghaffary, The algorithms that detect hate speech online are biased against black people, VOX, 15 August 2019, available at: <https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter> [last visited on 11 February 2019].

[41] Facebook Publishes Enforcement Numbers for the First Time, https://about.fb.com/news/2018/05/enforcement-numbers/.

[42] One method involves detecting and automatically removing content that matches existing hate speech violations in database. Another method involves proactively detecting potentially violating content and then giving it a score according to its similarity to content already removed for violating hate speech policy. Starting in Q2 2019, FB systems began removing posts automatically when they received very high scores or matched existing hate speech in database.

[43] Facebook Community Standards Enforcement Report: in Q1 2018 only 38% of the hate speech were removed automatically while this percent has raised to 80.2% in Q3 2019 https://transparency.facebook.com/community-standards-enforcement#hate-speech.

[44] T. Davidson et al, Racial Bias in Hate Speech and Abusive Language Detection Datasets, 29 May 2019.

trending topics and recommended content are not neutral, but reflect corporate and profit-oriented values.[45]

Besides amplifying the reach of “clickbait” content, AI provides users with content that is not merely based on their data but also on the characteristics of the group to which – according to the AI – the user belongs. It is essential to understand that AI, in this context, is merely a tool governed and operated by private companies, while the ranking of content is regularly based on users’ preferences and behavioral data, and again, to increase the time the users spend on the platform.

Drawing upon this analysis, it is evident that dominant online platforms have changed the ways that we access, receive and impart information, which lays the foundation for how we form our opinions. Due to a lack of transparency as well as awareness, most AI processes behind content curation lack the scrutiny of users and the general public, including researchers and regulators. However, content curation is an essential issue for freedom of expression. It needs to be addressed primarily by state actors, but also by non-state actors, including intermediaries that have a positive obligation to create an enabling environment that ensures diversity and pluralism of sources and views.[46]

A. Challenges to pluralism and diversity

In general, algorithms and AI are often deployed to categorize individuals into groups and to determine their particular political and commercial preferences. Based on this assessment, AI targets each individual with specifically curtailed content. As a result, such a process of social sorting may expose users to similar content, which tends to correspond with, or strengthen, their existing interests, and amplify their views and preferences, rather than to offer a variety of (alternative) information and sources that challenge and oppose their views.[47]

This process is often referred to as an “echo chamber”, a process whereby “individuals are increasingly cocooning themselves in the informational and communicational universe of their own creation”.[48] While social media has provided minorities and other marginalized voices with a myriad of opportunities to connect and engage, “echo chambers” are especially worrisome as they can reinforce societal power balances.[49]

At the same time, media outlets and journalists are struggling to adjust to the new dissemination practices underpinned by these AI processes. Under the “new communication order”, intermediaries, most notably social media platforms, decide, with the use of algorithms and AI, which information particular users will have the opportunity to access. Thus, while it may be easy to speak in cyberspace, it remains difficult to be heard.[50] Aspects of gender inequalities also need to be taken into account and explored further, in particular with regards to inequalities in access and production of information, as well as how AI technologies can reproduce gender biases.[51] Against this background, intermediaries as information gatekeepers are in a position to potentially hinder the public’s right to access pluralistic and diverse information.

[45] H. Bloch Wehba, Automation in moderation, Cornell International Law Journal (forthcoming), 2020, p.6.

[46] See for example, The United Nations (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, 20th anniversary joint declaration: challenges to freedom of expression in the next decade, 10 July 2019, para 1.

[47] Studying algorithmic agents and the ways in which they potentially “shape” the opinion-making process is tied to a number of ethical, legal and methodological challenges. Thus, this field is still exploring the right methodological approach. For further discussion see: B. Bodó et al., Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents, Yale Journal of Law and Technology, 19, 2017., See also a study on “personalized communication”: B. Bodó et al., Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization, Digital journalism, 2019.

[48] Tarlach McGonagle, Minority rights, freedom of expression and of the media: dynamics and dilemmas (Intersentia, Cambridge, 2011); Mike Cormack and Niamh Hourigan (eds.), Minority Language Media: Concepts, Critiques and Case Studies (Multilingual Matters Ltd., Clevedon, etc., 2007), p.157.

[49] Bojana Kostic and Tarlach McGonagle, How Social are New and Social Media for National Minorities? Perspectives from the FCNM, European Yearbook of Minority Issues (Vol.16), 2019, p.11-14.

[50] M. Hindmann, The myth of digital democracy, Princeton and Oxford University press, 2009, p.142.

[51] Noble, Safiya. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. 10.2307/j.ctt1pwt9w5, WIRED, “Machines Taught by Photos Learn a Sexist View of Women” https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/; Collett, Clementine and Dillon, Sarah (2019). AI and Gender: Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence. http://lcfi.ac.uk/media/uploads/files/AI_and_Gender_4_Proposals_for_Future_Research_yaApTTR.pdf.

However, recent research findings about the actual role of echo chambers and their impact on democratic discourse are rather inconclusive.[52] In any case, it is a fact that AI ranking practices, which often go hand-in-hand with political and commercial behavioral targeting, have changed the way information is consumed and may impact the way opinions can be formed.

B. Impact of surveillance, including surveillance capitalism, on freedom of expression

Machine-learning technologies require large amounts of data. This fact has impacted business models, especially in the media field. Information, as well as its curation via apps and social media platforms, is offered to users “for free” in exchange for their behavioral data and other data externalities. It has become more lucrative for internet companies to collect users’ data than to collect users’ money. In addition, large amounts of both personal and non-personal data enable data mining and therefore become a competitive advantage. Consequently, the development and sustainability of the online media and e-commerce market also need to be assessed from the perspective of competition law. As dominant platforms are the largest holders of data, there is a need for such data to be openly accessible, in order to enable free competition and further innovation while avoiding the network effect.[53]

This business model of intermediaries enables a profiling of individuals, even if individual citizens undertake all precautions to protect their privacy and shield their data from data processing. Personal digital footprints, even if small, will be sufficient for various online services powered by AI to classify users in already developed profiles and to predict needs based on the data of other people, supposedly similar to them. Too often, citizens are neither informed that these processes are taking place nor are they aware of how they work and their potentially discriminatory aspects.

Constant surveillance, online as well as offline, has a chilling impact on human rights, in particular the right to privacy and freedom of expression. This could be particularly true if state actors introduce smart video surveillance technologies in public spaces with facial recognition capacities, which could endanger not only freedom of expression but also freedom of assembly and other human rights. Special concerns arise if citizens’ data that is in the possession of state institutions is merged with the digital profiles of citizens to create AI-powered social credit systems.[54] There is a need to further explore the link between online profiling and surveillance and state surveillance as the online data infrastructure is constructed to service a data-driven business model, which could facilitate state surveillance.

These permanent surveillance practices, coupled with profiling, can have dangerous consequences on how journalists perform their work as well as on their safety. These risks are evident in connection with the protection of journalists’ sources and whistleblowers. However, there are also less evident but equally threatening issues, such as the use of facial recognition to identify journalists, for example, reporting from protests, or tracing back the digital footprints of individual journalists, especially those of marginalized groups. Additionally, when combined with restrictive legislation, algorithms and AI, which track newsgathering activities, can have a detrimental impact on newsgathering and investigative journalism.[55]

IV. Conclusion: Risks posed by AI to freedom of expression

This non-paper outlines the ways in which non-state and state actors deploy algorithms and AI to address concerns stemming from the online ecosystem that are able to make semi-autonomous decisions on filtering, ranking, removal and blocking of content. Automated measures engage with a wide spectrum of content, from “extremist” and terrorist content to hate speech and potentially harmful,

[52] Elizabeth Dubois and Grant Blank, “The echo chamber is overstated: the moderating effect of political interest and diverse media”, 21(5) Information, Communication and Society (2018).

[53] The network effect is a phenomenon whereby increased numbers of people or participants improve the value of a good or service. A social media platform might therefore grow in popularity because it has achieved a critical mass of users and new users will be deterred from using another platform.

[54] This process is already taking place in China: “China’s social credit system is the epitome of the disastrous consequences of technological advancement without a commensurate commitment to human rights.”, Oxford Human Rights Hub, 6 September 2019, available at: <https://ohrh.law.ox.ac.uk/the-human-rights-implications-of-chinas-social-credit-system/> [last visited on 12 February 2020].

[55] This can capture situations when journalists are trying to take informed views about terrorist groups’ motivations and actions without the intent to commit a terrorist offence, c.f. Article 575(1) of the Spanish Penal Code.

but lawful, content. Through the process of profiling, AI curtails online public forums and decides which information users are able to access online, while exacerbating the existing risks of surveillance.

Key challenges to freedom of expression stem from the lack of transparency and explainability of algorithms and AI, from the outsourcing of judicial responsibilities and protection of human rights to private entities, as well as the lack of oversight, accountability and correction mechanisms. It is essential that any measure, technological and regulatory, that seeks to manage or control "public forums" is human rights-based, proportionate, and incorporates checks and balances, in order not to limit freedom of expression, media pluralism, the free flow of information and other fundamental rights.

It is therefore crucial, as a first step, to establish and promote a clearer understanding of the policies and practices in place in the use of AI. It is equally important to understand better the impact they have on the future of media and quality information and the realization of human rights online. As a next step, policy recommendations need to be developed to ensure that freedom of expression and media freedom is safeguarded when using machine-learning technologies, such as AI. Looking forward, it is crucial to:

- promote a better understanding of the algorithmic decision-making and AI policies/practices in place (by both state and non-state actors) and how they impact freedom of expression;
- initiate a multi-stakeholder dialogue (including with industry and states, addressing their legitimate concerns to address security threats and hate speech online);
- develop recommendations to mitigate the negative impacts of automated tools and to prevent the infringement of free speech and media freedom;
- research and assess how automation affects media freedom and how journalism can benefit from algorithms and AI;
- measure the impact of legislation or policies mandating removal of content in short time periods on deployment of algorithms and AI by platforms;
- explore discriminatory effects of content moderation technologies, especially in the context of digital inclusion and marginalized voices;
- conduct studies on the effectiveness of automated measures specifically designed to identify illegal content, as well as to explore alternative measures to combat hate speech, for instance, how interface design impacts users' behaviors and how algorithms and AI could be used to counter hate speech;
- map out the current use of machine-learning technologies by law enforcement agencies and their potential impact on freedom of expression; and
- organize discussions and workshops about the positive and negative implications of automated measures for identification of illegal content on online platforms specifically targeting law enforcement in selected countries, as well as on how they impact freedom of expression.