SPOTLIGHT ON

ARTIFICIAL INTELLIGENCE AND FREEDOM OF EXPRESSION

#SAIFE #AIFreeSpeech

osce
The Representative on
Freedom of the Media

Spotlight on Artificial Intelligence and Freedom of Expression #SAIFE

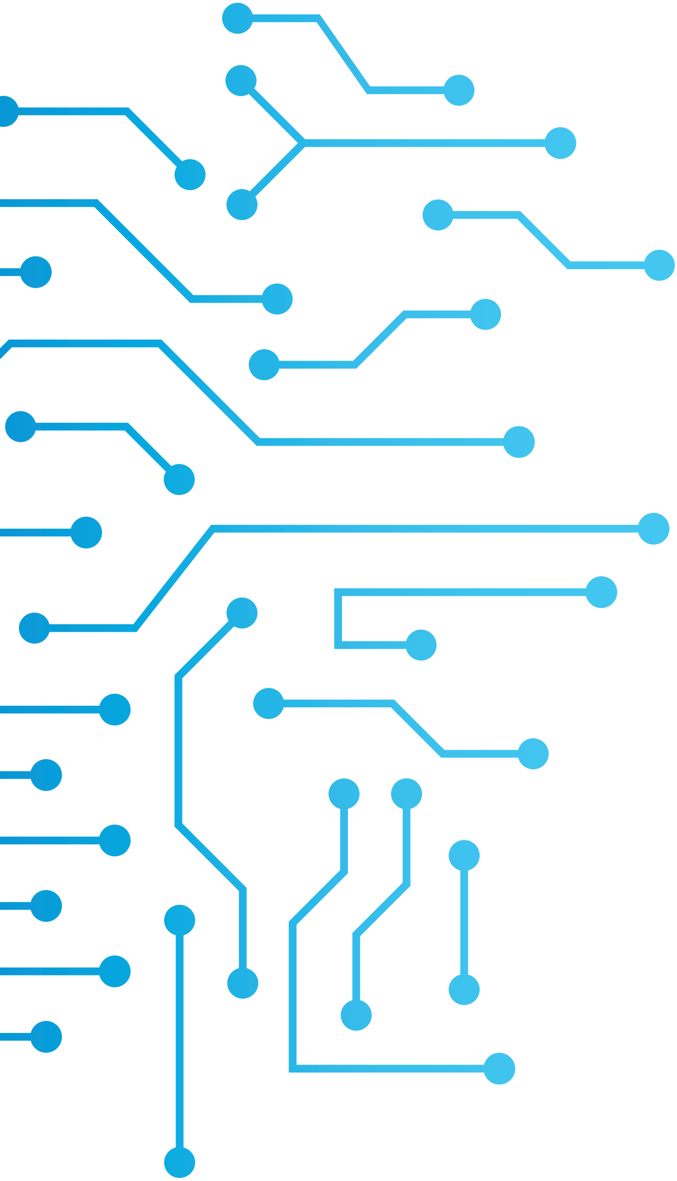
Author: Barbora Bukovska
Contributors: Amy Brouillette, Barbora Bukovska, Julia Haas, Fanny Hidvégi, Nani Jansen, Lorena Jaume-Palasi, Carly Kind, Bojana Kostic, Đorđe Krivokapic, Emma Llanso, Victor Naumov, Eliška Pírková, Krisztina Rozgonyi, and Martin Scheinin
Edited by: Julia Haas
Design & Layout by: Peno Mishoyan

Commissioned by Harlem Désir, the OSCE Representative on Freedom of the Media

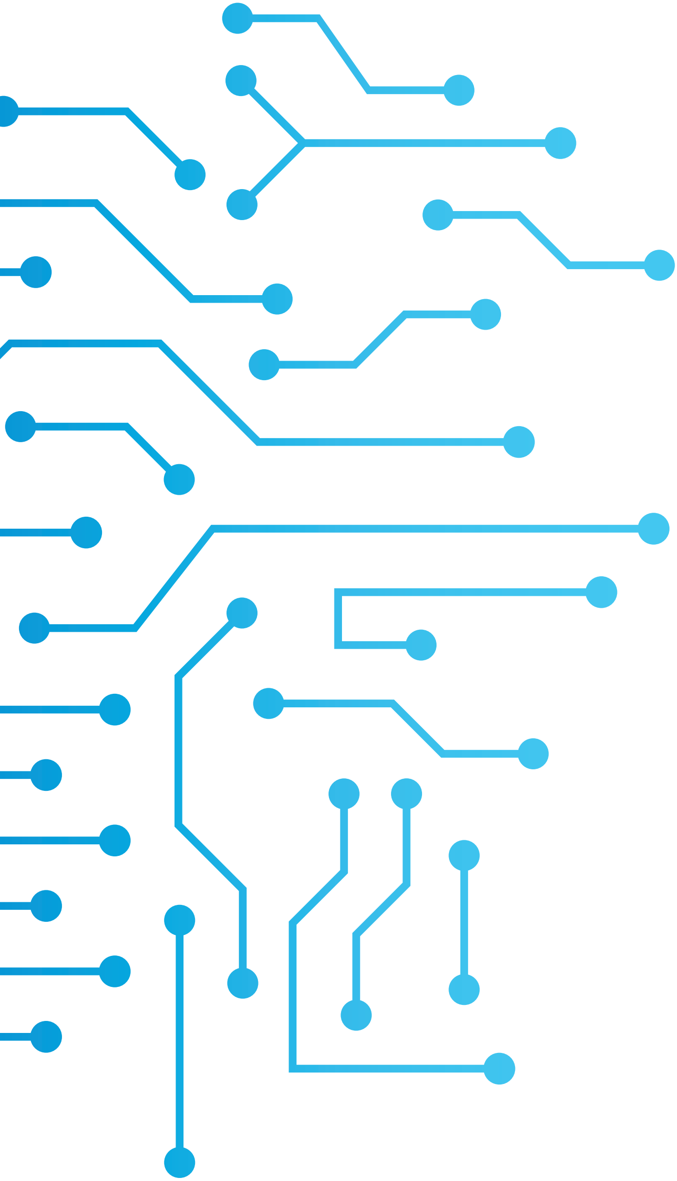
Wallnerstrasse 6
A-1010 Vienna, Austria
Tel.: +43-1 514 36 68 00
Fax: +43-1 514 36 6802
E-mail: pm-fom@osce.org
http://www.osce.org/fom

ISBN: 978-3-903128-54-5

Table of Contents



Foreword



Foreword

Since the early stages of the internet, various technologies have been deployed to enable and facilitate online communication. Over the last years, machine-learning technologies, such as artificial intelligence (AI), have become increasingly important tools for shaping and arbitrating online information. AI-powered tools rely on the collection and processing of vast amounts of data, which in turn are frequently monetized, and are often used for detecting, evaluating and moderating content at scale, oftentimes with a view to identify and filter out illegal and potentially harmful content. At the same time, AI is used for ranking, promoting and demoting massive amounts of content online.

AI has become a major reality of the information sphere, of online content moderation, and of the prioritization of information. How do we want to use algorithms and machine-learning tools in such an important domain of our lives? How will these tools be controlled and by whom? There is a genuine risk that such technologies could have a detrimental impact on fundamental freedoms, especially when driven by commercial or political interests. The use of AI could seriously jeopardize the enjoyment of our human rights, in particular the freedom of expression. Moreover, given that most AI-powered tools lack transparency and accountability, as well as effective remedies, their increasing use risks exacerbating existing challenges to free speech, access to information and media pluralism.

For these reasons, last year, I launched a project to put a spotlight on AI and freedom of expression (#SAIFE). This #SAIFE initiative focuses on the profound impact that the use of AI has on seeking, receiving and imparting information and ideas. In March 2020, I published an introductory non-paper to promote a clearer understanding of the policies and practices of governments, regulators, and internet intermediaries in their use of AI. I hope that the introductory non-paper contributed to unveiling the complexity of the impact that these technologies can have on freedom of expression and access to information.

Based on the introductory non-paper and discussions within the project, I am pleased to present this Paper now, which I hope will provide guidance for further discussions and actions to prevent any intentional or unintentional negative implications of the use of AI on free speech. It is only through close co-operation at both national and international levels, as well as among various stakeholders, including civil society and the tech industry, that a responsible and human rights-friendly use of AI can be ensured. Only then can illegal and potentially harmful content, such as speech presenting “security threats” or racism, anti-Semitism and “hate speech”, be tackled, and pluralistic democratic discourse be promoted without harming democracy as such.

While many international actors discuss important questions around the impact of AI on the enjoyment of human rights, my Office’s #SAIFE project focuses specifically on the impact of AI on freedom of expression, including the impact of AI on journalistic work and the overall media environment. While the challenges vary from country to country, with diverse national practices and different internet intermediaries prevalent, it is clear that challenges to free speech stemming from an increased use of AI exist across the entire OSCE region.

I hope that this publication, with its preliminary recommendations, will serve as a useful reference for much-needed discussions and for identifying a way forward to safeguard free speech when deploying AI.

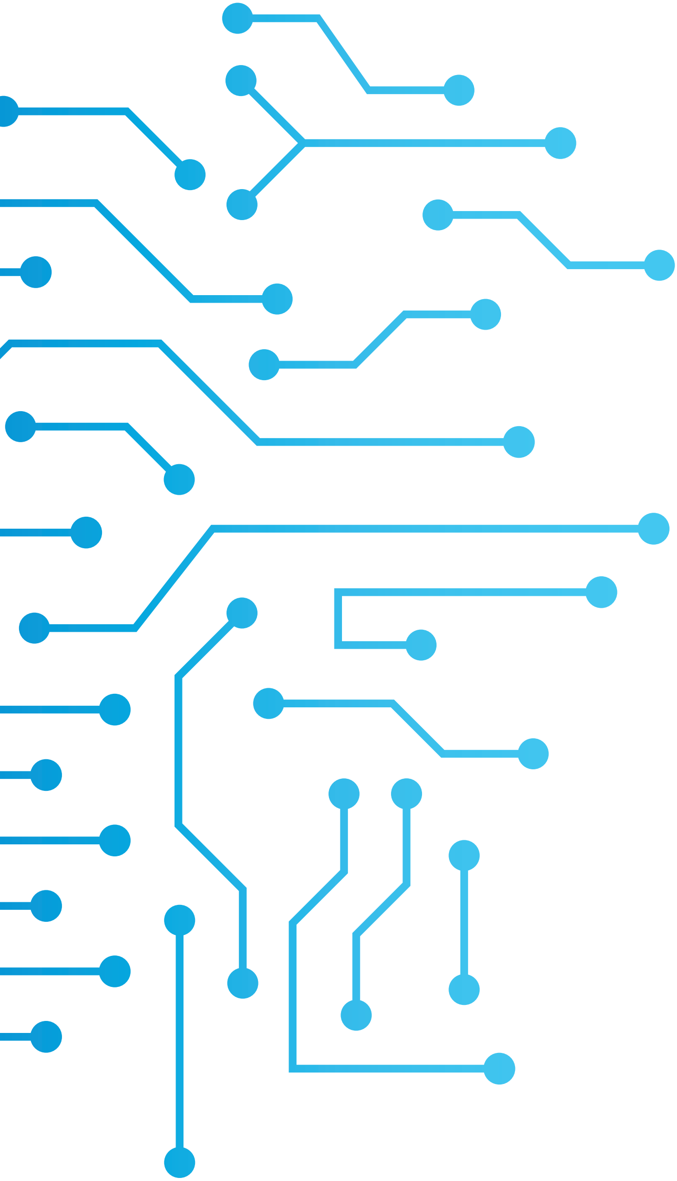
As a next step, this Paper, which will be discussed at an online event on 8 July,[1] will be followed by a public consultation phase, which will provide the foundation for a further elaboration of concrete policy recommendations.

I want to sincerely thank all the contributors to this paper,[2] especially Barbora Bukovska who drafted this Paper, and Đorđe Krivokapic, the main author of the introductory non-paper that has been integrated into this publication. Finally, a special thanks to Julia Haas of my Office, and all colleagues who have contributed to making this publication possible.

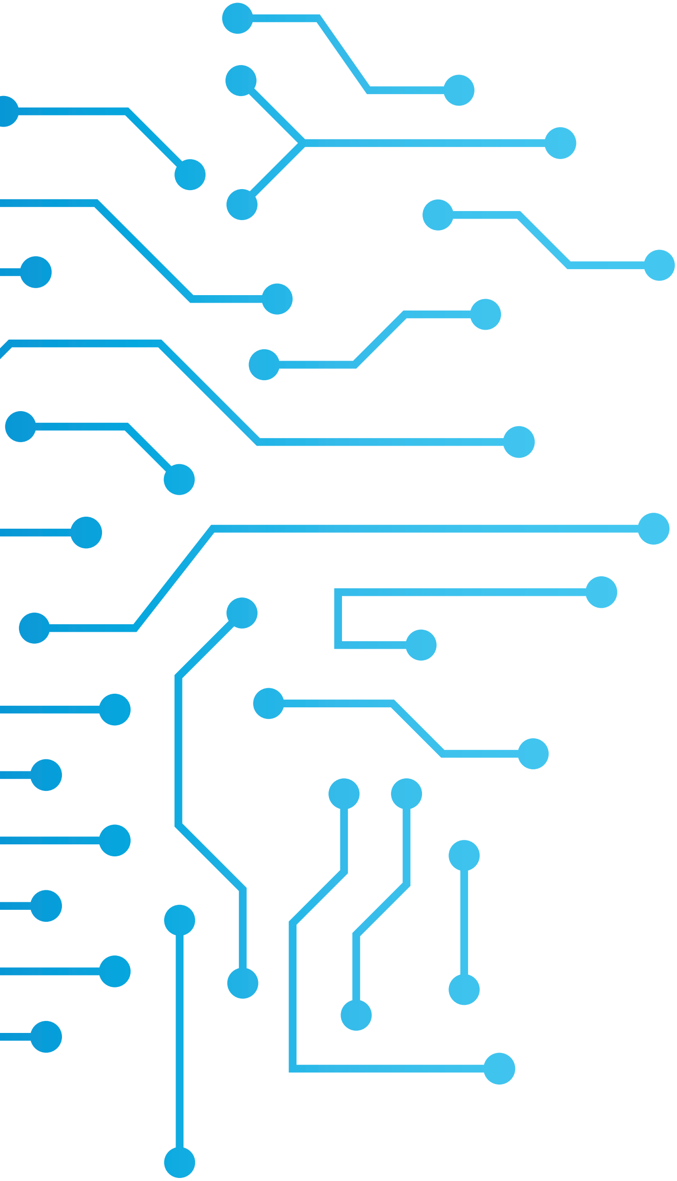
3 July 2020
Harlem Désir, OSCE Representative on Freedom of the Media

[1] OSCE RFoM event on The Rise of Artificial Intelligence & How it will Reshape the Future of Free Speech, 8 July 2020.

[2] Amy Brouillette, Fanny Hidvégi, Nani Jansen, Lorena Jaume-Palasi, Carly Kind, Bojana Kostic, Emma Llanso, Victor Naumov, Eliška Pírková, Krisztina Rozgonyi, and Martin Scheinin.



Executive Summary



Executive Summary

Artificial intelligence (AI) – a broad concept used in policy discussions to refer to many different types of technology – greatly influences and impacts the way people seek, receive, impart and access information and how they exercise their right to freedom of expression in the digital ecosystem. If implemented responsibly, AI can benefit societies, but there is a genuine risk that its deployment by States and private companies, such as internet intermediaries, could have a deteriorating effect on human rights.

When considering the impact of AI on freedom of expression, underlying issues that need to be taken into account include the business model of most internet intermediaries, based on the collection and processing of massive amounts of data about their users. Individual users are constantly surveyed for their online behaviour. Personal digital footprints, even if small, are sufficient for various online services powered by AI to classify users in pre-developed profiles and to predict customer "needs" based on the data of other, supposedly similar people. Most individuals are neither informed that these processes are taking place, nor are they aware of their workings and potentially discriminatory aspects.

Internet intermediaries use AI in various stages of content moderation and content curation on their platforms: from uploads, to making certain content more visible to users, to the removal of content. The deployment of AI in these processes creates risks for freedom of expression. Whether particular content should be removed (either for violation of the law or community guidelines) often depends on a number of issues, including the context, which is difficult to assess without human involvement. This could potentially lead to the removal of legitimate expression, or failure to remove content that could have a detrimental impact on many users. When it comes to content curation, the monetization of users' attention and engagement has had a great impact on diversity and pluralism online. This is all the more poignant, since the criteria that internet intermediaries apply are usually not open to the public. These problems cut across all types of content but are most prominent in the area of “security threats” and “hate speech”.

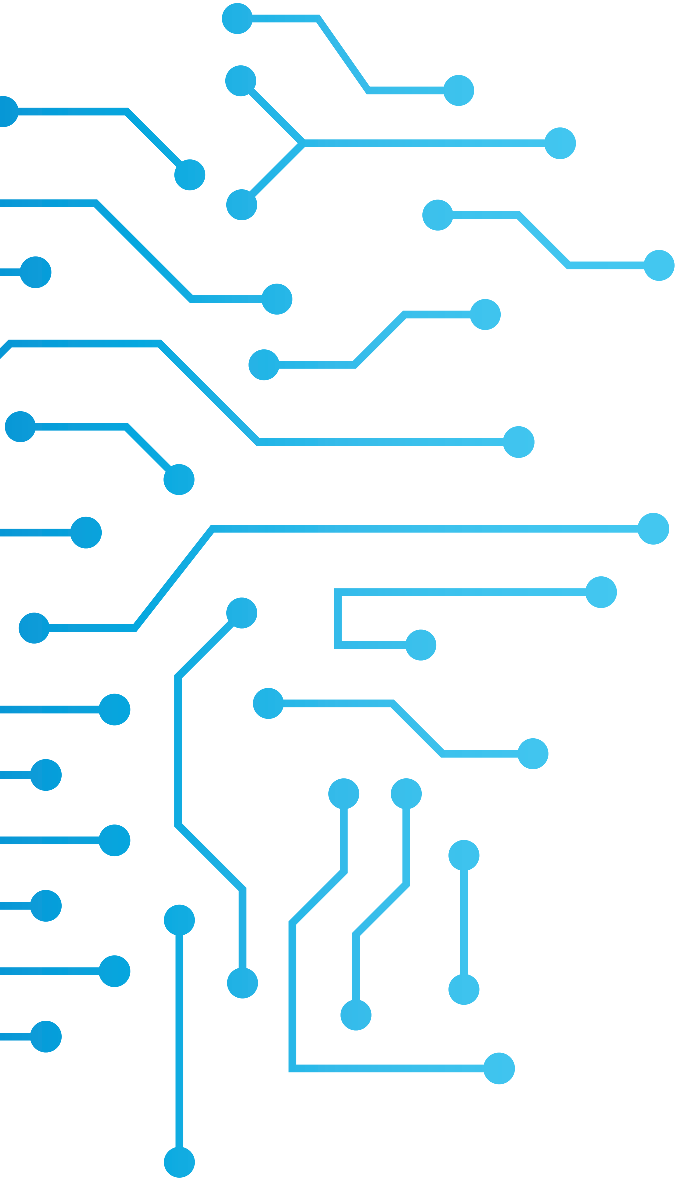
This Paper addresses these challenges, building on the initial work of the OSCE Representative on Freedom of the Media. It maps the key challenges to freedom of expression presented by AI across the OSCE region, in light of international and regional standards on human rights and AI. It identifies a number of overarching problems that AI poses to freedom of expression and human rights in general, in particular:

- The **limited understanding** of the implications for freedom of expression caused by AI, in particular machine learning;
- **Lack of respect for freedom of expression** in content moderation and curation;
- State and non-State actors **circumventing due process and rule of law** in AI-powered content moderation;
- **Lack of transparency** regarding the entire process of AI design, deployment and implementation;
- **Lack of accountability** and independent oversight over AI systems;
- **Lack of effective remedies** for violation of the right to freedom of expression in relation to AI.

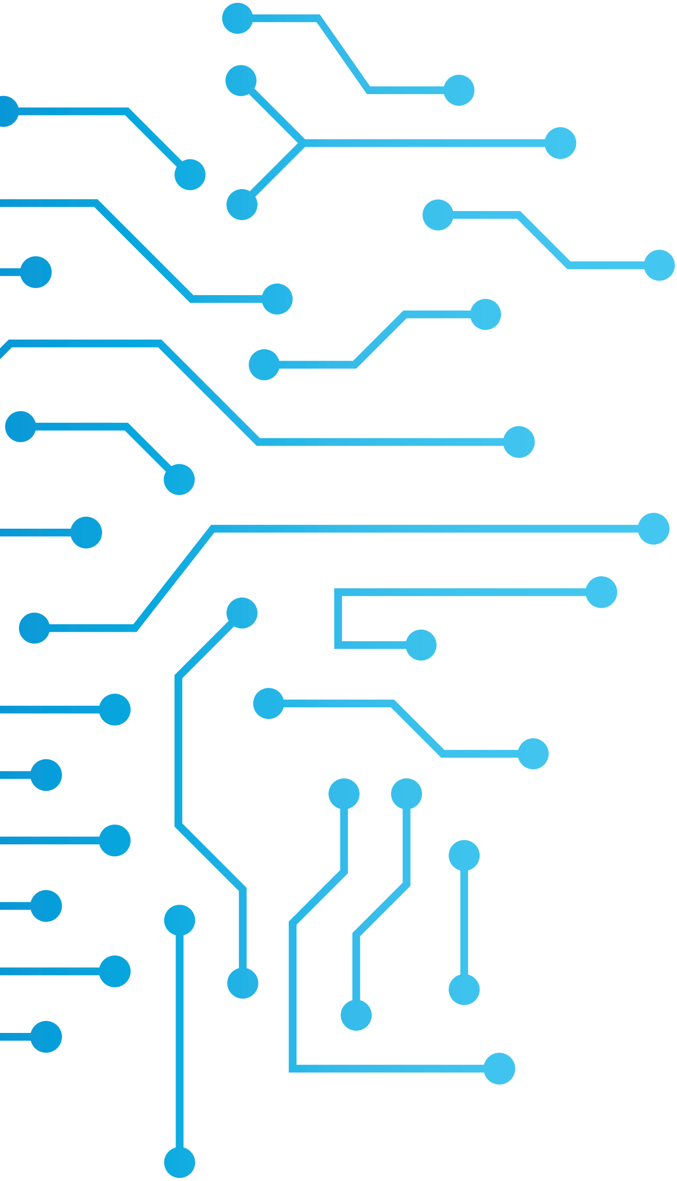
This Paper observes that these problems became more pronounced in the first months of 2020, when the COVID-19 pandemic incentivized States and the private sector to use AI even more, as part of measures introduced in response to the pandemic. A tendency to revert to technocratic solutions, including AI-powered tools, without adequate societal debate or democratic scrutiny was witnessed.

Using four specific case studies (“security threats”; “hate speech”; media pluralism and diversity online; and the impact of AI-powered State surveillance on freedom of expression), this Paper shows how these problems manifest themselves.

This Paper concludes that there is a need to further raise awareness, and improve understanding, of the impact of AI related to decision-making policies and practices on freedom of expression, next to having a more systematic overview of regional approaches and methodologies in the OSCE region. It provides a number of preliminary recommendations to OSCE participating States and internet intermediaries, to help ensure that freedom of expression and information are better protected when AI is deployed.



Introduction



Introduction

The development and use of artificial intelligence (AI) – a broad concept used in policy discussions to refer to many different types of technology – has rapidly expanded over the last two decades. The primary cause of this development has been the mainstream adaptation of widely accessible and affordable technology. The availability of large amounts of data, collected mostly by private actors based on their data-driven business models, has been a driving factor.[3] Consequently, AI has become a part of many aspects of people's daily lives – ranging from commerce, traffic management, policing and law enforcement, health diagnostics and health care, to public services and governance.

The use of AI has further increased with the exponential growth of the sharing and re-sharing of content generated by internet users.[4] Internet intermediaries, in particular social media platforms and search engines, now typically deploy AI systems to manage information flows and to shape and arbitrate content online.[5] AI is used both in content curation (supporting the distribution of content to audiences, such as content ranking or editorial data analysis), as well as in content removal (filtering and taking down illegal or otherwise problematic content). The use of AI-powered technologies by internet intermediaries has also been fostered by increased pressure from

[3] The problems of data collection and the underlying business model of internet intermediaries are discussed in greater detail in the subsequent sections.

[4] According to available data, every single hour, more than 500hours of videos are uploaded onto YouTube and 14.58 million photos on Facebook. For more information, see, e.g., Omnicore statistics.

[5] This Paper uses the term "AI" as an umbrella term that encompasses various concepts. It also acknowledges that the largest internet intermediaries (in particular social media companies Facebook, YouTube and Twitter) do routinely use machine-learning classifiers and algorithmic filtering techniques to detect problematic content (i.e., "hate speech" or spam), and co-ordinated inauthentic behaviour, or ranking and recommendation algorithms to promote and target content; while these systems are not necessarily "AI" in stricto sensu. Many smaller intermediaries, due to their resources capabilities, use much simpler algorithms to organize content.

States on them to remove certain contentwithin very short and strict time periods.[6]

While – if implemented responsibly – AI can benefit societies and provide some positive changes,[7] there is a genuine risk that underlying political, commercial or other interests could have a deteriorating effect on human rights.[8] A number of reports and studies, which examine the impact of AI on human rights – in particular of those groups in society that are in danger of discrimination – have already documented these risks.[9]

As AI greatly influences the way people seek, receive, impart and access information in the digital ecosystem, the ever increasing, pervasive

[6] See, e.g., the Network Enforcement Act (Netzwerkdurchsetzungsgesetz) of Germany, adopted on 17 June 2017; Directive on Copyright and related rights in the Digital Single Market, (EU) 2019/790, European Parliament, 17 April 2019; the EU Code of conduct on countering illegal hate speech online, European Commission, Twitter, Facebook, Microsoft and YouTube, 30 June 2016.

[7] For example, in the field of the media, it has been recognized that the automation of some tasks (such as speech-to-text transcription) can lead to better productivity, improved ability to predict demand to adjust resources, or better access to relevant data; see, e.g., D. James, How artificial intelligence is transforming the media industry, The Record, 7 September 2018; or R. Shields, What the media industry really thinks about the impact of AI, The Drum, 6 July 2018.

[8] See, e.g., R. F. Jørgensen, Human rights in the age of platforms, MIT Press, 2019.

[9] See, e.g., Ranking Digital Rights, Human rights risk scenarios: Algorithms, machine learning and automated decision-making (Consultation Draft), 2020; the Committee of Experts on Internet Intermediaries (MSI-NET), Algorithms and Human Rights, Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications, DGI(2017)12, March 2018; the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) of the Council of Europe, Responsibility of AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, DGI(2019)05, September 2019; Privacy International and ARTICLE 19, Privacy and Freedom of Expression In the Age of Artificial Intelligence, April 2018; AccessNow, 26 recommendations on content governance: a guide for lawmakers, regulators, and company policy makers

and often invisible use of AI by both public authorities[10] and private companies, coupled with their ability to identify and track people, can have a chilling effect on the right to freedom of expression and information. It can lead to self-censorship and altered behaviour, both in online and offline spaces, especially of dissenting voices. This, in turn, may lead to less plurality and diversity of speech, which would ultimately impede on the free flow of information and democratic discourse.

The impact of AI on freedom of expression has been recognized by the international community, and a number of international and regional bodies have called for respect of human rights in the context of AI.[11] However, there is still a need for all stakeholders to understand better what the specific implications of AI are for freedom of expression and freedom of the media, and how the existing freedom of expression framework applies to instances where AI is used.

This need for a better understanding became even more clear in the first months of 2020, when the global outbreak of COVID-19 led many States to introduce emergency powers. During this period, there was an exponential increase in the use of AI-powered surveillance by States, and an increase in reliance on AI in online content moderation by internet intermediaries.[12] Looking ahead, there is a strong tendency to implement AI across the board more often, making its potential, for good and for bad, even more pronounced.

[10] See, e.g., M. Kuziemski & G. Misuraca, AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, Telecommunications Policy, Volume 44, Issue 6, 2020. It concludes that "power has proven to be a central consideration for the use cases of AI in the public sector – by embracing automated methods, one gains control over the physical space, vital resources and information," p.10.
[11] See Chapter V. Applicable International and Regional Standards and VI. Freedom of Expression and AI: Overall Problems.
[12] See, e.g., ARTICLE 19, Coronavirus must not be used as an excuse to entrench surveillance, 20 March 2020; Privacy International, Telco data and Covid-19: A primer, 21 April 2020; or EDRI, COVID-Tech: Surveillance is a pre-existing condition, 27 May 2020.

It is crucial that all State and non-State actors critically evaluate the impact of AI on freedom of expression and freedom of the media. As an international security organization with a strong human rights component, the Organization for Security and Co-operation in Europe (OSCE), and especially the Office of the Representative on Freedom of the Media (RFoM), is well placed to help ensure consistency with global and regional human rights standards in this area. Building on the initial work of the RFoM in this area and initial discussions in 2019 and early 2020,[13] this Paper seeks to provide guidance to participating States in this process. Next to OSCE commitments, the Paper refers to relevant recommendations, developed within other regional bodies, in particular the EU and the Council of Europe.

The structure of this Paper is as follows:

- It first examines some **key contextual issues** as well as the **societal and legislative landscape** that must be considered when developing free speech-compliant actions and policies for the development and use of AI.

- Second, it provides a brief overview of the **applicable standards** for the protection of freedom of expression in the context of the deployment and use of AI at international and regional levels. This is followed by an examination of the main challenges that AI poses to freedom of expression and freedom of the media overall – in particular the potential lack of transparency, accountability and respect for the rule of law in AI processes.

- Subsequently, the Paper offers **case studies** on how AI impacts the right to freedom of expression in specific areas of concern:

[13] OSCE RFoM, Non-paper on the impact of artificial intelligence on freedom of expression, 4 March 2020.

  - › First, it explores the challenges to freedom of expression stemming from AI-powered moderation of certain content (the content that presents "**security threats**" and "**hate speech**"), the associated "**surveillance capitalism**", and the challenges this presents to media pluralism and diversity online.

  - › Second, it outlines the impact of AI-powered **State surveillance** on the right to freedom of expression, the freedom of the media, and the ability of journalists to carry out their work.

- Finally, the last section provides **preliminary recommendations** to OSCE participating States that can be translated into tangible freedom of expression commitments, as well as to internet intermediaries.

This Paper recognizes that the implications of AI on human rights, and especially on freedom of expression, are far broader and go beyond the specific issues examined in the following chapters.[14] The case studies in this Paper were chosen as a reflection of the four main areas of concern identified in the project.

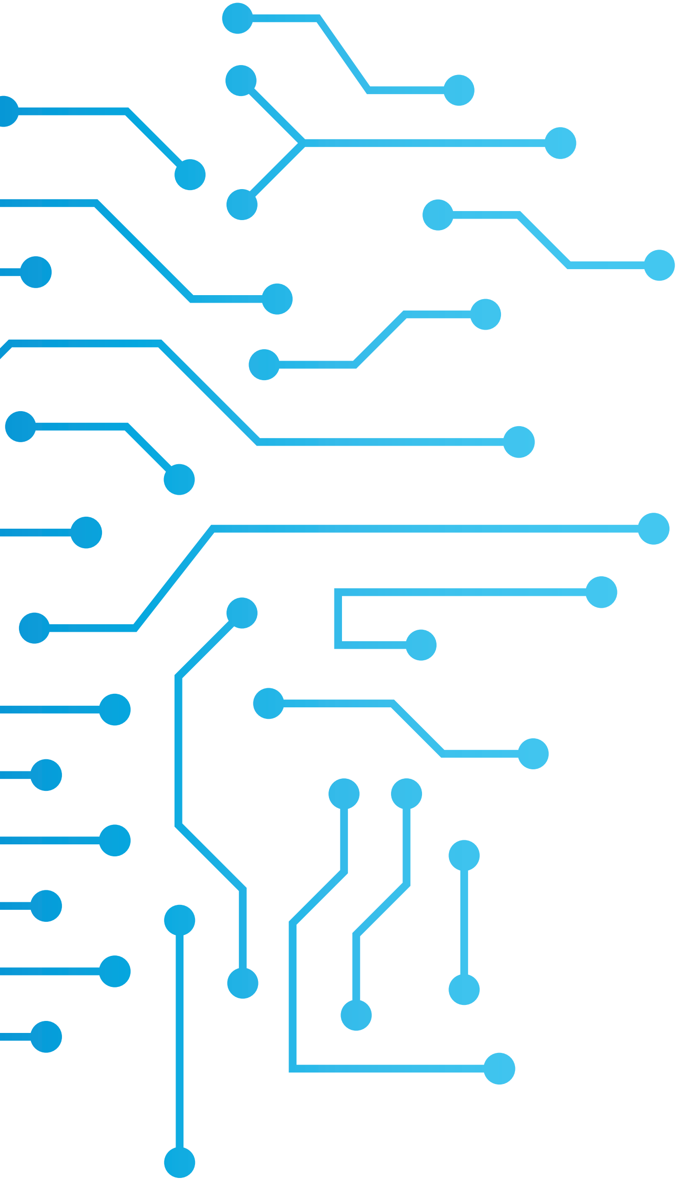
It should be noted that various terms used in this Paper, such as "artificial intelligence", "content moderation", "internet intermediaries" or "hate speech", are very broad concepts that are not uniformly defined in the international human rights framework.

[14] This Paper does not address issues of mis-information/disinformation or information related to public health, measures against inauthentic behaviour, commercial spam, bots or impersonation.

This Paper employs these terms in the way they are most often characterized in expert literature and in documents of other international and regional bodies.[15]

[15] For example, the term "artificial intelligence" can encompass different concepts, in particular, "algorithm" (a computer code that carries out some set of instructions, and is essential to the way computers process data); "encoded procedures for transforming input data into desired output, based on specific calculations"; "automatic decision-making execution", "artificial narrow intelligence" or "artificial general intelligence". It is theorized that the creation of "artificial general intelligence" could lead to the "singularity", or a period of runaway technological growth that profoundly changes human civilization. The term AI is used in this Paper to collectively refer to various types of these concepts. C.f. Privacy International and ARTICLE 19, op.cit. "AI in content moderation" refers to the use of a variety of automated processes at different phases of content moderation; these include content removals, prioritization, de-prioritization, promotion and demotion, monetization, and demonetization of online content. See, e.g., E. Llanso, J. van Hoboken, P. Leerssen & J. Harambam, Artificial Intelligence, Content Moderation, and Freedom of Expression, February 2020. This Paper focuses mainly on content removal and content curation as the most visible techniques of content moderation. The term "internet intermediaries" is used to refer to various actors in the digital ecosystem, including internet service providers, search engines, web hosting providers or "hosts" and social media platforms. The term "content moderation" is used here to refer to sets of measures and tools used by internet intermediaries to enforce their community standards and curate content on their service.

Background



Background

Artificial Intelligence

There is no universally agreed definition of AI. The term is typically used to refer to systems designed by humans operating with varying levels of autonomy, which, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing offline or virtual environments.[16] This broad usage encompasses machine learning, domain-specific AI algorithms, fully autonomous and connected objects, and even the futuristic idea of an AI "singularity".[17]

AI systems act by perceiving their environment through data acquisition; interpreting the collected data; reasoning on the basis of this knowledge, or processing the information derived from this data; and deciding on the best course of action(s) to achieve a given goal. Self-learning forms of AI can adapt their actions by analysing how the environment was affected by their previous actions.[18] In other words, AI is an umbrella term to describe processes that essentially delegate decision-making and execution activities, partially or completely, from humans to software systems.

AI is based on algorithms, which are sets of human-designed instructions with encoded procedures for transforming input data into a desired output, based on specific calculations.[19] 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

[16] The OECD Principles on Artificial Intelligence, May 2019.
[17] See footnote 15 for explanation of AI "singularity".
[18] European Commission's High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019.
[19] See, e.g., T. Gillespie, The relevance of algorithms, Media technologies: Essays on communication, materiality, and society, MIT Press, 2014, p. 167; algorithm is defined as "encoded procedures for transforming input data into the desired output, based on specific calculations."

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 in machine learning is access to big data; extremely large datasets characterized by the volume (amount), velocity (speed) and variety of data.[20] 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 on which to base more precise decisions in the future.

The Role of Internet Intermediaries as Gatekeepers of Freedom of Expression

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 internet intermediaries. Internet intermediaries play an essential role in enabling the flow of information between the two other categories of actors, without any, or little, contribution to the content itself.[21] They enable and manage interactions online, host content online, enable access to platforms, or carry out multiple of these roles.[22]

Internet intermediaries, especially social media platforms, act as "information gatekeepers" by engaging in the selection of information that is published, in the ranking and editorial control over it, as well as

[20] H. Surden, Machine Learning and Law, 89 Wash. L. Rev. 87 (2014).
[21] C.f. European Commission, A Digital Single Market Strategy for Europe - Analysis and Evidence, Brussels, SWD (2015) 100 final, 6 May 2015; or D. Trottier & C. Fuchs, Theorising social media, politics and state, January 2017.
[22] Council of Europe, Role and responsibilities of internet intermediaries.

in the removal of content.[23] They are in a unique position to prevent or mitigate risks that may be inflicted by users' illegal activity.[24] As such, they may, under certain circumstances, be liable for the content of others, and are inevitably put under pressure by public authorities, including law enforcement, to control the content.[25] As a result, they manage processes that could have a great impact on the freedom of expression, other human rights, and democracy at large.

Context of the AI Processes

When considering the impact of AI on freedom of expression in the aforementioned internet ecosystem, there are several underlying issues that need to be taken into account.

The first issue concerns the business model of most internet intermediaries, which is based on a number of key factors. These factors are not unique to AI. However, the specific deployment and use of AI amplify the human rights concerns about the **business model**, in particular:

- The collection and processing of vast amounts of data of users: this happens across the board, even when individuals undertake precautions to protect their privacy and shield their data. Personal digital footprints, even if small, are sufficient for various AI-powered

23 See, e.g., E. B. Laidlaw, A framework for identifying Internet information gatekeepers, International Review of Law, Computers & Technology, 2010, p.16; which stated "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)".

24 A. Savin, EU Internet Law (second edition), Elgar European Law series, Edward Elgar Publishing, 2017, p.143.

25 Ibid. It is also noted that this risk of pressure to act as gatekeepers by government actors is precisely why most democratic countries have legal frameworks that specifically limit the liability that internet intermediaries face for third-party content; c.f. for example, the Center for Internet and Society, Stanford Law School, World Intermediary Liability Map (WILMap).

online services to classify users in pre-developed profiles and to predict customer needs based on the data of other, supposedly similar people. Very often, individual users are not informed that these processes are taking place and that their data is used for these processes, are not aware of how these processes work and which potentially discriminatory aspects are involved, and do not have any user agency or choice with regards to the processes' application.

- Users' attention and engagement are treated as an economic resource: the time users spend on online platforms is one of the key factors that determines the platforms' economic gain. Therefore, most online platforms curate their news feeds and search results in order to increase engagement and time spent on the platform, using AI solutions that determine trending topics and recommended content. These solutions are not neutral, but reflect corporate and profit-oriented values,[26] aiming to increase profit by amplifying sensational or potentially harmful content, so-called "clickbait" content.[27]

- The monetization of users' data through online (targeted) advertising: information and its curation via applications and social media platforms are oftentimes offered to users "for free" in exchange for their behavioural data and other data externalities. Collecting users' data has become more lucrative than collecting users' money, with large amounts of both personal and non-personal data enabling data mining and providing a competitive advantage. This issue is even more poignant in light of the dominant position that certain platforms possess.

[26] H. Bloch-Wehba, Automation in moderation, Cornell International Law Journal, 17 January, 2020, p.6.

[27] See case studies below for how this impacts media pluralism and diversity.

At the same time, it should be noted that all user-generated content sites offer an overwhelming amount of potential content for any one user to view and access, which is why internet intermediaries determine ways to help users narrow down the specific content they see. Internet intermediaries' decisions about which content to amplify on their platforms is also determined by the relevance of the content and users' interest in using the service. AI-powered tools can provide a technically appealing solution in this context. However, the use of AI, even if not paired with the above-mentioned business model, can seriously impede on freedom of expression.

The second key issue concerns the **dominance of a few internet intermediaries**[28] in the digital markets, which makes it possible for these companies to lock in users and to compel them to follow arbitrary rules that only the intermediaries control ("network effect").[29] Therefore, a small number of dominant companies decide what information the majority of internet users get to see and share. This small group is able to do so on the basis of AI-powered technology, the aforementioned massive troves of data, and network effects. In this sense, platforms are not only gatekeepers of the information itself, but also of innovation and the industry's new approaches.

The third underlying issue is the **resulting surveillance** of individuals' activities through AI technologies. On the State level, AI enables both

[28] This Paper notes that intermediaries vary across the region – with some companies operating on the global level, while some are specific to certain markets. The critique of dominance concerns primarily companies operating on the global level.

[29] See, e.g., EDRI, Platform Regulation Done Right: EDRi Position Paper on the EU Digital Services Act, 9 April 2020. 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.

mass (untargeted) and targeted surveillance of individuals.[30] AI-powered mass surveillance conducted by States mostly relies on data collected through, and shared by, private companies. Corporate surveillance (often referred to as "surveillance capitalism") refers to the aforementioned business model of internet intermediaries.

AI and Content Moderation

Today, internet intermediaries are often called upon by States to play a more active role in monitoring the content on their platforms and to make decisions on the content's permissibility.[31] Such moderation typically takes place on three different levels.

- In many instances, various types of 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, AI is supposed to automatically block such content from being published.

[30] Mass surveillance refers typically to automated gathering and bulk interception, storage and analysis of internet communications not necessarily linked to particular individuals; while targeted surveillance is only aimed at a specific person or entity. Where accompanied by appropriate legal and procedural safeguards, targeted interception of an individual's communications is a legitimate act of a democratic government. On the other hand, even when a legitimate purpose for the use of AI-powered surveillance is identified, its deployment has to meet a narrowly constructed test of legality, necessity and proportionality: the technology has to be absolutely necessary to achieve the scope and there should be no other less invasive means to do so. If this test is not passed, the use of the technology shall not be allowed, irrespective of its availability. See, e.g., the European Court of Human Rights, Klass and Others vs. Germany, App. No. 5029/71, 6 September 1978; Liberty & Other Organisations v. the UK, App. No. 58243/00, 1 July 2008; or Roman Zakharov v. Russia, App. No. 47143/06, 4 December 2015.

[31] At present, there is a number of legislative or regulatory proposals considering requiring monitoring/filtering of content that are a significant threat to freedom of expression. See, e.g., in the EU, the Copyright in the Digital Single Market Directive, the Terrorist Directive and pending Terrorism Regulation.

- Due to vast amounts of user-generated content, the 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 and 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 content, or previous interaction by a "similar user".[32] The criteria used in the algorithmic decision-making are usually not made public, meaning that "black boxes"[33] employing AI technologies determine which content is available to whom. These processes are not value-neutral and are oftentimes driven by private profit rather than being public value-oriented.[34]

- On the third level, and mostly with human intervention, content moderation is based on reporting mechanisms, established under a legislative framework or the internal policies of companies. These mechanisms usually allow any user to report "inappropriate" content (based on the platform's internal rules), which triggers a review 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.

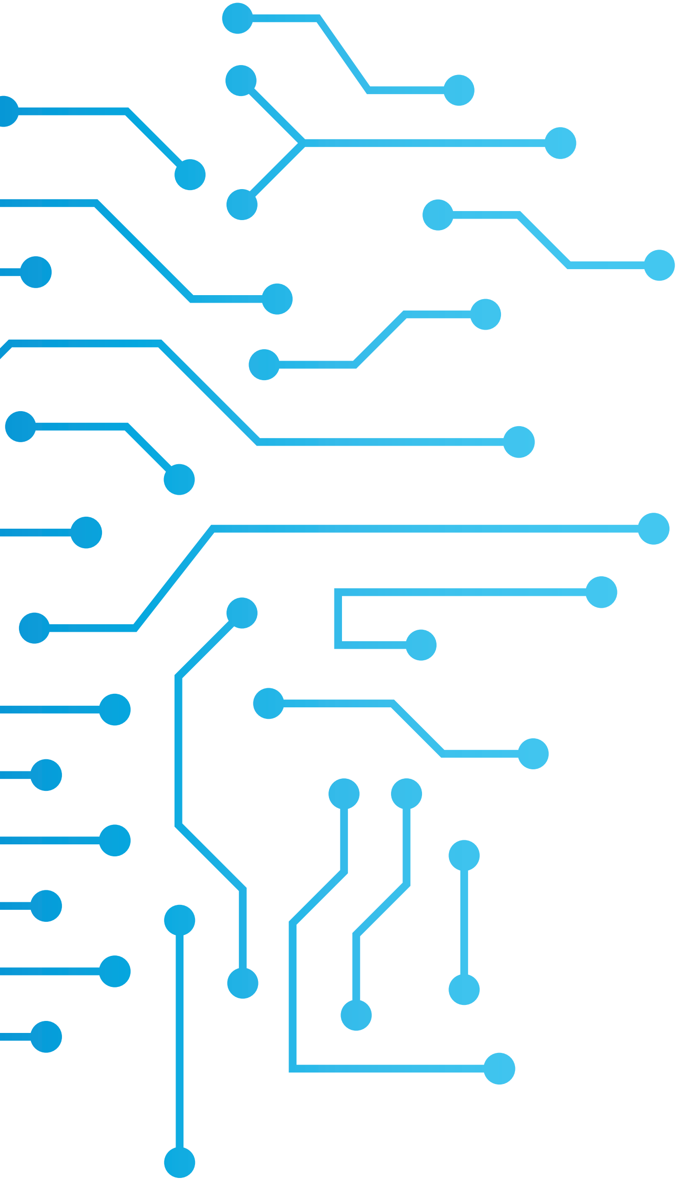
[32] When deciding which content to show to individual users, the following factors are important (not exclusive): character of the person who wants to distribute the content (user, page, group, business, etc.); form of content (text, video, audio, photo, etc.); interest in content from other network 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.

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

[34] Value-neutrality is a well-known discourse in the philosophy of science. It is also suggested that digital technology is not value neutral but an "embodiment of a complex system of political, social, economic, and technical priorities and philosophical stances;" see W.D. Surry and F.W. Baker III, The Co-dependent Relationship of Technology and Communities. British Journal of Educational Technology. 2016, 47(1):13-28.

The aforementioned issues with the usage of AI can be demonstrated in particular in key areas of concern, and will be addressed in this Paper. These include content removal of speech presenting "security threats" and "hate speech", the ramification on media diversity, and the impact of surveillance.

Applicable International and Regional Standards



Applicable International and Regional Standards

Standards on AI and Human Rights

A number of international and regional human rights bodies have recognized the impact of AI and automation on human rights. For example, in its 2017 Resolution, the United Nations Human Rights Council (UNHRC) noted with concern that "automatic processing of personal data for individual profiling may lead to discrimination or decisions that otherwise have the potential to affect the enjoyment of human rights, including economic, social and cultural rights."[35] Similar concerns were raised by both the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (UN Special Rapporteur on Freedom of Expression) and the UN Special Rapporteur on Freedom of Association and Assembly. They also expressed concerns over how untargeted AI-powered surveillance impacts those exercising their rights in both physical and digital spaces,[36] in particular journalists, human rights defenders, or even UN investigators.[37]

On the regional level, the bodies of the Council of Europe and the European Union have adopted several documents and recommendations on AI and human rights.[38]

[35] UN Human Rights Council Resolution on the Right to Privacy in the Digital Age, U.N. Doc. A/HRC/34/L,7, 23 Mar. 2017, para. 2.

[36] Report of the Special Rapporteur on the rights to freedom of peaceful assembly and of association (2019).

[37] OHCHR, UN expert calls for immediate moratorium on the sale, transfer and use of surveillance tools (2019).

[38] See, e.g., Recommendation n°2102(2017) about Technological convergence, artificial intelligence and human rights, (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); Declaration Decl(13/02/2019)1 on the manipulative capabilities of algorithmic processes, adopted by the Committee of Ministers on 13 February 2019; Recommendation on Artificial Intelligence and Human Rights "Unboxing artificial intelligence: 10 steps to protect Human Rights,(2019); 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; or The European Commission, White Paper on Artificial Intelligence: a European approach to excellence and trust, (2020).

In particular, they raised concerns about the challenges posed by AI to democracy and human rights, including the right to privacy. They also provided some guidance on measures that national authorities could take to prevent or mitigate the negative impact of AI on people's lives and rights, as well as guidelines on issues such as transparency, accountability and independent oversight over AI systems.

Aside from these documents, international and regional bodies have created dedicated committees or task force groups to develop recommendations in the area of AI, including on human rights and "trustworthy AI".[39] For example, the European Commission's High Level Expert Group on AI adopted Ethics Guidelines for Trustworthy AI that focuses on respect for human autonomy, prevention of harm, fairness, and explicability of AI.[40] In addition, several countries have published generic national AI strategies.

Standards on AI and Freedom of Expression

Although the aforementioned documents raise concerns on the impact of AI on human rights, including the rights to freedom of expression, privacy and the right to equality and non-discrimination, and there are various international initiatives around AI and human rights, there is only a limited number of international standards that explicitly deal with AI and the right to freedom of expression and freedom of the media.

The only dedicated documents from intergovernmental organizations on AI and freedom of expression so far are the report by the UN Special Rapporteur

[39] See, e.g., the Sub-Committee on artificial intelligence and human rights of the Committee on Legal Affairs and Human Rights of the Parliamentary Assembly of the Council of Europe; or the EU High-Level Expert Group on Artificial Intelligence.
[40] Ethics Guidelines for Trustworthy Artificial Intelligence, 8 April 2019.

on Freedom of Expression to the UN General Assembly in August 2018,[41] and the report by the Council of Europe on AI and the media.[42]

The UN Special Rapporteur on Freedom of Expression recommends that States create a policy and legislative environment conducive to a diverse, pluralistic information environment in the AI domain. Such measures could include the regulation of technology markets to prevent the concentration of AI expertise and power in the hands of a few dominant companies. Such measures could also include regulation to increase interoperability of services and technologies. For the private sector, the UN Special Rapporteur on Freedom of Expression recommends they should, inter alia, make explicit where and how AI technologies are used on their platforms, services and applications, as well as publish data on content removals and case studies. Moreover, they are recommended to provide education on commercial and political profiling; and to give individual users access to remedies for adverse human rights impacts of AI systems.

The report by the Council of Europe analyses the use of AI-powered tools in light of Article 10 of the European Convention on Human Rights. It highlights the role of member States in ensuring that access to innovative technologies, training data, digital skills, and media literacy education is also open to smaller intermediaries, not just the most dominant ones. It also stresses the need to produce concrete proposals for the development of professional ethics on AI-powered tools and their compatibility with human rights and fundamental freedoms. It recommends that States

[41] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on Artificial Intelligence technologies and implications for the information environment, A/73/348, 29 August 2018.
[42] Council of Europe, Artificial Intelligence – Intelligent Politics Challenges and opportunities for media and democracy, Background Paper, Ministerial Conference, Cyprus, 28-29 May 2020.

identify clear conditions for the responsibility and (editorial) oversight of automated processes in the media, and study new digital inequalities and unequal opportunities of access to information.

Additionally, a number of other standards are relevant to the deployment and usage of AI, in particular those related to the role and responsibility of internet intermediaries,[43] a right to anonymity,[44] and privacy and data protection.[45]

AI and the Responsibilities of the Private Sector

International human rights law acknowledges that States are the prime duty bearers in respect to human rights, putting on them the obligation to protect, promote and respect human rights. A number of documents on the international and regional levels, however, also recognize that the private sector bears a responsibility to respect human rights. These documents are applicable to the deployment and use of AI in their practices.

[43] For instance, States should not impose a general obligation on intermediaries to monitor the information that they transmit, store, automate or otherwise use, see Directive 2000/31/EC June 8, 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive). See also Report of the Special Rapporteur on Freedom of Expression to the Human Rights Council on Freedom of expression, states and the private sector in the digital age, A/HRC/32/38, 11 May 2011, para. 47; Regulation (EU) 2016/679, 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1; Recommendation CM/Rec (2008)6 of the Committee of Ministers to Member states on measures to promote and respect for freedom of expression and information with regard to internet filters, s.1; 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.

[44] Report of the Special Rapporteur on freedom of expression to the Human Rights Council on the use of encryption and anonymity to exercise the rights to freedom of opinion and expression in the digital age, A/HRC/29/32, 22 May 2015.

[45] See, e.g., Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, GDPR).

The UN Guiding Principles on Business and Human Rights offer a non-binding vehicle for applying human rights standards to corporations, including internet intermediaries.[46] The Guiding Principles state that businesses should respect human rights and enumerate further duties for companies. Among these are duties to apply internationally recognized human rights standards; mitigate adverse human rights impacts; develop policies that promote human rights; carry out due diligence to identify human rights risks; and provide remedies for human rights violations.

The OSCE RFoM and the UN Special Rapporteur on Freedom of Expression have also addressed the role of social media platforms in promoting freedom of expression, and have recommended that said platforms respect and promote human rights in their practices.[47] The UN Special Rapporteur on Freedom for Expression, for example, recommended that intermediaries should establish clear and unambiguous terms of service, in line with international human rights norms and principles; produce transparency reports; and provide effective remedies for affected users in cases of violations.[48] The UN Special Rapporteur on Freedom of Expression also emphasized that companies should embark on radically different approaches to ensure transparency at all stages of their operations, and to open themselves up to public accountability.[49] As companies are typically ill equipped to make determinations of the (il)legality of content,

[46] Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect and Remedy' Framework, developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, 7 April 2008, A/HRC/8/5A/HRC/17/31. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

[47] See, e.g., Key conclusions and recommendations, Conference on internet freedom, The role and responsibilities of internet intermediaries, co-organised by the Austrian Chairmanship of the OSCE and the Czech Chairmanship of the Council of Europe Committee of Ministers 13 October, 2017.

[48] Report of the Special Rapporteur on FoE, A/HRC/32/38, 11 May 2016.

[49] Ibid.

the OSCE RFoM and the UN Special Rapporteur also called on States not to require from the private sector to take steps that unnecessarily or disproportionately interfere with freedom of expression, whether through laws, policies, or extra-legal means.[50] Similar recommendations have been made by of the Council of Europe[51] and the European Union Agency for Fundamental Rights (FRA).[52]

AI and Ethics

Additionally, ethical codes on AI, some of which are general and others sectoral, have been developed by private sector initiatives, technical standard-setting bodies, and States. The tech industry has also undertaken "self-regulatory" initiatives regarding AI. Overall, these initiatives focus on eight key themes: privacy; accountability; safety and security; transparency and explainability; fairness and non-discrimination; human control of technology; professional responsibility; and promotion of human values.[53] To name a few:

[50] Ibid. See also, Joint Declaration on Challenges to Freedom of Expression in the Next Decade, Declaration by the United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe 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 Special Rapporteur on Freedom of Expression and Access to Information, 10 July 2019; or OSCE and the Council of Europe, Key Conclusions and Recommendations of the Internet Freedom Conference, 2017.

[51] See, e.g., the Committee of Ministers of Council of Europe, Recommendation CM/Rec (2012)4 of the Committee of Ministers to Member States on the protection of human rights with regard to social networking services, adopted by the Committee of Ministers on 4 April 2012. These recommendations were further echoed in the Committee of Ministers, Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, p. 4; Council of Europe Recommendation CM/Rec (2018)2, op.cit., or The Council of Europe Commissioner for Human Rights, The Rule of law on the Internet and in the wider digital world, Issue paper, CommDH/IssuePaper (2014) 1, 8 December 2014.

[52] Fundamental Rights Agency (FRA) Proposal for a Regulation on preventing the dissemination of terrorist content online and its fundamental rights implications (2019).

[53] See Berkman Klein Center, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI, 2020.

- The Global Initiative on Ethics of Autonomous and Intelligent Systems of the Institute for Electrical and Electronics Engineers (IEEE)[54] focuses on developing technical standards that embed ethics in AI systems. The initiative also aims to raise awareness among the AI community about the importance of prioritizing ethical considerations in the development of technology;

- The Partnership on Artificial Intelligence to Benefit People and Society – originally established by Microsoft, Google, Amazon, Facebook, and IBM – aims to "study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society."[55]

Some companies have also adopted internal policies on AI. Most notably, in June 2018, Google published its Principles on Responsible Development of AI,[56] which contain an overview of the company's approach to the issue. Some companies also established internal review processes to help avoid bias, to test rigorously for safety, and to design AI with privacy as a key consideration.

At the same time, there is growing criticism about the lack of clear commitment to human rights in these initiatives (even so-called "ethical white-washing[57]). In particular, there are calls that human rights should

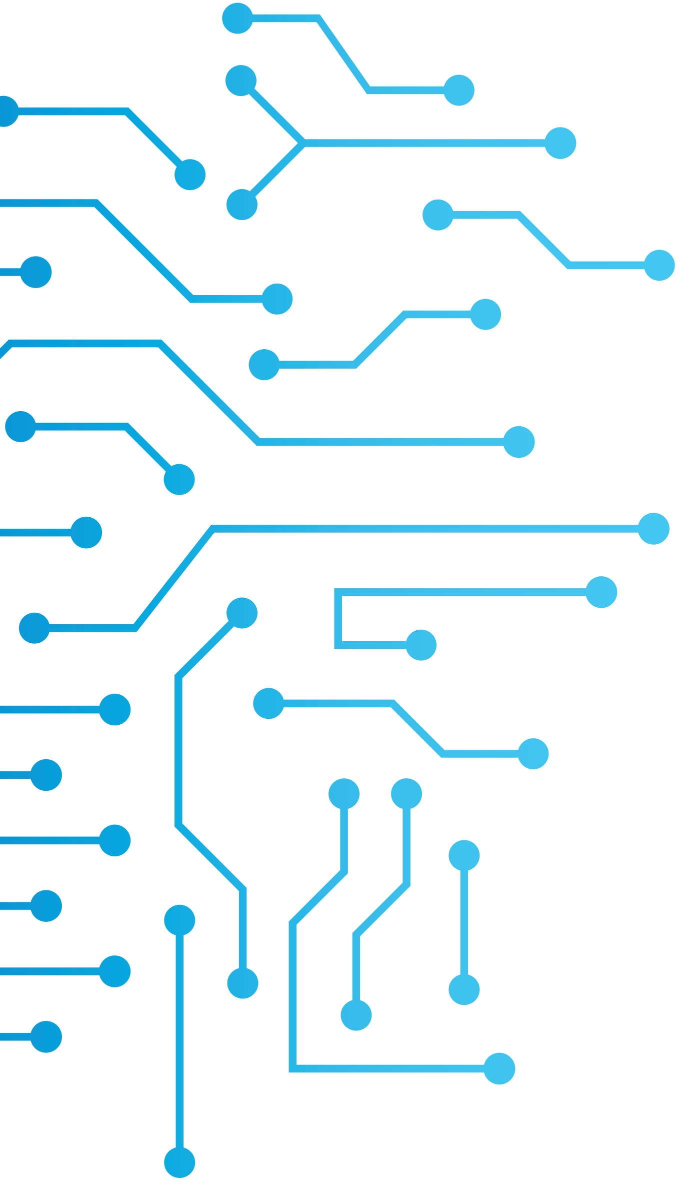
[54] For more information, see IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems or IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being, 2019.
[55] See Partnership for Artificial Intelligence website.
[56] Google, Responsible Development of AI.
[57] For example, Prof. Metzinger called the EU ethics guidelines for AI as "ethical white-washing", arguing "the underlying guiding idea of a 'trustworthy AI'" he argues, "is conceptual nonsense. Machines are not trustworthy; only humans can be trustworthy (or untrustworthy). If, in the future, an untrustworthy corporation or government behaves unethically and possesses good, robust AI technology, this will enable more effective unethical behaviour;" see T. Metzinger, Ethics washing made in Europe, Der Tagesspiegel, 8 April 2019.

form the basis of any ethical standards and protocols, or that companies should adopt clear human rights-centred AI principles as policies.[58]

In the subsequent section, this Paper outlines how the existing human rights framework applies, or falls short, in four specific areas of concern (see below under "case studies").

[58] See, e.g., Ranking Digital Rights, 2020, op.cit.; ARTICLE 19, Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence, April 2019.

Freedom of Expression and AI: Overall Problems



Freedom of Expression and AI: Overall Problems

AI poses several overarching problems to freedom of expression, and human rights in general, common to all areas of concern explored in the four case studies. Most of these problems are not specific to AI. However, the use of AI does amplify the concerns over the challenges they present to human rights.

- At present, society in general, but also many actors deploying AI, only have a very limited understanding of the legal (and ethical) implications of the development and control of AI, in particular machine learning. **Machine learning** uses statistics to find patterns in the data amassed by internet intermediaries. Studies show that the choice of dataset, and the methods used, can have a discriminatory impact on some groups in society. This creates concerns over the privacy of individuals and their engagement in civic space in general (e.g., through the identification of individuals in public spaces or protests), which also impacts freedom of expression. Further, machine-learning methods are used for advanced profiling of individuals, based on their engagement through technologies (e.g., sensitive personal data, such as sexual orientation or political beliefs).[59] These profiles are used to categorize, sort, assess or rank individuals, oftentimes without their knowledge. This creates concerns both for privacy and individual autonomy, as well as for freedom of expression.

- The aforementioned **business model**, combined with the fact that **a few internet intermediaries dominate** the digital markets, is inherently problematic from a human rights perspective, and particularly freedom of expression,[60] leaving AI as a powerful instrument in the

[59] See, e.g., Privacy International, op.cit.
[60] See, e.g., Amnesty International, Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights, November 2019; or . Privacy International, Response to the Open Consultation on the Online Harms White Paper, July 2019.

hands of a small number of private companies. For these reasons, several civil society organizations have called for a radical overthrow of this business model as a prerequisite for human rights protection.[61]

- **Lack of respect for freedom of expression:** AI is deployed by companies to identify and remove specific content – be it illegal content or content that violates internal rules and procedures (i.e., Community Standards and Terms of Services) – from their platforms.[62] Whether certain content should be considered "illegal" typically depends on the context in which the content is presented.[63] This is a complex task, which is dependent on local context, local languages, and other societal, political, historical and cultural nuances.[64] Numerous studies show that automated decisions for content removal can fail to understand nuances underpinning the pieces of content,[65] resulting in the filtering and taking down of legitimate content.[66] Furthermore, cultural and legal differences across the world put into question the application of systems trained on data

[61] See, e.g., Amnesty International, Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights, November 2019; or Privacy International, Response to the Open Consultation on the Online Harms White Paper, July 2019.

[62] It should be noted that the internal rules and guidelines often go beyond what is "illegal" according to national legislation or international frameworks; see, e.g., Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/HRC/38/35, 6 April 2018.

[63] This is particularly the case for "hate speech" or "extremist" content. Thus, some forms of content might have a potentially harmful effect, but they can still be protected under international human rights standards and removing such content is harmful to freedom of expression.

[64] For example, detecting bullying online requires an understanding of the relationship between two or more users, their age, the number of exchanged messages, the nature of their connection, as well as previous interaction history and shared connections. See, e.g., OFCOM and Cambridge Consultants, Use of AI in Online Content Moderation (2019). See also, N. Duarte, E. Llanso, Mixed Messages? The Limits of Automated Social Media Content Analysis, CDT, 28 November 2017.

[65] See, e.g., Tech Dirt, YouTube Takes Down Ariana Grande's Manchester Benefit Concert On Copyright Grounds, 17 June 2017.

[66] This is the case, for instance, of automated takedowns of political speech and marginalised voices based on copyright upload filters. See, e,g., J. Reda, When filters fail: These cases show we can't trust algorithms to clean up the internet, 2017.

from one region to be able to work effectively in other regions. All of this emphasizes the importance of genuine human involvement, referred to as the "human in the loop" principle, which should guarantee that the accuracy of AI would remain amenable to human intervention.[67]

- **Lack of respect for the rule of law:** The deployment and usage of AI in content moderation often leads to the circumvention of due process and legal safety. This happens in two instances. First, law enforcement and other public authorities often delegate the identification and removal of certain content, including "hate speech" and "security threat" (see below), to private actors, excluding due process. Even though internet intermediaries are usually not obliged, in principle, to comply with such requests without a legitimate decision issued by courts or adjudicatory bodies, this does create pressure on them to remove specifically categorized content, mostly in a certain, often short, time-frame. Given the vast amount of content on some platforms, this necessitates, or at least incentivizes, the usage of AI for detecting such content in time. Second, most intermediaries do not provide any due process guarantees when they remove the content under their own community guidelines.

- **Lack of transparency:** Transparency is essential for freedom of expression, as it enables the scrutiny of users, the media and the general public, including researchers and regulators. Several studies, however, point out that States and corporations are deploying AI in ways that are non-transparent and inscrutable.[68] This lack of transparency spreads over the entire process of AI design: from the initial idea through

[67] See, e.g., G. Wang, Humans in the Loop: The Design of Interactive AI Systems, 20 October 2019.

[68] See, e.g., Ranking Digital Rights, 2020, op.cit.; Panoptykon, Who (really) targets you?; or E.J. Llanso, No amount of "AI" in content moderation will solve filtering's prior-restraint problem, Big Data & Society, 23 April 2020.

to its implementation. There is insufficient accessible information about who is developing which AI systems, what kind of technology is being developed and how, or for which purposes. The tech industry develops, or owns, the majority of AI technology, which complicates the assessment of many of these tools by "trade secrets rules and high barriers to transparency around application and development, as well as the inherent complexity of these systems".[69] Lack of transparency also permeates the partnership between States and the private sector in the deployment of these technologies, typically because of "state-protected secrets". For these reasons, proposals have been made for a multi-tiered approach to transparency, meaning that users would be provided with substantial information to help them understand more fully the operation of the systems they use; government experts would be provided with more detailed terms of reference by platforms to describe the operation of their systems; and researchers would be provided access to data to conduct studies.[70]

- **Lack of accountability:** The ability of AI systems to be invisible and opaque, as well as inscrutable, lead on to efforts to make companies accountable and to attempts to form independent oversight of these processes. There have been several proposals for independent oversight on so-called algorithmic accountability, such as the principle that an algorithmic system should employ a variety of controls, to ensure that the operator can verify that these algorithms work in accordance with its intentions, and to ensure that the operator can identify and rectify harmful outcomes and reproduction inequalities.[71]

[69] C.F. Privacy International and ARTICLE 19 report, op.cit.
[70] M. MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, 12 February 2020.
[71] See, e.g., the proposal for the Algorithmic Accountability Act, 2019 in the USA; or the Government of France, Creating a French framework to make social media platforms more accountable: Acting in France with a European vision, May 2019.

- **Lack of effective remedies:** In cases of violation of their rights, including the right to freedom of expression, international human rights standards provide individuals with the right to an effective remedy. As noted in some studies, "the precondition to the establishment of effective remedy processes is ensuring that individuals know that they have been subject to an algorithmic decision (including one that is suggested by an artificial intelligence system and approved by a human interlocutor) and are equipped with information about the logic behind that decision" ("explainability").[72] A number of civil society organizations have demanded, therefore, that internet intermediaries, in cases of violations caused by the deployment of AI, should guarantee their users with a right to appeal and effective remedy.[73]

These problems became more pronounced in the first months of 2020, as the **COVID-19 pandemic** has incentivized States and the private sector to use AI even more. Digital technologies can be vital for many people during a pandemic, as they enable access to information, communication with others, or access to education.[74] AI and other technologies can play an important supportive role in the efforts to protect public health, such as by helping to spread public health messages, or increasing access to healthcare. At the same time, however, there has been a tendency to revert to technocratic solutions without adequate societal debate or democratic scrutiny.[75]

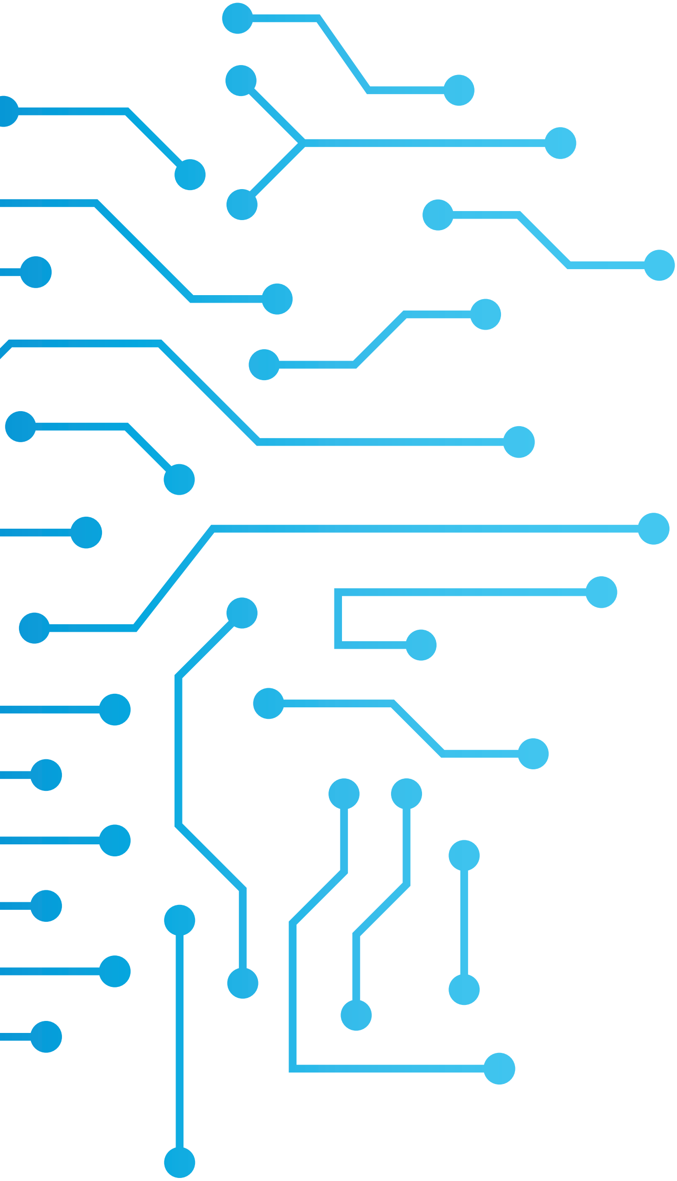
[72] Privacy International, op.cit.
[73] See, e.g., Santa Clara Principles, On Transparency and Accountability in Content Moderation, 2018; AccessNow, 26 Recommendations, op.cit.
[74] See, e.g., ARTICLE 19, Coronavirus: Access to the internet can be a matter of life and death during a pandemic, 31 March 2020.
[75] Ibid

In particular, the first months of 2020 were characterized by attempts to increase the tracking and monitoring of individuals and populations, ostensibly to protect public health.[76] The pandemic disrupted the way in which internet intermediaries, especially social media platforms, undertake content moderation on their platforms, as it has accelerated the use of AI. In March 2020, for instance, all major social media companies (Facebook, Twitter, Google and YouTube) announced that, due to their reduced workforce and staff physically available to review content during curfews and other containment measures, they would rely more on AI for the removal of some content without human review.[77] YouTube publicly stated that the lack of human review would also result in delays in appeals against content removals.[78] During this period, it also became clear that, because of a higher reliance on AI, a large number of content was incorrectly removed.[79] Independent researchers and other stakeholders have raised concerns about the removal of large amount of content by AI during the pandemic, and the inability of independent scrutiny over this practice.[80]

It is important that any AI-supported measure adopted in response to the COVID-19 pandemic fully complies with international human rights standards. It is equally important that such measures or practices are not further entrenched in legislation or in practice.

76 See, e.g., L. Taylor, G. Sharma, A.K. Martin, and S.M. Jameson (eds), Data Justice and COVID-19: Global Perspectives. London, Meatspace Press (forthcoming).
77 See, e.g., Facebook, Keeping People Safe and Informed About the Coronavirus, 4 May 2020; Twitter, An update on our continuity strategy during COVID-19, 16 March 2020; or Youtube, Protecting our extended workforce and the community, 16 March 2020.
78 Youtube, op.cit.
79 See, e.g., tweets of Guy Rosen, Vice President of Integrity at Twitter, The Guardian, Facebook says spam filter mayhem not related to coronavirus, 18 March 2020; or WZB, COVID-19 and the Future of Content Moderation, 15 April 2020..
80 See, e.g., CDT, COVID-19 Content Moderation Research Letter, 22 April 2020.

Case Studies



Case Studies

The Use of AI and "Security Threats"

AI is often deployed to detect content that – under most laws and platform standards – is perceived as a threat to national security.[81] One AI tool deployed in this area is the "Hash Database", which was initially developed in 2016 by Facebook, YouTube, Microsoft, and Twitter.[82] It contains digital hash "fingerprints" of images and videos that platforms have identified as being "extreme" terrorist material, based on various criteria.[83] Platforms use automated filtering tools to identify and remove duplicates of the hashed images or videos. Other 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.

There are several problems with the use of AI in this area:

- First, there is no universally adopted definition of various categories of content removed due to security reasons. Terms such as "extremism", "violent extremism", "terrorism" or "incitement or condoning of terrorism" are not uniformly or comprehensively defined under international human rights law, or are weak terms,[84] despite

[81] See, e.g., Recommendation of 1.3.2018 on measures to effectively tackle illegal content online (C(2018) 1177 final); or Proposal for a Regulation of the European Parliament and the Council on preventing the dissemination of terrorist content online, A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September 2018, COM/2018/640 final.

[82] It has been reported that the Database now contains hashes representing over 200,000 images or videos. At least 13 companies now use the Database, and some 70 companies have reportedly discussed adopting it. See Global Internet Forum to Counter Terrorism, Hash Sharing Consortium.

[83] According to GIFCT transparency report, "for the purposes of the hash sharing database, and to find an agreed upon common ground, founding companies in 2017 decided to define terrorist content based on content relating to organizations on the UN Terrorist Sanctions lists. Companies also agreed upon a basic taxonomy around the type of content ingested relating to these listed organizations. The taxonomy includes the following labels that are applied to the content when a company ads hashes to the shared database;" see GIFCT, Transparency Report - July 2019.

[84] See, e.g., OHCHR Factsheet on Human Rights, Terrorism and Counter-terrorism; Report by the UNSR on the Promotion and Protection of Human Rights and Fundamental Freedoms

being employed routinely based on national laws and community guidelines or terms of services of internet intermediaries. This poses the risk of arbitrary application of these laws and guidelines. The increasing pressure that States put on internet intermediaries to take a more proactive role in policing such vaguely defined "terrorist" or "extremist" content, and to develop proactive automated measures to identify content falling under these categories, further exacerbates this problem.

- Second, as already noted above, AI has shown to be prone to error, particularly when an analysis of the context is necessary. There is always a certain grey area regarding the question of whether content presents a security threat, including context and nuances that call for a sophisticated and balanced assessment. There have been cases of a "false negative", with AI incorrectly identifying illegal content to be permissible, or a "false positive", with the removal of legitimate content.[85] Videos and other content, which may be used to advocate terrorist violence in one context, may be essential for news-reporting elsewhere,[86] for combating terrorist recruitment online, or for research work. Technical filters are blind to these contextual differences. It was reported, for example, that YouTube's AI-powered filtering tools deleted over 100,000 videos of the Syrian Archive, a civil society organization preserving evidence of human rights abuses in Syria.[87]

while Countering Terrorism, A/HRC/43/46, 2020, or OSCE, Preventing Terrorism and Countering Violent Extremism and Radicalization that Lead to Terrorism, 2014.

[85] Ibid., OFCOM and Cambridge Consultants, op.cit.

[86] C.f. CPJ, EU online terrorist content legislation risks undermining press freedom, 11 March 2020.

[87] See, e.g., D. Kayyali and R. Althaibani, Global Witness, Vital Human Rights Evidence in Syria is Disappearing from YouTube.

There are only a few studies[88] on the effectiveness of AI-powered identification of illegal content presenting "security threats". An evidence, or research, based justification for the swift removal of online content is currently missing. There is also a lack of evidence demonstrating that the successful removal of such content in fact results in reducing the security threat. This lack of evidence is problematic, as international freedom of expression standards mandate that restrictions on freedom of expression must be necessary and proportionate to achieve a legitimate aim.

More transparency and research is needed to understand the possible impact of the usage of these tools on content and freedom of expression. It is important to ensure that more extensive usage of AI in this context will not amplify the risk of harm to users whose messages and communications, about matters of urgent public concern, may be wrongly removed by internet intermediaries, including news-reporting or journalistic content.[89]

The Use of AI and "Hate Speech"

Problems with the deployment of AI in the identification and removal of "hate speech" are largely similar to those of speech presenting "security threats".

- There is no uniform definition of "hate speech" under international human rights law. International and regional human rights instruments imply varying standards for defining and limiting "hate speech", with a wide range of hateful expressions requiring different

[88] **88** 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.

[89] **89** See, e.g., Brittan Heller, Combating Terrorist-Related Content Through AI and Information Sharing, 26 April 2019.

responses based on the severity of the speech in question.[90] The context of the speech, the role of the speaker and the likelihood that the speech results in harm are among the key factors to determine whether the speech in question should be restricted.[91] Legal prohibitions and community guidelines of internet intermediaries often fail to reflect these complex aspects of the international human rights framework. This poses problems for the deployment of AI in policing "hate speech" online, 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 are not capable of recognizing figurative speech, or of discerning mockery or offensive language that sometimes follows heated public debate over issues of public importance. This often leads to the removal of legitimate content. Facebook's 2018 report, for instance, agreed that technology still does not work well in detecting contextually complex "hate speech".[92]

- Studies show that bias in AI design can disproportionately affect the removal of minority groups' content.[93] There is an additional risk of unwanted consequences when AI, having been trained in a certain

[90] 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. See, e.g., ARTICLE 19, 'Hate Speech' Explained, 2015; or Susan Benesch, Dangerous Speech Project.
[91] Ibid. see also OHCHR, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.
[92] Facebook, Facebook Publishes Enforcement Numbers for the First Time, 15 May 2018.
[93] See, e.g., M. Sap et al, The Risk of Racial Bias in Hate Speech Detection, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678 Florence, Italy, July 28 - August 2, 2019; S. Ghaffary, The algorithms that detect hate speech online are biased against black people, VOX, 15 August 2019; or Algorithm Watch, Automated moderation tool from Google rates People of Color and gays as "toxic", 19 May 2020.

cultural setting, is being used in societies with different cultural communication rules. A recent study of Twitter content, for instance, 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 that "these systems may discriminate against the groups who are often the targets of the abuse we are trying to detect".[94] Besides the problem of over-removals, AI can also fail to remove genuinely illegal content, leading to serious harm for groups at risk of discrimination.

AI and Media Pluralism and Diversity

Media pluralism is a crucial component of the right to freedom of expression. It is defined as "the diversity of media supply, reflected, for example, in the existence of a plurality of independent and autonomous media [...] as well as a diversity of media types and contents made available to the public".[95] The Council of Europe recommended in its 2018 recommendation on media pluralism that "the automated decision-making processes that govern the distribution of online content should integrate the objective of improving the effective exposure of users to the broadest possible diversity of media content online."[96]

As "information gatekeepers", internet intermediaries are in a position to promote or hinder the public's right to access pluralistic and diverse information.[97] As noted above, the current business model does not

[94] T. Davidson et al, Racial Bias in Hate Speech and Abusive Language Detection Datasets, 29 May 2019.
[95] G. Doyle, Media Ownership, London: Sage Publications Ltd., 2002, 12.
[96] Council of Europe, CM/Rec(2018)1 / Recommendation of the Committee of Ministers to member States on media pluralism and transparency of media ownership, adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies.
[97] This Paper recognizes that different internet intermediaries have structurally a greater ability to act as gatekeepers compared to others. For application-layer services, ability to act as a significant gatekeeper instead correlates with market dominance.

encourage internet intermediaries to expose their users to all potentially available content.[98] While some sets of tools for sorting through and finding information is needed to help users find their way in an overwhelming amount of content, internet intermediaries oftentimes use AI to increase users' time spent on the platform in order to better monetize their data. With this in mind, intermediaries deploy AI to categorize individuals into groups and to determine their particular political, commercial and other preferences. Based on the users' data, as well as on the characteristics of the group to which – according to the AI code – the user belongs, intermediaries target each individual with specifically curtailed content. Social media platforms do not have economic incentives to expose their users to plurality and diversity. On the contrary, since users' attention is usually best kept with only a tiny portion of the available information, platforms have economic incentives not to do so. As a result, this process of social sorting exposes users to content that 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.[99]

This process, often referred to as an "echo chamber", leads to a situation where "individuals are increasingly cocooning themselves in the informational and communicational universe of their own creation."[100]

[98] C.f. Context of AI processes (above), about the need for intermediaries to offer users system for sorting out information on their platforms.

[99] 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. See also N.Statt, Facebook reportedly ignored its own research showing algorithms divided users, The Verge, 26 May 2020.

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

This creates information asymmetry, as users are prevented from accessing diverse information, and are also not even aware of the fact that the content they see does not reflect all available information, which might stop them from looking for diverse information.[101]

Media outlets and journalists are struggling to adjust to these new dissemination practices underpinned by AI. As some experts have pointed out, "it may be easy to speak in cyberspace, it remains difficult to be heard".[102] The already complex relationship between internet intermediaries and media outlets is further complicated by the deployment of AI-powered tools. Aspects of societal inequalities also need to be taken into account and explored further in connection with AI. This is particularly important with regards to inequalities in access and production of information, and how AI technologies can reproduce gender biases.[103]

AI-Powered State Surveillance

State mass surveillance has a chilling effect on freedom of expression, as it intersects with the right to privacy online and offline.[104] Freedom of expression and the right to privacy mutually reinforce one another,

[101] B. Kostic and T. 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.
[102] M. Hindmann, The myth of digital democracy, Princeton and Oxford University press, 2009, p.142.
[103] S. Noble, Algorithms of Oppression: How Search Engines Reinforce Racism, 2018; WIRED, Machines Taught by Photos Learn a Sexist View of Women; C. Collett and S. Dillon, AI and Gender: Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence, 2019.
[104] C.f. e.g., UN, Human Rights Committee, draft General Comment No. 37, Article 21: right of peaceful assembly, draft prepared by the Rapporteur, Christof Heyns, July 2019, para 69; EDPB, Guidelines 3/2019 on processing personal data through video devices, Version 2.0, adopted on 29 January 2020. See also the European Court of Human Rights, Copland vs the UK, App. Nos. 62617/00, 3 July 2007, para. 42.

with privacy being a prerequisite to the exercise of freedom of expression.[105] The use of AI-powered surveillance techniques by governments and the private sector, and a wide merging of data in public-private partnerships, specifically impacts the right to freedom of expression in the following ways:

First, AI-powered surveillance techniques affect the right to remain anonymous along with people's ability to seek and receive information without being identified on- and offline. Just as people are much more likely to speak freely when they know that their privacy is protected, the knowledge that their communication is highly likely to be inspected risks having a profoundly damaging effect on the free flow of information and ideas.

Second, permanent surveillance practices, coupled with profiling,[106] can have dangerous consequences for media and journalism, both on- and offline. In particular:

- The use of AI-powered surveillance techniques, as shown in several reports, severely **impede** the ability of journalists to conduct their **research and investigations**, to publish their work to specific or general audiences, and may lead to self-

105 See, e.g., Report of the Special Rapporteur on FoE on the implications of States' surveillance of communications on the exercise of the human rights to privacy and to freedom of opinion and expression, A/HRC/23/40, 17 April 2013, para 79; OHCHR, The right to privacy in the digital age, A/HRC/27/37, 30 June 2014, para 25; European Parliament, Report on the US NSA surveillance programme, surveillance bodies in various Member States and their impact on EU citizens' fundamental rights and on transatlantic cooperation in Justice and Home Affairs, 2013/2188(INI), 21 February 2014; Council of Europe Commissioner for Human Rights, The rule of law on the Internet and in the wider digital world, Issue Paper, December 2014, p. 16. Council of Europe, Resolution 2045 (2015) – Mass surveillance, 21 April 2015, para 4; or ARTICLE 19, The Global Principles on Protection of Freedom of Expression and Privacy, 2017.

106 Profiling has been defined as "automated decision-making, about people, from people's data, will shape their lives -- what they have access to, what they can do, and what they may become"; see Privacy International, Profiling.

censorship.[107] This can then inhibit the important functions that the media has in holding governments to account.

- The **chilling effect** of the use of surveillance was demonstrated by a 2016 study published in the Berkeley Technology Law Journal, which found a dramatic fall in monthly traffic to Wikipedia articles about terrorist groups and their techniques after Edward Snowden had disclosed information in 2013 about the U.S. domestic surveillance program.[108] The study found that article views dropped by 30 percent after June 2013, supporting "the existence of an immediate and substantial chilling effect."[109] The 2016 study on the impact of government surveillance on social media users,[110] in which the participants were informed of monitoring by the U.S. National Security Agency and showed a fictional Facebook post regarding U.S. airstrikes against the terrorist group ISIL/DAESH, provides another example. The study showed that people who are aware of government surveillance are significantly less likely to speak out when their views differ from what they perceive to be the majority opinion. The study concluded that "the government's online surveillance programs may threaten the disclosure of minority views and contribute to the reinforcement of majority opinion."[111]

[107] See, e.g., Privacy International, Two sides of the same coin – the right to privacy and freedom of expression, 2 February 2018; or Association for Progressive Communications, The right to freedom of expression and the use of encryption and anonymity in digital communications, Submission to the UN Special Rapporteur on Freedom of Expression, February 2015, p. 11.

[108] J. Penney, Chilling Effects: Online Surveillance and Wikipedia Use, Berkeley Technology Law Journal, Vol. 31, No. 1, 2016, p. 117.

[109] Ibid

[110] E. Stoycheff, Under Surveillance: Examining Facebook's Spiral of Silence Effects in the Wake of the NSA Internet Monitoring, Journal of Mass Communication Quarterly, 93(2), 296-311.

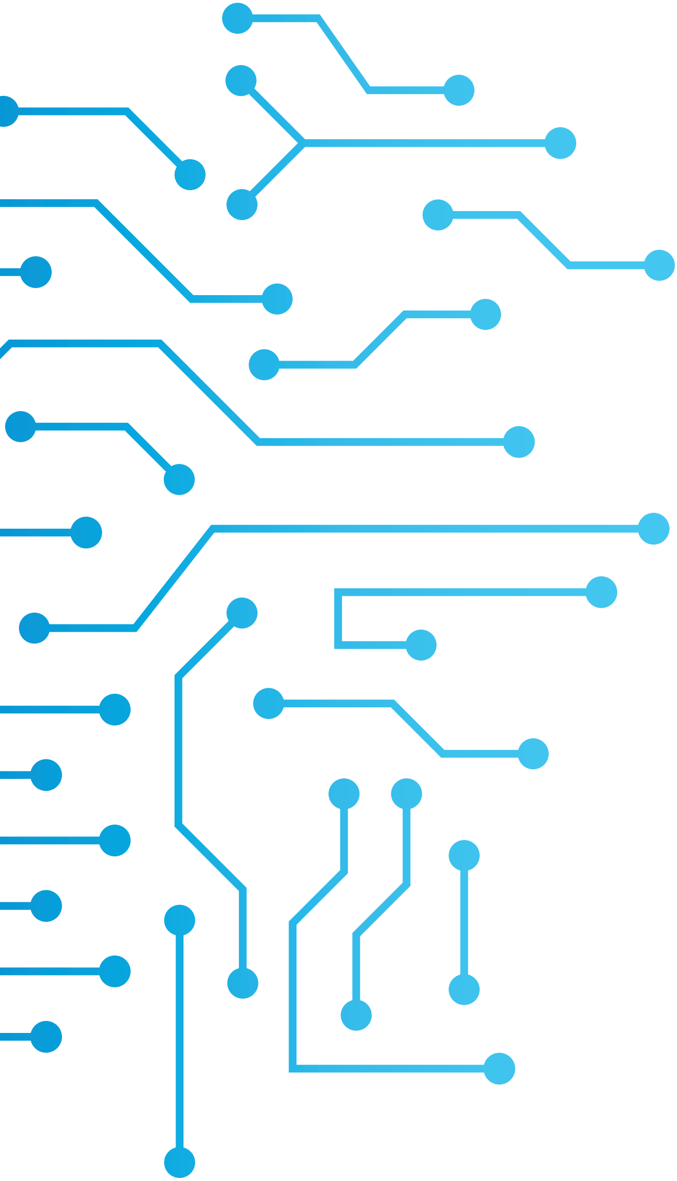
[111] Ibid.

- The **impact of surveillance on journalistic work** is evident in connection with the protection of journalists' sources and whistle-blowers. The already mentioned 2018 report by Citizen Lab, for example, documents that the mere perception of being under potential surveillance would lead to self-censorship of journalists and their sources.[112] Studies also show that there is a real risk that State actors could pass on the communications of journalists and whistle-blowers to a foreign government, with further risks of retaliation for the individuals concerned.[113] These concerns go beyond online activities and can have an impact on journalistic activities offline. AI-powered facial recognition, for example, can be used to identify journalists reporting on protests, or tracing back the digital footprints of individual journalists, especially those with dissenting views, and poses risks for the protection of journalists' confidential sources.

[112] Op. cit.

[113] See, e.g., Association for Progressive Communications, The protection of sources and whistleblowers Submission to the United Nations Special Rapporteur on the Right to Freedom of Opinion and Expression, 29 June 2015; ARTICLE 19, Response to the Special Rapporteur Consultation on Protection of Journalists' Sources and Whistleblowers, July 2015; or Center for Constitutional Rights, Written Submission on the Protection of Sources and Whistleblowers to the UN Special Rapporteur on Freedom of Expression, 22 June 2015.

Conclusion and Recommendations



Conclusion and Recommendations

This Paper outlined key implications of AI on freedom of expression and media freedom, and identified key issues that should be further studied and monitored. This Paper aims to contribute to ensuring that freedom of expression and media freedom are safeguarded in the deployment of AI, both by States and private actors.

There are a number of regional standards developed on AI and human rights, which are applicable in parts of the OSCE region, in particular by the Council of Europe and the EU. However, it is necessary that awareness and a better understanding of the impact of AI related to decision-making policies and practices on freedom of expression are promoted across the entire OSCE region. In particular, it is important to have a more systematic overview of the regional approaches and methodologies. To this extent, more regional and country-specific studies on positive and negative practices, as well as the exchange of knowledge and expertise at all levels, both horizontally and vertically, should be encouraged. This could include research into how AI-powered surveillance, content moderation and content curation affect freedom of expression and media freedom in the region, how journalists could benefit from AI in their work, and how to mitigate the possible discriminatory effects of AI on marginalized groups. In order to prevent problems in content removal of specific content (in particular "hate speech" and speech presenting "security threats"), studies on the effectiveness of AI tools designed to identify illegal content and promote counter speech online should be developed.

Addressing the impact of AI on freedom of expression is primarily the responsibility of States, as they have an obligation to create an enabling environment for freedom of expression that ensures diversity and pluralism of sources and views.[114] However, as demonstrated in this Paper,

[114] See e.g., 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

it must also be addressed on the level of corporate responsibility of internet intermediaries.[115] Given the role of the State and the public sector in this regard, effectively safeguarding freedom of expression in the development and deployment of AI requires multi-stakeholder dialogues and co-operation. Hence, multi-stakeholders' processes, designed to address concerns in the deployment and use of AI in surveillance, content moderation and curation, are critical.

Greater transparency about the question of how AI is deployed by State and private actors is needed, in terms of users, researchers and regulatory bodies. Further research on data access and transparency is needed.

In light of recent developments and measures adopted to address the COVID-19 pandemic, it is also important that new AI-powered surveillance measures and companies' enhanced reliance on AI in content moderation are used in a strictly temporary manner, and will not be normalized.

Looking forward, the OSCE RFoM makes the following preliminary recommendations:

OSCE participating States should:

- Evaluate whether their domestic legal and policy frameworks applicable to AI fully incorporate international human rights standards on freedom of expression, privacy and data protection.
- Adopt legal and policy frameworks that fully enable freedom of expression in the digital ecosystem. This includes updating and applying existing regulation, particularly data protection regulation, to AI.

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.

[115] C.f. The Guiding Principles, op.cit.

- Ensure that any technological and regulatory measures regarding AI are human rights-based, in order not to limit freedom of expression, media freedom and other fundamental rights. All restrictions must serve a legitimate aim, as explicitly promulgated by international human rights standards. They must have a proper legal basis and be necessary in a democratic society in the pursuit of the legitimate aim and the use of minimally intrusive measures to achieve such an aim. Any possible human rights harm caused must remain proportionate to the actual benefit obtained towards the legitimate aim. In particular, legal and policy measures restricting speech (including "hate speech" and speech presenting "security threats") should be evidence-based and strictly necessary. In this context, participating States should distinguish clearly between the need to combat illegal speech and the risk of chilling lawful speech.
- Refrain from arbitrary or unlawful interference with journalists' use of encryption and anonymity technologies, and refrain from employing unlawful or arbitrary surveillance techniques.[116]
- Be transparent about the use and underlying functionalities of AI, especially regarding data sharing, the question of which datasets feed into AI systems, and potential biases and inaccuracies of the systems used, as well as proposed means to overcome them. Be transparent about public-private partnerships in this area, and conduct regular assessments of the impact that these partnerships have on freedom of expression and freedom of the media.
- Put in place policy and regulatory frameworks on AI-driven algorithmic transparency and explicability. In particular, States should ensure that all national oversight mechanisms (i.e., such as public defenders of human rights, or independent public control of security services) regularly monitor, record and include in their

[116] Organization for Security and Co-operation in Europe (OSCE), Ministerial Council, Decision No. 3/18, 'Safety of Journalists', 7 December 2018, Milan.

reports their findings and concerns on the use of AI in the contexts of surveillance or general "security threats".

- Adopt legal and regulatory frameworks that would require internet intermediaries to conduct human rights impact assessments in relation to AI systems, in particular those procured or used by public authorities. When doing so, States should acknowledge the diversity of audiences and pay due attention to groups of users at risk of discrimination and their special needs, and ensure specific and tailor-made interventions enabling access to, and use of, AI-based communication processes.
- Ensure that research, development, and use of AI fully complies with international human rights standards. This should include jointly developing an understanding of what constitutes "AI human rights critical systems", and ensuring that laws and regulations, codes of conduct, ethical codes, and self-regulatory and technical standards meet the threshold set by international human rights standards, as well as conducting periodical reviews to ensure compliance.
- Ensure a competitive field in the AI domain. In this respect, States should adopt regulation to prevent the concentration of AI expertise within a few dominant companies. They should also introduce regulation designed to increase interoperability of services and technologies, and adopt policies supporting network and device neutrality.[117]
- In order to mitigate the impact of AI on pluralism and diversity, States should support the development of sustainable and alternative business models to enhance the availability of diverse sources of information, including quality news sources and public service media. They should identify and promote AI-supported content production and share practices that promote an enabling environment for freedom of expression, especially for marginalized

[117] The August 2018 Report of the UN Special Rapporteur on Freedom of Expression on artificial intelligence technologies, op.cit.

voices. States should also adopt policy measures and make funding available for promoting diversity in access to innovative technologies, including those focusing on local media, start-ups, and non-major language media providers.

- Support the development of ethical standards for AI that reflect the protection of human rights in all stages of development and implementation of AI. Ethical standards could include prohibition of manipulation of peoples' behaviour through AI technologies, and ensure protection of human rights by AI design.
- Ensure that all national strategies on AI give full consideration to respect for human rights, including protection of the right to freedom of expression, access to information, and freedom of the media.
- Encourage platforms to enhance users' agency and choice, and implement the principle of privacy by design in respect of any AI system, and ensure that such techniques are fully compliant with the relevant privacy and data protection law and standards.[118]

Internet intermediaries are recommended to:

- Ensure that the protection of human rights is central to private sector design, deployment and implementation of AI systems. Internet intermediaries should affirm, or reaffirm, their commitment to the UN Guiding Principles on Business and Human Rights. Ensure that the Principles guide all of their operations and activities in the development and use of AI systems, and embed international human rights standards, in particular those on freedom of expression.
- Undertake rigorous human rights assessments of all AI systems developed and deployed by intermediaries, throughout their entire life cycle (from the design, the use of datasets to the deployment of the AI systems), and create feedback and continuous auditing mechanisms for

[118] Council of Europe, Recommendation on media pluralism and transparency of media ownership, op.cit.

AI systems and usage. The results of human rights impact assessments and public consultations should be made public.[119]

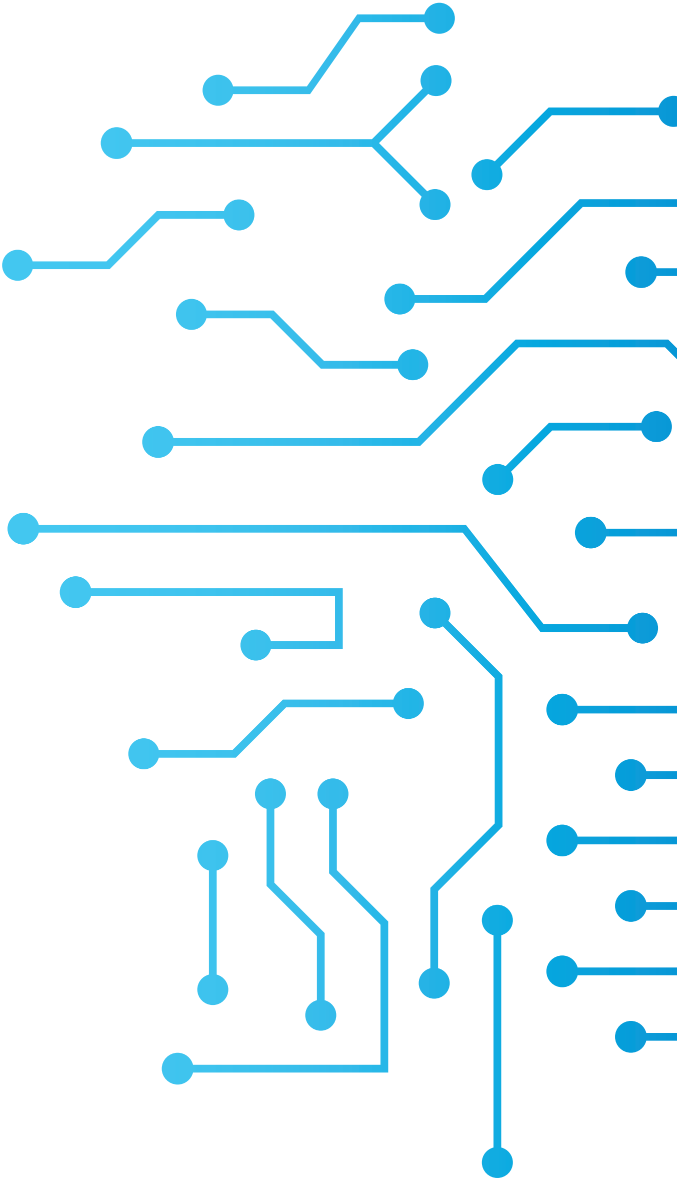
- Adopt codes of conduct, or codes of ethics, on AI that are grounded in human rights principles.
- Enhance transparency and accountability of intermediaries' use of AI in their operations. Intermediaries should publish transparency reports with aggregated statistics, accurately reflecting the usage of AI in content moderation and curation. These reports should include information on how internet intermediaries rank and profile content, how they target users with certain content on their own initiative, and how they moderate the content on their platforms. The transparency reports should be released periodically and in a timely manner, so that they can be scrutinized while they are valid.[120] Tiered access to information should be considered, distinguishing between access for users (to help them understand more fully the systems they use), access for researchers to conduct studies, and access for independent regulators.[121]
- When setting up industry and technical standards on AI, engage in multi-stakeholder initiatives, and provide opportunities for engagement for a wider range of stakeholders, including civil society and representatives of groups at risk of discrimination. In particular, non-binding frameworks must be accompanied by strong accountability and oversight measures.
- Adopt measures to enhance user's agency and choice, and implement the principle of privacy by design in respect of any AI system, in full compliance with standards on privacy and freedom of expression.[122]

[119] The August 2018 Report of the UN Special Rapporteur on Freedom of Expression on artificial intelligence technologies, op.cit.

[120] M. MacCarthy, op.cit.

[121] The French Interim report, op.cit.

[122] Council of Europe, Recommendation on media pluralism and transparency of media ownership, op.cit.



Artificial intelligence (AI) has become an increasingly important tool for shaping and arbitrating online information. It is increasingly, and often invisibly, used by both public authorities and private companies, and greatly impacts the way people seek, receive, impart and access information. Coupled with its ability to identify, surveil and track people, AI can seriously impede on the right to freedom of expression. This #SAIFE Paper puts a spotlight on AI and freedom of expression, and provides guidance and preliminary recommendations on how to effectively safeguard free speech when AI is deployed.

osce
The Representative on
Freedom of the Media