Mia Georgievska*

SAFEGUARDING FREEDOM OF EXPRESSION IN THE AGE OF ARTIFICIAL INTELLIGENCE

Abstract

The world in the 21st century is moving at an unprecedented speed in the field of technological development, and new technologies based on artificial intelligence have now become deeply rooted in human reality. Independence and autonomy of computer technologies offer speed, efficiency and lightning-fast solutions for filtering, ranking and sharing information on the internet, while profoundly transforming the way people communicate, consume and engage with media, governance and society.

The ability to rapidly collect and analyze personal data inevitably impacts the freedom of expression, as, to be able to freely form political, religious or other opinions and views, and to freely express them, an individual needs a private domain without interference from the state, private sectors or society. AI-driven surveillance and monitoring tools, when used without transparency or accountability, increase the risk of profiling, censorship, and repression. In the absence of strong data protection laws and independent oversight, individuals may even self-censor, fearing they are being watched or judged for their views. These consequences can potentially undermine the democratic values of accountability, integrity, adherence to the rule of law and respect for human rights, particularly in contexts where legal protections and institutional capacities are still evolving.

This paper offers a comprehensive analysis of how the right to freedom of expression is challenged by artificial intelligence systems and algorithms used on the online platforms. It further analyzes the obligations that online platforms, *i.e.,* Internet intermediares, own in respect to the use of AI-driven systems, and how different legal systems address the novel challenges that generative AI poses to the freedom of expression. The paper aims to demonstrate the need to properly and carefully address and adopt concrete legal frameworks to regulate the use of artificial intelligence, while also arguing that the present interpretation of freedom of expression is able to address the unique challenges posed by AI technologies.

Keywords: Freedom Of Expression; Artificial Intelligence; Legal Accountabily; Online Platforms; Human Rights;

I. INTRODUCTION

The world today is moving at an unprecedented speed in the field of technological development, and new technologies based on AI are deeply rooted in human reality. The modern information society has imposed a new way of conducting human everyday life: connecting through the development of the Internet and online communication platforms. Online platforms occupy a central place in the way people connect, communicate, socialize, act, get informed and share information. Traditional media are being replaced by social media,

* Mia Georgievska, Attorney at Law, Ph.D. Candidate in International Public Law, Ss. Cyril and Methodius University in Skopje, Iustinianus Primus Faculty of Law, email: mia@georgievskalaw.com; ORCID 0009-0005-6945-6094.

news "aggregators", search engines and online platforms. The European Court of Human Rights recognizes the Internet as crucial in "improving public access to news and facilitating the dissemination of information in general",[1] but also as "one of the main means through which individuals exercise their right to freedom of expression."[2] Unlike the past, this 'new world' offers every person the right and opportunity to express their opinion and share opinions and ideas on a globally. Today, everyone can be informed about daily, national, regional and global events with just one online click. It is precisely the ability of each individual to express themselves and share opinions and ideas through online platforms that creates huge amount of news and information, texts, images and videos available online. The use of artificial intelligence in the online space has enabled a further transformation of the way people communicate, consume content and engage in content creation. AI applications and algorithms "are found in every corner of the internet, on digital devices and in technical systems, as well as in search engines, social media platforms, messaging applications and public information mechanisms."[3] From the multitude of information available, AI offers attractive solutions for filtering and ranking "what seems like" an infinite amount of content and information created by Internet users.[4]

In this sense, there are three key areas of use of AI on online platforms. First, the use of AI to personalize the experience of each user by filtering, ranking and displaying content that would interest a particular user, *i.e.* not displaying or excluding potentially "unwanted" content (content curation).[5] The second area of use of AI is the moderation and removal of content that does not comply with the standards and rules of a particular platform, namely, the evaluation of available content by AI-algorithms and the removal, suspension, blocking or deletion of a particular user from the platform.[6] The third area of application concerns profiling, targeting and advertising, that is, the use of algorithms as business models for offering services by analyzing huge data bases on each individual and targeted marketing.[7]

II. DEFINING ARTIFICIAL INTELLIGENCE AND ITS APPLICATION ON ONLINE PLATFORMS

There is no universally accepted definition of what artificial intelligence is, nor is there a global international agreement regulating artificial intelligence. One of the international legal instruments, which also provides a comprehensive definition of what AI is, is the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.[8] The Draft Convention was created with the intention of becoming the first international agreement of a binding nature that aims to "ensure that activities within the life cycle of artificial intelligence systems are fully compatible with human rights, democracy and the rule of law", while also being conducive to technological progress and innovation.[9] Within the framework of this international document, an artificial intelligence system is defined as "a

[1] European Court of Human Rights, *Times Newspapers Ltd v UK*, Case No. 3002/03, 23676/03, 2009, ¶27.
[2] European Court of Human Rights, *Ahmet Yildirim v. Turkey*, 2012, Case No. 3111/10, ¶54.
[3] David Kaye, Promotion and Protection of the Right to Freedom of Opinion and Expression, United Nations, 2018, A/73/348, ¶9, [Hereinafter "Kaye, 2018"]
[4] Kaye, 2018, ¶9.
[5] Kaye, 2018, ¶10.
[6] Julia Haas, Freedom of the Media and Artificial Intelligence, Global Conference for Media Freedom, 2020, p. 3 [hereinafter "Haas"]; Kaye, 2018, ¶13.
[7] Kaye, 2018, ¶17.
[8] Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Council of Europe, 2024.
[9] Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024.

machine-based system that, for explicit or implicit purposes, infers, from the input data it receives, how to generate outputs such as predictions, content, recommendations or decisions that can affect physical or virtual environments. Different artificial intelligence systems differ in their levels of autonomy and adaptability after deployment."[10] However, it is necessary to mention that this definition was "prepared for the purposes of the Framework Convention and is not intended to give a universal meaning to the relevant term […], although this definition provides a common understanding among the Parties of what artificial intelligence systems are," and it is drafted in such a way as to allow for its further specification in domestic legal systems for additional legal certainty and precision, without limiting its scope."[11] The definition given in the EU AI Act as a second significant instrument of a binding nature for the member states of the European Union, as well as for all those actors who apply an AI system on the territory of the EU (extraterritorial application).[12]

In its essence, AI is about creating intelligent machines, *i.e.* technology that allows computers and machines to simulate human thought and learning, understanding, problem solving, decision-making processes, creativity and autonomy. Thus, AI envisages a new concept of individuality of the human mind, which is no longer unique in performing tasks that require cognitive functions such as thinking, learning, predicting and solving a certain problem. These systems analyze a huge amount of data, analyze examples, learn and discover patterns, and in the end, offer a certain solution, a desired result. These systems recognize images, events, people and human behavior, and they, through algorithms of the so-called "machine learning", independently develop techniques for further learning and intelligence, that is, continuously learn from the data they independently collect, and make appropriate decisions based on what they have learned.

It is precisely these characteristics of AI that make these systems attractive for use on online platforms, precisely as a result of their ability to adapt, search, filter, rank and (not) display content at a speed that is impossible for the human mind. As previously established, there are several areas of application of AI systems on online platforms, namely: personalization (filtering) of content according to the interests of the user (content curation) and content moderation (content moderation).

III. CONTENT CURATION

Content personalization, also known as content curaton, refers to the use of AI-driven algorithms to filter news and information for users on the online platforms based on their personal prefrences and desires.[13] The Internet is clasified as one of the "most widespread forms of personalization and content display in the history of artificial intelligence."[14] Internet intermediaries, *i.e.,* the entities that own the online platforms, carefully observe and store every piece of data about each user's online activity, while parallely using this additional data to create so-called "individual user models."[15] Based on these individual user models, AI-driven

[10] Article 2, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
[11] Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024, ¶ 24.
[12] Article 3, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).
[13] Emina Kolarevic, The Influence of Artificial Intelligence on the Right of Freedom of Expression, 2022, p. 115, available at https://orcid.org/0000-0002-3270-6740, [hereinafter "Kolarevic"]; Haas, p. 3.
[14] Kaye, 2018, para 11.
[15] Kolarevic, p. 115.

systems predict which news and information are relevant to the user and which are not, thus, the system displays relevant information and omits irrelevant information.[16] As a result, having in mind that that each internet search engine differs from user to user, a situation occurs, in which individuals, with the same friends on social media, of the same or similar age and background, could have access to different news and different information based on the previous interaction with the internet platform.[17]

Essentially, "artificial intelligence algorithms that determine how much, when, with which audience and with which individuals [share] content."[18] Such individual user models are trained to use data that is combined with browsing history, to analyze the demographics of the user and his or her semantic abilities, to analyze the content searched, as well as many other factors that AI-driven systems use to rank and filter information that would be useful for each specific user. Namely, the information and news are displayed based on how interesting or attractive the content could be for the specific user, based on the substantive assessment of the AI system used.[19] The process of (non)displaying certain content offers little or no exposure to the user to news and updates that are different from the preferences of the specific individual, and news or information which may have a different attitude, criticism or perspective than what is offered to the individual.[20] Such personalized could lead to strengehing "users' pre-existing views, creating "echo chambers" and "filter bubbles", while "decreasing the likelihood of individuals' exposure to diverse media content".[21] Shaping the information by the AI systems is often not visible to the users, and many times, it is not visible even to the platform itself, as these systems operate autonomously.[22]

IV. CONTENT MODERATION

Content moderation refers to the process of using AI-driven systems to identify and remove content generated and shared by users on online platforms, *i.e.* examining whether a certain shared content complies with legal requirements or social media standards and rules.[23] Content moderation, in its essence, means assessing, analyzing and suspension, blokage or removal of content that threatens national security, content that incites hatred, violence or terrorism, content that qualifies as hate speech, that shows nudity, child exploitation, or content that is essentially prohibited under many laws around the world.[24]

In cases of content moderation, AI systems analyze the content shared by users, evaluate it through assessment filters according to defined criteria, and, if certain content is assessed as prohibited or undesirable, these systems automatically block the content and enable its deletion.[25] In certain cases, sharing prohibited content may result in suspension or temporary or permanent deletion of a certain user profile.[26] Content moderation allows the AI-driven

[16] Ibid.
[17] Kolarevic, p.115; Haas, p.3. See also: Engin Bozdag, Bias in algorithmic filtering and personalization, *Ethics and Information Technology,* 2013.
[18] Kaye, 2018, ¶ 10.
[19] Kaye, 2018, ¶ 10.
[20] Ibid.
[21] Haas, p.3. See also: C. R. Sunstein, *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*, Princeton University Press, 2001; B. Bodó et al., "Interested in Diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization," *Digital Journalism*, 2018; N. Helberger, "Challenging Diversity – Social Media Platforms and a New Conception of Media Diversity," *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, New York: Oxford University Press, 2018.
[22] Haas, p.3.
[23] Kolarevic, p. 119; Haas, p. 3.
[24] Ibid.
[25] Ibid.
[26] Ibid.

system to filter and prohibit sharing content that is illegal or that does not comply with the "terms of use" of online platforms.[27]

V. TARGETED ADVERTISING

AI-driven targeted advertising refers to "using artificial intelligence algorithms to deliver ads that match specific audiences with precision."[28] As such, AI-driven systems are trained to use a vast amount of data to "identify patterns, predict user behavior, and optimize ad placements" with the aim to ensure that "relevant messages reach receptive audiences at optimal moments."[29] Online platforms, *i.e.* internet intermediares, apply AI to disseminate information that is based on the predicted preferences of individual users, as this data also "facilitate[s] advertising, which is the basis of many internet intermediaries busines model."[30]

While the traditional concept of targeting used demographic segments to place certain products on the market, Ai-driven targeting operates on a different approach.[31] In traditional targeting, advertisers "chose broad categories like age ranges, gender, and location, then relied on third-party cookie data to retarget users who visited their websites."[32] Unlike traditional advertising, AI-targeted advertising allows algorithms to "automatically analyze hundreds of signals simultaneously: browsing behavior, content context, time of day, device type, previous interactions, and more."[33] Such algorithms analyze data and predict user intent "often uncovering micro-segments and behavioral patterns that human marketers would miss."[34] In that sense, AI-driven system might learn that a "user interested in hiking gear also responds positively to travel ads specifically in the evening hours", and "identifies it automatically through pattern recognition across millions of data points", while traditional marketing would "likely miss this nuanced timing and cross-category connection."[35]

VI. CHALLENGES TO THE FREEDOM OF EXPRESSION POSED BY AI-DRIVEN PERSONALIZATION AND MODERATION TOOLS

The use of AI-drive tools to personalize and moderate content on the online platforms poses certain concerns on safeguarding the right to freedom of expression, as a fundamental human right protected under international human rights law. Such challenge is the ability of AI systems to identify harmful content, which in turn affects the freedom of expression of users on online platforms.

The most famous example of content moderation, which also raised the question of the potential of AI in this domain, is the removal of the iconic Pulitzer Prize-winning photograph "The Terror of War", which shows a naked girl running after a bomb attack during the Vietnam War, by Facebook, on the grounds that the photo violated the social media standards, namely

---

[27] Giovanni De Gregorio, Pietro Dunn, Artificial Intelligence and Freedom of Expression, Artificial Intelligence and Human Rights, Oxford University Press, 2023, Chapter 5, p. 7.
[28] Mary Gabrielyan, *AI Targeted Advertising: How Smart Targeting is Redefining ROI and Personalization*, 2025, avaiable at: https://www.aidigital.com/blog/ai-targeted-advertising.
[29] Ibid.
[30] Haas, p. 2.

[31] Mary Gabrielyan, *AI Targeted Advertising: How Smart Targeting is Redefining ROI and Personalization*, 2025, avaiable at: https://www.aidigital.com/blog/ai-targeted-advertising.
[32] Ibid.
[33] Ibid.
[34] Ibid.
[35] Ibid.

the ban on nudity in the context of images of child abuse.[36] After a wave of negative publicity, Facebook reversed its decision and acknowledged the importance and value of the photo.[37]

This example best illustrates the limitations that artificial intelligence has in terms of the process of identifying and removing harmful or prohibited content created by users on online platforms. It is about the limited capacity of artificial intelligence to analyze content.[38] Namely, the evaluation of a particular speech depends largely on the context in which a particular speech is expressed, and this requires a deep understanding of cultural, linguistic, political and sociological factors.[39] In this sense, the development of AI has not yet reached the level at which these systems can distinguish "between news reporting, support [or promotion of a particular idea, group or right,] and satire on the one hand, and on the other, the very incitement to harm [violence, hatred]."[40] For these reasons, AI, when assessing the context of the content itself, "is prone to errors: it can identify illegal content as permitted, resulting in "false negatives" or "false positives" in the case of removal of legitimate content."[41] One such example is the videos uploaded by journalists and activists to YouTube in 2017, which depicted alleged war crimes in Syria, which were identified by YouTube's algorithms as terrorist propaganda and were removed from the platform.[42] In all of these situations, of paramount importance in assessing whether a piece of content is legal or not, *i.e.* whether it complies with the terms of use of a particular platform or not, is the context in which the content itself is displayed and shared, and thus, content provided in one context may promote violent and extremist behavior, while in another context, it may be crucial for reporting and news or for combating recruitment of extremists online.[43] Furthermore, "public debates on issues of public interest can often be heated in their language and accompanied by offensive or figurative language, with irony or mockery, which does not constitute hate speech, but which artificial intelligence systems can easily identify as such and remove."[44] What AI cannot yet identify is processing the context of a particular speech or content – it cannot "distinguish between a sarcastic response to hate speech from a genuine hateful comment."[45]

Additionally, AI-driven systems, when determining that certain content should not remain avaiable in the online space, have the "authority" to remove "unwanted" content, as well as the ability to suspend or block certain user profiles. The possibility of implementing such measures undoubtedly means a restriction of the freedom of expression of a certain individual, hence, the question of the legitimacy of the restriction of this freedom arises. According to the European Convention on Human Rights, freedom of expression may be subject to restrictions when such restrictions are prescribed by law and necessary in a democratic society for the achievement of one of the legitimate social aims.[46] The removal and blocking of content constitutes a restriction of the freedom to share information and ideas, and such a restriction

---

[36] The whole article is avaiable at: https://www.bbc.com/news/technology-37318031.

[37] The whole article is avaiable at: https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

[38] Haas, p. 3.

[39] Ibid.

[40] Hu, Neupane, Echaiz, Sibal & Lam, *Steering AI and Advanced ICTs for Knowledge Societies: A Rights, Openess, Access and Multi-stakeholder Perspective*, Paris: UNESCO Publishing, 2019, p. 38.

[41] Eliska Pirkova, Matthias Kettemann, Marlena Wisniak, Martin Scheinin, Emmi Bevensee, Katie Pentney, Lorna Woods, Lucien Heitz, Bojana Kostic, Krisztina Rozgonyi, Holli Sargeant, Julia Haas, and Vladan Joler, *Spotlight on Artificial Intelligence and Freedom of Expression, A Policy Manuel,* OSCE: The Representative on Freedom of the Media, 2020, p. 56, [hereinafter: "Pirkova and others"].

[42] See: https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.

[43] Pirkova and Others, p. 58.

[44] Kolarevic, p. 120.

[45] Hu, Neupane, Echaiz, Sibal & Lam, Steering AI and Advanced ICTs for Knowledge Societies: A Rights, Openess, Access and Multi-stakeholder Perspective, Paris: UNESCO Publishing, 2019, p .40.

[46] Article 10 (2), European Convention on Human Rights, Council of Europe.

should respond to a "pressing social need" and pass the proportionality test, *i.e.* it should be proportionate to a certain legitimate aim.[47] The European Court of Human Rights found that the measure of blocking Google user accounts in Turkey "produced arbitrary effects" that led to the restriction of users' rights because a huge amount of information became inaccessible to the public, and considering that "delaying the publication [of the removed news], even for a short period, is likely to deprive them of their value and interest."[48] In addition, the removal of online content or the blocking of access to the internet requires a legal framework that would ensure strict control over the scope of the restrictions and effective judicial review aimed at preventing abuse of power, while "judicial review of such a measure, based on a weighing of the competing interests at stake and designed to strike a balance between them, is inconceivable without a framework that establishes precise and specific rules regarding the application of preventive restrictions on freedom of expression."[49]

In terms of content curation, it undoubtedly affects the right to freedom of expression in several ways. A personalized online experience, which is always based on the personal preferences and interests of each user, placed in a space where users are offered little or no exposure to opinions contrary to their own, undoubtedly affects and potentially threatens freedom of thought, which includes the protection of the individual's right to hold an opinion and "hold opinions without interference" as an absolute right whose limitation is not permitted. Thus, in one of the Reports of the United Nations Special Rapporteur on Freedom of Expression, it was noted that "the ways in which information is stored, transmitted and provided in the digital age have a significant impact on the exercise of the right to hold opinions," and that internet searches and browsing, text communications, as well as documents and data files stored in online clouds, collectively referred to as digital activities and records, form the structure of individuals' pre-formed opinions.[50] This approach allows for interference in freedom of thought not only by the state, but also by non-state actors who can influence the mechanisms and processes of forming and maintaining an already formed opinion.[51]

Additionally, the enjoyment of freedom of expression is closely linked to the exercise of other rights and is a basis for the efficient functioning of democratic institutions, which in its essence implies the promotion of the diversity and independence of the media and the provision of efficient access to information.[52] By personalizing the experience of each user, this diversity is restricted, the quality of information that individuals receive is reduced, and the information requested and received is reduced to confirming pre-existing attitudes and already formed opinions, without the possibility of accessing another attitude or opinion, which could lead to the formation of a vulnerable political discourse.[53] This directly affects the freedom to share and receive information. Additionally, in developing and underdeveloped countries, the effects of personalization of online content can lead to "a worsening of the socio-political climate and further polarization and radicalization in society."[54]

---

[47] Kolarevic, p. 121.
[48] European Court of Human Rights, *Ahmet Yildirim v. Turkey*, 2012, ¶ ¶ 47, 66, 68.
[49] Ibid, ¶ 66.
[50] David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UNGA, A/HRC/29/32, 2015 [Hereinafter "Kaye, 2015"].
[51] Ibid.
[52] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Organization for Security and Co-operation in Europe Representative on Freedom of the Media, Organization of American States Special Rapporteur on freedom of expression and African Commission on Human and Peoples' Rights Special Rapporteur on freedom of expression and access to information, "Joint Declaration on freedom of expression and 'fake news'", disinformation and propaganda", 2017.
[53] Engin Bozdag, Bias in algorithmic filtering and personalization, *Ethics and Information Technology,* 2013, p. 2018.
[54] Kaye, 2018, ¶31.

i. The Obligation to Respect Human Rights by Online Platforms

In the context of the enjoyment of freedom of expression online, and in particular the fact that sharing and accessing information online and through online platforms remains a key area for the manifestation of this freedom today, the UN Human Rights Council "affirms that the same rights that people have offline must also be protected online, in particular [the right to] freedom of expression, which shall be exercised regardless of frontiers and through any media of [the individual's] choice, in accordance with article 19 of the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights."[55] This position of the UN Human Rights Council testifies that the same rights, obligations and responsibilities arising from international instruments for the protection of human rights must also be analyzed when using the Internet and online portals, especially in a situation where such use infringes on certain guaranteed human rights, such as, among others, freedom of expression. Consequently, the application of artificial intelligence in the online space, the application of which undoubtedly affects certain human rights, must also necessarily be analyzed from the perspective of existing international human rights law. Given that the current uses of artificial intelligence systems in the online space affect the right to freedom of expression, these technologies "must be designed, developed and deployed in a manner that is consistent with the obligations of States and the responsibilities of private actors under international human rights law."[56]

International human rights law offers "a consistent framework for evaluating paltforms that operate globally."[57] Both the UN Human Rights Committee and reports of the Special Rapporteour call for private actors to govern the content on the online platforms in terms of the relevant international human rights treaties and instruments.[58] Additionally, the UN Guiding Principles on Business and Human Rights ackolewdge that business "should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved."[59] Such obligation "exists independently of States' abilities and/or willingness to fulfil their own human rights obligations, and does not diminish those obligations," and it "exists over and above compliance with national laws and regulations protecting human rights."[60]

Furthermore, the 2024 Report of the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence "called for AI governance to be firmly grounded in the UN Charter, IHRL, and related international commitments."[61] Similarly, the joint declaration by the UN Special Rapporteour, the OSCE Representative of the Freedom of the Media, the Organisation of the American States Special Rapporteour, and the African Commision on Human and Peoples' Rights Special Rapporteour, ackowledged that "AI design, development and deployment must be rooted in IHRL" and "urged a shift […] toward the proactive

---

[55] Human Rights Council, 26/13 The promotion, protection and enjoyment of human rights on the Internet, A/HRC/RES/26/13, 2014.
[56] Kaye, 2018, ¶ 19.
[57] Jordi Calvet-Bademunt, Jacob Mchangama, Isabelle Anzabi and Carlos Olea, Freedom of Expression in Generative AI Models, in *That Violates My Policy: AI Laws, Chatbots, and the Future of Expression*, The Future of Free Speech, October, 2025, p. 21, [hereinafter "Calvet-Bademunt and others, Freedom of Expression in Generative AI Models"].
[58] Calvet-Bademunt and others, Freedom of Expression in Generative AI Models, p. 21.
[59] United Nations, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, UN Human Rights Office of the High Commissioner, 2011, p.13, [hereinafter "UN Guiding Principles on Business and Human Rights"]
[60] Ibid. p. 13.
[61] Calvet-Bademunt and others, Freedom of Expression in Generative AI Models, p.22.

embedding of freedom of expression and information integrity as foundational design principles."[62]

As States have the primary obligation to respect, protect and fulfill the human rights of individuals within their terriotory and/or jurisdiction, States are obligated "to protect against human rights abuse within their territory and/or jurisdiction by third parties, including business enterprises."[63] Such protection "requires taking appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication."[64] In that sense, "States should set out clearly the expectation that all business enterprises domiciled in their territory and/or jurisdiction respect human rights throughout their operations."[65]

Thus, different legal systems treat AI-driven algorithms used by online platforms differently, in accordance with their embedded legal guarantees of protection of freedom of expression. In the following chapters, this paper examines how different legal systems address the novel challenges that generative AI poses to the freedom of expression, with the focus of the practices of the European Union (EU), the United States of America (USA) and North Macedonia.

VII. AI-DRIVEN SYSTEMS AND FREEDOM OF EXPRESSION IN THE EUROPEAN UNION (EU)

The EU has taken a cautious stance in the development, deployment and accountability for the use of AI-drive systems, with comprehensive and developed legal framework on the use of AI-drive systems while protecting the freedom of expression.

The freedom of expression, in the European context, applies to information and ideas that the majority considers "acceptable" or with "inoffensive" content,[66] while also extending to "the transmission of ideas that offend, shock, or challenge the established order."[67] The latter are considered crucial as they essentially represent a demand for pluralism, tolerance and broad-mindedness, "without which a democratic society cannot exist."[68] However, the right to freedom of expression is not an absolute right, and can be restricted provided that such restrictions are "prescribed by law", and "necessary in the democratic society in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary." [69] The EU Charter of Fundamental Rights, the European Convention on Human Rights,[70] as well as national laws of the EU member States constitute and represent the legal framework for protecting and promoting freedom of expression in the EU. At the particular moment, the EU has adopted the AI Act[71], the Digital

[62] Ibid.
[63] UN Guiding Principles on Business and Human Rights, p. 3
[64] Ibid.
[65] Ibid.
[66] European Court of Human Rights, *Handyside v. the United Kingdom*, 1976, ¶49; European Court of Human Rights, *Palomo Sanchez and Others v. Spain*, Case No. 28955/06, 28957/06, 28964/06, 2011, ¶53.
[67] European Court of Human Rights, *Women on Waves and Others v. Portugal*, No. 31276/05, 2009, para 42.
[68] European Court of Human Rights, Handyside v. the United Kingdom, 1976, ¶49.
[69] Article 10, European Convetion on Human Rights.
[70] Although the European Convention on Human Rights is not an EU Document, it applies to all member States of the EU, as all have ratified it. Additionally, the EU Charter of Fundamental Rights explicitly provides that the meaning and scope of the rights protected in the Charter must align with those laid out in the Convention.
[71] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU)

Services Act (DSA),[72] and has a growing body of related policies that tackle certain issues of importance, such as disinformation or hate speech.[73] Finally, the EU has signed[74] the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.[75]

Generally, the framework shows that EU has taken "an early and leading role in regulating AI, with the EU's AI Act at the forefront of globl efforts."[76] This Act is based on a rist-based approach, as it bans "few extreme use cases deemed to pose unacceptable risk, heavily regulates 'high-risk' applications, and imposes transparency requirements on more general systems with limited risk."[77] Although such attempts to regulate AI at the EU level is of paramount importance for the way its use will be shaped, there are several aspects within the regulation itself that indicate an increased risk to the right to freedom of expression.

One of these concerns is embedded in the DSA, legislation that has the aim to protect user rights by placing a regulatory requirement on online platforms to idetify and mitigate risks resulting form their online services, *i.e.,* systemic risk assesments.[78] The concept of "systemic risks" is left undefined in the DSA, and, as such, it could potentially be interpred as to encompass a wide range of controversial or unpopular ideas and speech, protected under the right to freedom of expression.[79] By encouraging preemtive moderation, it could "lead providers and platforms into over-moderation as a defensive measure."[80] Such concept, *i.e.*, "better safe than sorry" approach of certain platforms, could protect against some harms but it could also risk creativity, pluralism and political discurs as essential for a democratic society.[81] This argumenation is in line with the UN Special Rapporteour on Freedom of Opinion and Expression, who highlighted that imposing obligations on companies to restrict content on the basis of "vague or complex legal criteria without prior judicial review and with the threat of harsh penalties" could pose certain risks to freedom of expression, as it puts significant pressure on companies such that they may remove lawful content in a broad effort to avoid liability.[82] In that context, "the 2016 European Union Code of Conduct on countering illegal hate speech online involved agreement between the EU and four major companies to remove content, committing them to collaborate with 'trusted flaggers' and promote 'independent counter-narratives'." It was noted that "while the promotion of counter-narratives may be attractive in the face of 'extremist' or 'terrorist' content," the "pressure for such approaches runs the risk of

---

No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), [hereinafter "AI Act"]

.[72] REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), [hereinafter "Digital Services Act"].

[73] Jordi Calvet-Bademunt, Jacob Mchangama, Isabelle Anzabi, Artificial Intelligence and Freedom of Expression in the European Union, in *That Violates My Policy: AI Laws, Chatbots, and the Future of Expression*, The Future of Free Speech, October, 2025, p. 84, [hereinafter "Calvet-Bademunt and others, AI and Freedom of Expression in the EU"].

[74] Although not yet ratified.

[75] See Chart of signatures and ratifications of Treaty 225, avaiable at https://www.coe.int/en/web/Conventions/full-list/?module=signatures-by-treaty&treatynum=225.

[76] Calvet-Bademunt and others, AI and Freedom of Expression in the EU, p. 87.

[77] Ibid.

[78] Digital Services Act, ¶¶80-92.

[79] Calvet-Bademunt and others, AI and Freedom of Expression in the EU, p. 105.

[80] Ibid, p. 84.

[81] Ibid.

[82] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 2018, Section III. A. ¶¶15-18.

transforming platforms into carriers of propaganda well beyond established areas of legitimate concern."[83]

Another issue that is emerginmg is the liability for AI-generated content. The Eruopean Commsion proposed an AI Liability Directive, as a complement to the AI Act, however, it later withdrew it.[84] Adopting this Directive "could have raised concerns about the freedom of expression, as it might have incentivized companies to preemtively censor their models to avoid liability."[85] However, it updated its Product Liability Directive[86] to classify software and AI as "products" which would hold providers and developers liable for defects. This Directive complements the AI Act as it ensures that "individuals harmed by AI technologies have clear legal avenues for compensation," still, it is of crucial importance that "liability rules are interpreted in a manner that does not chill access to legitimate information by prompting companies to withhold content out of liability concerns."[87]

These examples point to an emerging practice of taking aside the requirements of necessity and proportionality when restricting speech and building a "framework where lawful expression is filtered out preemtively, not because it violates the law, but because it is safer for intermediaries to take it down to avoid legal and reputational risks."[88] This places the EU before a crucial crossroad: will the future AI Regulation lead the way to a strong precedent in safeguarding the freedom of expresson, or will it practical implementation gradually narrow its scope of application?

VIII. AI-DRIVEN SYSTEMS AND FREEDOM OF EXPRESSION IN THE UNITED STATES OF AMERICA (USA)

The legal framework of protecting the freedom of expresion in the USA is embedded in the First Amendment of the US Constitution: "Congress shall make no law […] abridging the freedom of speech, or of the press."[89] Practice of the Supreme Court shows a broad interpreation of this right extending it to "new communication technologies and safeguarding both the right to speak and the right to receive information and ideas."[90] As the USA relies on rapid private sector innovation and AI development, courts in the USA face the question whether AI-generated content shall recive the protection of the First Amendment and the constitutional limits on regulating content that is illegal, harmful or deceptive.

The US has signed[91] the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.[92] However, on a national level, it has not yet adopted a comprehensive federal framework that regulates the use of AI.

---

[83] Ibid, Section III. A. ¶¶20, 21.
[84] Avaiable at https://iapp.org/news/a/european-commission-withdraws-ai-liability-directive-from-consideration.
[85] Calvet-Bademunt and others, AI and Freedom of Expression in the EU, p. 88.
[86] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, Effective December 2026.
[87] Calvet-Bademunt and others, AI and Freedom of Expression in the EU, p. 88.
[88] Ibid, p. 105.
[89] US Constitution, 1st Amendement.
[90] Isabelle Anzabi, Jordi Calvet-Bademunt, Jacob Mchangama, Artificial Intelligence and Freedom of Expression in the United States, in *That Violates My Policy: AI Laws, Chatbots, and the Future of Expression*, The Future of Free Speech, October, 2025, p. 59, [hereinafter "Anzabi and others"]; Brown et al. v. Entertainment Merchants Assn. et al., 564 U.S., 786 (2011).
[91] Although not yet ratified.
[92] See Chart of signatures and ratifications of Treaty 225, avaiable at https://www.coe.int/en/web/Conventions/full-list/?module=signatures-by-treaty&treatynum=225.

As of 2025, AI governance is regulated through America's AI Action Plan,[93] which was introduced with the aim to articulate a "deregulatory philosophy rooted in global competitivness and national sovereignty."[94] The Action Plan aims to, *inter alia,* ensure that AI systems are built "from the ground up with freedom of speech and expression in mind, and that U.S. government policy does not interfere with that objective," as well as to ensure that "AI procured by the Federal government objectively reflects truth rather than social engineering agendas."[95] In the absense of a federal AI regulation, all 50 States have taken legislative steps to shape the legal and normative dimensions of the use of AI.[96] Despite steps to (de)regulate, several questions on the use of AI-driven systems and how they affect freedom of expression remain debated at a national level in the USA.

One of the ongoing legal debates is whether AI-generated content constitutes speech under the First Amendment.[97] On one hand, legal scholars argue that AI-generated content must be protected and the focus must be on the listener's right to receive information, regardless of whether the source of that information is human or artificial, thus, comparing AI models to the press or cameras as tools for creating expressing content.[98] Thus, "users have a right to obtain information from the AI models."[99] On the other hand, some scholars argue that AI-generated content is not "inherently expressive" or that it lacks "the human intentionality" that is connected to free speech rights.[100] In other words, AI-driven models do not "speak" in a way as to award constitutional protection, but rather produce automated results based on training and algorithm data.[101]

Another important matter is the unsettled legal framework regarding liability for AI-generated content, and the question arises on who is to be held legally responsible when and AI system generates harmful or unlawful speech? Traditionally, when a person publishes a false statement of fact about another person, that causes harm on the reputation, and actual malice or reckless disregard for the truth is established, liability arises.[102] In the context of AI-generated speech, "a user who knowingly promts an AI system to generate and then publicly shares false and injurious statement could be held liable" under the liablility rules.[103] However, a complication occurs when the harmful speech is generated autonomously from the AI system. In this scenario, on one hand, the user intent to defame is absent, and, on the other hand, the AI system lacks "the mental state or fault traditionally" required for liability.[104] Hence, another question arises: to what point and under what circumstances AI developers and deployers could be held liable for the content their systems generate? Additionally, these models often produce fabricated information without intent or factual grounding. US Courts have dealt with similar cases and hold the opinion that "given the growing public awareness that AI output may be unrealiable or speculative […], such statements [are] less likely to be interpreted by a

[93] The whole text of the AI Action Plan is avaiable at: https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[94] Anzabi and others, p. 62.
[95] AI Action Plan, p. 4.
[96] Anzabi and others, pp. 63, 64.
[97] Anzabi and others, p. 61.
[98] Jane R. Yakowitz Bambauer, Negligent AI Speech: Some Thoughts About Duty, Journal of Free Speech Law, April 28, 2023; Toni Marie Massaro, Helen L. Norton, and Margot E. Kaminski, "SIRI-OUSLY 2.0: What Artificial Intelligence Reveals about the First Amendment," Minnesota Law Review 101, 2017, p. 2481; Volokh, Lemley, and Henderson, Freedom of Speech and AI Output, p. 658.
[99] Anzabi and others, p. 61.
[100] Peter Salib, AI Outputs Are Not Protected Speech, Washington University Law Review, University of Houston Law Center Research Paper no. 2024-A-5, January 1, 2024; Anzabi and others, p. 61.
[101] Anzabi and others, p. 61.
[102] Anzabi and others, p. 66.
[103] Ibid.
[104] Ibid.

reasonable person as factual assertations",[105] as reflected in the *Walter v. OpenAI* case.[106] Worth to mention is also the case of *Robby Starbuck v. Meta*, which alleged that Meta's AI platform produced defamatory and false statements about Starbuck in response to user promts.[107] The case was settled with the aim of Starbuck working with Meta to address ideological and political bias in its AI.[108]

The third major field being discussed is hate speech. Namely, the US Constitution has one of the most robust protections on the freedom of expression, as US has no general statutory prohibition on hate speech, with many cases where the US Supreme Court has rejected government efforts to restrict speech based on hateful or offencive nature.[109] In that sense, AI-generated content would likely be protected, unless its falls in the narrow categories of unprotected speech,[110] such as incitment to imminent lawless action,[111] true threats,[112] or obscenity.[113] However, AI developers and platforms, as not bound by the First Amendment, could design and enforce their own moderation policies with the aim to filter out hate speech or other forms of content that is deemed offencive.[114] Detecting hate speech in such circumstances might prove to be difficult, as "definitions differ over which groups are protected", or as the assessment which "speech is merely offensive, satirical or part of legitimate discussion" is highly contextual.[115] In that sense, summarizing historical writings or political rethoric that contains offensive language might be unpleasant, but it could serve an educational or research purpose in certsin context.[116]

It could be concluded that the relationship between AI and the freedom of expression in the US continously evolves as the use of AI on the online plaforms continues to develop and progress.[117] Although the principles embedded in the First Ammendment would likely extend to the speech generated by AI, still, how they will be applied in the context of AI still remains to be developed and witnessed.

IX. AI-DRIVEN SYSTEMS AND FREEDOM OF EXPRESSION IN NORTH MACEDONIA

North Macedonia lacks regulation that tackle AI-driven systems despite initiated efforts to create a National Strategy for AI. The reason for the slow legal development in the field of AI in the country could be found in several challenges such as insufficient data, human resources, and tehcnical capabilities.[118] As a country that aims to become a member State of the EU, North Macedonia is faced to develop a comprehensive strategy aligned with EU standards. However,

---

[105] Ibid, p. 67.
[106] More informationa avaiable at: https://www.loeb.com/en/insights/publications/2025/05/walters-v-openai-llc.
[107] Sarah Nassauer and Jacob Gershman, "Activist Robby Starbuck Sues Meta Over AI Answers About Him," Wall Street Journal, April 29, 2025.
[108] Joseph De Avila, "Meta, Robby Starbuck Settle AI Defamation Lawsuit," Wall Street Journal, August 8, 2025.
[109] Anzabi and others, p. 72.
[110] Ibid.
[111] Brandenburg v. Ohio, 395 U.S. 66 (1969).
[112] Counterman v. Colorado, 600 U.S. 66 (2023).
[113] Miller v. California, 413 U.S. 15 (1973).
[114] Anzabi and others, p. 72.
[115] Ibid.
[116] Ibid.
[117] Ibid, p. 81.
[118] Andrea Radonjanin, Andrea Lazarevska,The Status and Future Prospects of AI Regulation and Development in North Macedonia, avaiable at https://ceelegalmatters.com/briefings/27312-the-status-and-future-prospects-of-ai-regulation-and-development-in-north-macedonia.

it faces challenges on developing "the necessary infrastructure for AI education, fostering innovation and adopting a National Strategy along with ethical guidelines" and risks "falling behind in AI advancement, despite its strong IT sector and investment appeal".[119] Drafting AI-related legal framework must address the human rights obligations and its potential risks to the freedom of expression.

X. CONCLUSION

The Internet and online platforms have quickly become a central forum for exchanging ideas and opinions, sharing news and updates, and connecting individuals from different parts of the world, and as such, it is today considered a leading prerequisite for the enjoyment of freedom of expression. As such, the online space has a profound value for freedom of thought and expression because it can disseminate information very quickly and make it accessible to anyone who has access to it.

In terms of the countries assessed, USA could be acknowledged as the most speech-protective country in relation to generative AI, with the First Amendment providing strong protections. Similarly, the European Convention on Human Rights and the EU Charter of Fundamental Rights ascertain solid protections on the right to freedom of expression, however, broad hate speech rules and poorly defined "systemic risk assessments" should be properly addressed to avoid disproportionate speech restrictions. North Macedonia must commence the assessment and drafting of legal frameworks that balance global needs and provide effective protection of human rights and alignment with international human rights standards.

It is a fact that the use of artificial intelligence systems affect the freedom of expression and irresponsible use could pose a major risk to its enjoyment by individuals. Therefore, it is crucial to cautiously establish a framework for harmonizing the use of artificial intelligence with the principles of international human rights law and to encourage states to reflect on their obligations and responsibilities for the protection of human rights in the digital world, as well as to take measures to encourage and control private companies to respect human rights in an era in which "the power, reach and scope of artificial intelligence technology are growing."[120] In doing so, it must not be forgotten that the obligation to respect freedom of expression is binding on every state. All authorities (executive, legislative and judicial), as well as all other state bodies or private actors exercising public authority, at local and national levels, have a responsibility to protect human rights, and that responsibility includes protecting every individual from any actions by private persons or entities that would violate their full enjoyment of freedom of expression.

[119] Ibid.
[120] Kaye, 2018, ¶61.