Artificial Intelligence and Hate Speech: Exposing Facebook's Biased Algorithms and Threats to the Freedom of Expression

Katsiaryna Bareika

Submitted in partial fulfillment of the requirements for the degree of MA

Erasmus Mundus Human Rights Practice and Policy Masters Programme

School of Global Studies, University of Gothenburg

Pedro Arrupe Human Rights Institute, Deusto University

School of Humanities and Social Sciences, University of Roehampton

Department of Social Sciences, University of Tromsø – Arctic University of Norway

22.05.2023

Dissertation Module (30 ECT)

Supervisor: Dolores Morondo

Spring semester 2023

Abstract

Despite the great variety of technologies available today, our attention is mostly attracted to those that cover one of our primary needs: communication. For this reason, the Internet and Internet-related technologies have become an integral part of our daily existence. The foundation of our online presence is social media and our interaction therein, which includes obtaining information, personal communication, and public expression. Thus, when using Internet technologies, we exercise our fundamental right - the right to freedom of expression.

Since the Internet came into our everyday life, the right to freedom of expression and its perception by law and society has changed significantly. Today social networks are the leading platforms for expressing opinions publicly. However, publicity on the Internet differs considerably from the generally accepted concept of publicity: when speaking any opinion or idea online, the user feels neither direct danger nor possible consequences for expressing his thoughts. Thus, Hate Speech online was created as a particular category of public statements that require special legal regulation. Currently, the biggest platforms for social interaction on the Internet (social networks) are actively cooperating with human rights institutions to comply with the right to freedom of expression. They engage in dialogue with governments, constantly update their anti-hate speech policies based on legal requirements, introduce their own hateful content tracking mechanisms and insist on following international recommendations for identifying online hate speech.

However, some current research shows that the activism of such companies to protect the right to freedom of expression has a massive dark side. Social networks use the soft nature of human rights regulations and the lack of clear legal definitions to violate the right to freedom of expression in their favour. They promote hate and drive toxic and harmful content for their own benefit, make deals and go for concessions with the governments, and close their eyes to the holes and failures of their own automotive systems. Therefore, this kind of conduct demonstrates an "unhealthy relationship" between human rights and Big Tech and reveals significant socio-legal gaps that have become the main focus of this study.

Word count: 17419

Keywords*: hate speech, freedom of expression, social media, biased algorithms, hate speech identification, hate speech criteria, legal standards.*

Acknowledgements

To my strong, beautiful mother and grandmother who dedicated their lives to making mine better.
I'm so grateful to you, and you are my real-life heroes.

To my dearest friends who became my wings through these two hardest years.
I love you, and I miss u.

To my past, present and future self.
I'm so proud of you.

List of abbreviations

AI – Artificial Intelligence

BSR – Business for Social Responsibility

CEO – Chief Executive Officer

CoE – Council of Europe

ECHR – European Court of Human Rights

ICCPR – International Covenant on Civil and Political Rights

IT – Information Technology

MIT – Massachusetts Institute of Technology

MSI – Meaningful Social Interactions

SEC – The United States Securities and Exchange Commission

UN – United Nations

UDHR – Universal Declaration of Human Rights

Table of Contents

Chapter 1. Introduction

Online hate speech is a complex, multifaceted, and at the same time, extremely vulnerable concept, which can be easily and invisibly neglected by unscrupulous IT companies. For this reason, any misinterpretation or misuse of the concept constitutes a massive threat to the freedom of expression. Internet giants such as Facebook can easily use the soft nature of human rights regulations and the lack of clear legal definitions of Hate Speech to violate the right to freedom of expression in their favour. This paper is going to focus on a new loophole that is hard to notice for those who are not related to big data and engineering sciences – artificial intelligence (AI).

Biased algorithms of artificial intelligence used as hate speech detection mechanisms are a newly raised issue. From time to time, the evidence of intentionally biased AI used by such Big Tech players as Google, Twitter, Meta and others was found but created no particular resonance. The problem hit the wide public differently in 2021 when numerous high-profile media sources published evidence of intentionally biased detection mechanisms used by Facebook. As follows, big data studies started to appear to demonstrate that unscrupulous AI programming creates multiple ways to achieve certain goals beneficial for the company's business and allows them to keep such bias unnoticed by human rights monitoring mechanisms.

Despite the novelty of the topic, we can find various studies related to AI in the context of human rights and hate speech. However, due to the low level of Facebook's transparency, the work of its artificial intelligence tools lacks studies from various disciplines. As long as Facebook remains the world's most popular social media network, it is vital to understand the actual challenges, consequences and threats this company brings to one of the fundamental human rights – freedom of expression. For this reason, this research aims to finish the following debate: "Are Facebook's algorithms aimed to counter online hate speech intentionally biased and create a threat to the freedom of expression?".

Chapter 2. Literature Review

Today, special attention must be paid to the direct technical development of society. We appreciate the technical benefits that make our lives much easier, and every day we find new ways to use them. One of these ways has been to train artificial intelligence to detect and remove hate speech from the Internet space. This research aims to prove the hypothesis that Facebook's artificial intelligence algorithms for detecting hateful content are invisibly biased and threaten the right to free expression. Such a study is rather ambitious and challenging to implement. Due to the novelty of the topic and the constantly emerging evidence, we can now observe a need for studies explicitly aimed at assessing the impact of Facebook algorithms on human rights, especially on the right to free expression. There is plenty of research on hate speech theory and practice within the academic space, including online hate speech studies. However, these studies focus not on companies' intentionally wrongful actions but on categories and types of hate speech.

Facebook's lack of transparency is forcing researchers from various fields to scrape together hard evidence that the algorithmic abuses are intentional and, thus, a severe threat to the right to freedom of expression. Despite the use of various legal and social sources, the proof of the hypothesis will be based on media case studies and big data scientific evidence rather than on socio-legal human rights studies. Therefore, this literature review aims to explain the choice of sources to achieve the research goals and to prove the bias of Facebook's artificial intelligence algorithms in identifying hateful content. To clarify, as a reference point to "artificial intelligence" or "algorithms" in the context of Internet platforms, we use the definition of Tufekci (2015), referring to such technology as "computational processes that are used to make decisions of such complexity that inputs and outputs are neither transparent nor obvious to the casual human observer".

Identifying the challenges of dealing with online hate speech requires a detailed examination of international standards for its identification and the right to freedom of expression in general. On the one hand, proponents of a legal approach to human rights (Walker, 1994; Barendt, 2005; Rosenfield, 2005; Mihkailova et al., 2013) base their understanding of hate speech on the fundamental sources of international human rights law (UDHR, 1948; ECHR, 1950; ICCPR, 1950). They argue that the concept of hate speech, first and foremost, derives from the right to freedom of expression, which is directly regulated by international human rights law. On the other

hand, proponents of socio-legal studies point out that this phenomenon is integral to cultural, political, educational and technological development (Brown, 2017; Brown & Sinclair, 2019) and, therefore, also depends on their changes and trends. Nevertheless, the study of legal sources is the focus of this paper. Today's Big Tech companies' wide range of activities, affecting virtually every area of our lives, creates complexities in their legal regulation. The need to take account of the many nuances creates blind spots which attract unscrupulous actors who may take advantage of them for their benefit and may also threaten human rights. International human rights law remains the most effective mechanism for bringing accountability and creating preventive measures.

The other part of the legal approach used here is studying the so-called "soft law" sources. This study focused on two specifics: the identification of categories and criteria for the identification of hate speech and the specifics of the regulation of hate speech in the online space. The main criteria for the identification of hate speech are set by international legal standards based on judicial precedents in the field of human rights law (Handyside v. UK, 1976; Jersild v. Denmark, 1993; Otto-Preminger-Institut v. Austria, 1994; Smajić v. Bosnia and Herzegovina, 2018). Doctrinal and educational sources in this field[1] indicate that the primary task when dealing with hate speech is identifying its type. Such types include speech or expression that must be prohibited, speech or expression that may be prohibited, and permitted expressions (protected speech). The determination of the type of hateful content is, in turn, based on identification criteria. The courts precede three main criteria: intent, content (context), and consequence.

Further, the specifics of this study involved examining the distinctive features of hate speech online. International human rights mechanisms offer many guidelines to help identify such distinctions[2]. Despite their multiplicity, a detailed analysis reveals the following problems. Firstly, due to the constant development of Internet technology, the primary instruments regulating hate speech on the Internet are gradually becoming obsolete and need constant updating. Secondly, an insufficiently broad view of combating hateful online content was noted. What is meant here is that International human rights law emphasises introducing new regulations rather than effective enforcement and respect for human rights. The current development of AI allows Internet

[1] See Article 19 (2015). Hate Speech Explained Toolkit. Free World Centre; Weber, A. (2009). Manual on hate speech. Council of Europe.

[2] See UN Human Rights Council (2012). The promotion, protection and enjoyment of human rights on the Internet, 20/8; Council of Europe (2001). Convention on Cybercrime. European Treaty Series - No. 185; European Commission (2016). The EU Code of conduct on countering illegal hate speech online; European Commission's High-Level Expert Group on Artificial Intelligence (2018). A definition of AI: Main capabilities and scientific disciplines; Committee on Legal Affairs (2017). Report with recommendations to the Commission on Civil Law Rules on Robotics, European Parliament, 2015/2103.

companies to find tricks to circumvent human rights. In this regard, more effort must be made to explore the technical side of the issue to assess threats to human rights.

Of particular interest in using AL to deal with hate speech is the world's largest social network, Facebook, whose published data and reports leave many questions about the real state of the company. Consequently, the practical part of the study is based on a case study of Facebook. In its public materials (company's website, public statements and interviews, guidelines, and terms of use), Facebook emphasises the high effectiveness of AI in combating hate speech while blaming any shortcomings on the limits of technological capabilities. However, the qualitative methodological sources prove otherwise. Credible investigations by major media sources (Time, The Guardian, Financial Review, CBS News, The Intercept, Financial Review and others) have raised fears of the intentional misusing mechanistic technologies by Facebook. The abundant evidence of intentionally biased algorithms, inconsistent internal company policies, and evidence of manipulation of public policy and human rights demonstrate the application of these findings to the problems of combating online HC. At the same time, case studies on the challenges of computational agencies (Tufekci, Z., 2015; Zhang, Z. et al., 2019), statistical surveys (Ntoutsi, E. et al., 2020) on content moderation and data mining, studies on the mechanisms of CS distribution (Vosoughi, S. et al. 2018) have addressed the technical side of the issue and confirmed the existing threat. Finally, an assessment of the reasons for such bias and the desire to violate the right to freedom of expression found reasons such as a commercial gain from viral marketing (Chalermsook P. et al., 2015; Schwarz, O., 2019) and political influence (Vosoughi, S. et al., 2018; Papakyriakopoulos, O. et al., 2020).

The key findings from the reviewed literature resulted in the following.

1. The prevalence of the legal approach to hate speech is justified, and international human rights law remains the main safeguard for protecting our right to free speech and expression. However, current legislation on combating hate speech online has significant gaps, allowing companies like Facebook to neglect the rules with no significant consequences.

2. Current legal mechanisms of Internet hate speech control do not sufficiently address the activities of Big Tech giants from a technical perspective. They call for but do not oblige companies to be fully transparent and, therefore, cannot fully assess the detrimental impact on human rights. The key to proving intentional bias in algorithms lies not in trusting official

information and Facebook reports but in verifying studies and concerns from other related fields. The lack of consideration of technicalities in shaping the legal framework to combat hate speech in cyberspace demonstrates the need for cross-sectional research to understand both the roots and the actual and potential consequences of Facebook's actions.

3. Companies have the substantial benefit of ignoring human rights. Numerous pieces of evidence of intentionally biased algorithms, such as issues with performance, language detection, capacity limits, and overall controversial internal policies of the company, are a consequence of Facebook's two main goals - financial and socio-political influence. Furthermore, a "de jure" and "de facto" comparison of Facebook's anti-hate speech policies and existing international legal standards for identifying hate speech leads to the conclusion that companies primarily base their anti-hate speech efforts on their own criteria, which often contradict international human rights law standards.

All the facts above demonstrate the intimidation of one of the fundamental human rights - the right to freedom of expression. For this reason, this paper will further address the need for a comprehensive socio-legal analysis of Facebook's biased hate speech identification algorithms, identifying blind spots and potential threats and elaborating on possible solutions.

Chapter 3. Scope, Methodology and Outline

The hypothesis of this research is based on the following thesis: the algorithms of hate speech detection used by Big Tech companies are invisibly biased, do not comply with the international legal standards of hate speech identification and constitute a threat to freedom of expression. To support this thesis, we can outline the following research goals:

- to analyse current international legal regulations of freedom of expression and hate speech;
- to define the importance of the legislative approach to human rights in combating hate speech online;
- to outline the specific features of online hate speech identification methods and policies;
- to expose the biased algorithms of Facebook and its violation of freedom of expression;
- to propose solutions for combating online hate speech.

Regarding the methodology, the interdisciplinary character of the hate speech concept requires the use of combined methods of research.

The theoretical part of the paper is based on two research methods - legislative approach and doctrinal research. For the legislative part of the study, we examined both "hard" and "soft" law sources. Major international human rights treaties and conventions such as the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights (ICCPR), the European Convention on Human Rights (ECHR), and the Convention on Cybercrime were used to represent imperative legislation on the freedom of expression. However, the target of the theoretical part demanded an extensive understanding of the hate speech framework, its peculiarities and so-called "underwater stones". Thus, such soft law sources as general comments, additional protocols, recommendations, reports, and case studies emphasised this demand.

The qualitative method was used for the factual part of the research. The methods used for findings and their analysis are the following:

- a case study of the company Facebook, based on its policies and technical challenges (examination of the company's website, public statements and interviews, guidelines and terms of use), aimed to show the official position of the company on countering online hate speech;

- media research of the works of mass social media (investigations, interviews, officially published documents, leaked documents) aimed to expose the actual state of affairs inside the company and find human rights-related issues and controversies;

- research of doctrinal studies in the field of big data and engineering sciences aimed to understand the technical side of online hate speech detection algorithms and outline their strengths and weaknesses;

- analysis of socio-legal doctrinal sources aimed to understand the primary reasons for the existing algorithm bias.

Lastly, the findings from the previous parts were considered from the prism of relationships between Big Tech companies, their users and human rights law. Since the investigation of biased algorithms aims not only to prove the violation of the right to freedom of expression by Facebook but to evaluate the actual and potential consequences of such violations on society and human behaviour and to discuss possible loopholes and solutions, the socio-legal method was chosen for this purpose.

The dissertation is written in four chapters organised to develop the topic coherently. Each following part focuses on the aforementioned contributions.

The first chapter is based on the theoretical explanation of freedom of expression and hate speech in the international human rights legislation framework. This part provides a major understanding of the hate speech concept and explains the two main concepts essential for combating hate speech – hate speech typology and criteria of identification.

The second chapter concentrates on the specifics of online hate speech. Beginning with discussing the historical and legislative backgrounds of freedom of expression on the Internet, we will further explore Big Tech companies and their role as intersections between the state and the individual. Simultaneously, using the Facebook example, we uncover the technical background of hate speech identification algorithms and get closer to the most controversial part of the study – secretly biased algorithms.

The third chapter investigates the real situation with Facebook and its hate speech policy. This chapter reveals intentionally biased algorithms, their weak performance, and non-obvious problems of Facebook hate speech policies. Here we discuss the reasons for such bias and its impact on the users and elaborate on the actual and potential threat to freedom of expression.

In the last chapter, we take a socio-legal approach to the aforementioned blind spots, debating the importance of the legal approach to freedom of expression and discussing existing and potential solutions and improvements for online hate speech regulations.

Chapter 4. Hate Speech in International Human Rights Law

4.1 The Right to Freedom of Expression

We often encounter the concept of "hate speech" on the Internet. However, only some of the average users can give a precise definition of what this concept encompasses. While teaching the topic of hate speech in my 2021 Freedom of Expression course, I found that most of the students also failed to articulate their definitions. When searching for "What is hate speech?" on the Internet, the first result we get is a definition from the official United Nations website, which states that "In common language, 'hate speech' refers to offensive discourse targeting a group, or an individual based on inherent characteristics (such as race, religion or gender) and that may threaten social peace[3]. The definition formulated by the UN mostly derives from the substantial international human rights documents (UDHR, 1948; ECHR, 1950; ICCPR, 1950). At the same time, abstract from legal, social or other sciences, people generally perceive hate speech as any manifestation of negativity (equal to hatred) towards their actions, values, or themselves, which can be attributed to two main reasons.

First of all, the very concept of hate speech is relatively new in human rights-related sciences. The identification of hate speech is a complex area within the framework of the right to freedom of expression and opinion due to the fact that international standards and recommendations to combat it are mainly related to soft law, which means that they have more advisory than mandatory character. Second, such a simplistic perception results from the lack of public education on this issue. Therefore, for further context, the first thing to do is to find out what constitutes hate speech under human rights law and explore the criteria formed by legal practice for identifying and categorising speech.

[3] United Nations. What is hate speech? [online]. Available at: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech?gclid=Cj0KCQjw2v-gBhC1ARIsAOQdKY1-lyuALindV77gVDFVTCAoHjG8Z1Ft-a3TXmyMD-didrWl4cyUAnYaAuS4EALw_wcB [Accessed 13 September 2022].

Before going straight to the legal analysis, we have to address the intersectional character of the hate speech concept. Some scholars prefer a solid legal approach to human rights and thus perceive hate speech as a subject primarily related to freedom of expression[4]. However, our understanding of hate speech does not always correspond to what is implied in the human rights framework. According to Brown (2017), it is better to understand hate speech not only through the legal concept but take into consideration "a range of extralegal measures including counter-speech and education[5]". Together with that, Brown and Sinclair (2019) in their studies outline a range of other domains, such as the social, cultural, political and even technical character of hate speech[6]. So why is it necessary to rely on a legal approach in the first place?

Hate speech and everything related to this concept originates in one of the fundamental human rights necessary in a progressive democratic society: the right to freedom of expression. The high importance of this right is confirmed by its inclusion in most international legal documents in the field of human rights law[7]. The legal standard in this context includes the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or print, in the form of art, or through any other media of one's choice[8]. The right to freedom of expression is a prerequisite for the principles of transparency and accountability, which, in turn, are indispensable for promoting and protecting human rights[9]. Nevertheless, we must bear in mind that the exercise of the rights, as mentioned earlier, carries with it special duties and responsibilities. As such, the right to freedom of speech and expression cannot be exclusive and absolute because of the many factors that directly or indirectly affect its operation and exercise. That is why the right to freedom of Speech and expression is subject to certain restrictions, which,

[4] See Walker, S. (1994). Hate speech: The history of an American controversy. U of Nebraska Press, pp. 101 - 127; Barendt, E. (2005). Freedom of speech. OUP Oxford, pp. 39 – 73; Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: a comparative analysis. Cardozo L. Rev., 24, 1523, pp. 1525 - 1529; Mihkailova, E., Bacovska, J. and Shekerdjiev, T. (2013). Freedom of expression and hate speech. [online] Skopje: OBSE, p.6.
[5] Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. Law and Philosophy Vol. 36 No. 4. Springer, p.422.
[6] Brown, A., & Sinclair, A. (2019). The politics of hate speech laws. Routledge, p.3. See also Weinstein, J., & Hare, I. (2009). General Introduction: Free Speech, Democracy, and the Suppression of Extreme Speech Past and Present in Extreme Speech and Democracy. OUP Oxford.
[7] See United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December), Article 19 (2); UN General Assembly (1948). Universal Declaration of Human Rights, 217 A (III), Article 19; Council of Europe (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, ETS 5, Article 10.
[8] United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December), Article 19 (2).
[9] UN Human Rights Committee (2011). General comment no. 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34, para. 3.

however, must be established by law and be necessary both for respect of the rights and reputations of others and for the protection of national security, public order, public health or morals[10].

Freedom of expression applies to all persons and means that everyone is free to express their thoughts, ideas and opinions in any way without any infringement, censorship or other state interference, subject only to a narrow range of exceptions as defined in the articles of an international legal instrument. This right applies not only to individuals but also to organisations, media corporations and campaigns. Such a right provides protection against any kind of undue interference, including infringement and censorship, not only to individual citizens but also to specific groups of people acting together both in working and in public to achieve their goals, which is essential in the context of the inextricable link between free expression and the functioning of organisations, corporations and other companies operating both offline and on the Internet.

4.2 Criteria of Identification

The lack of a definition of hate speech generally accepted under international human rights law results in multiple problems regarding its identification. Although, at first glance, the identification of hate speech is a simple task, practice shows that a superficial glance at a statement is insufficient to determine its legitimacy or illegitimacy. At the moment, several criteria can be used to determine whether a particular utterance falls under the concept of hate speech, but such measures can be elusive or contradictory. To formulate the typology of Hate Speech and the criteria for its identification, we will further use the Hate Speech Manual, published with the support of the Council of Europe[11], together with the precedent rulings of the European Court of Human Rights (ECHR).

In general, opinions on what constitutes "incitement to hatred" vary widely. In simplistic terms, hate speech is any expression of discriminatory hatred against people that do not necessarily entail inevitable consequences. This most simplistic definition reflects an overly broad range of expressions, including expressions whose restriction is not legitimate. In order to determine what

[10] United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December), Article 19 (3).
[11] Weber, A. (2009). Manual on hate speech. [online] Strasbourg: Council of Europe. Available at: http://icm.sk/subory/Manual_on_hate_speech.pdf [Accessed 29 September 2022].

type of expression is of a type, courts first use the standard criteria established by the Rabat Plan of Action[12], viz:

- The context and form of the expression;
- The status of the speaker and the degree of their influence on the audience;
- The tone and content of the expression;
- Intent to incite hatred;
- Audience coverage;
- Probability and imminence of harm.

Based on International Human Rights case law, the courts give precedence to three criteria: intent, content (context), and consequence[13].

1. Intent to spread hatred

As mentioned in the definition, hate speech is defined as speech behind which there is an intention to incite, promote or justify hatred of persons associated with a particular group (not only religious but also including racial, ethnic, LGBT+ community and other kinds). The ECHR first applied such an interpretation in Jersild v. Denmark (1994) and afterwards included this element in the main characteristics of prohibited hate speech.

2. The context of the specific statement

For assessing the legitimacy of a particular expression or opinion, international law considers not only what is expressed but also the specific circumstances of how the statement was used. One commonly used example of such a context is an individual's profession. For example, statements made by a politician, a journalist, an artist, and an ordinary citizen will be evaluated differently by both competent judges and society. The circumstances, time, place and other factors also play a significant role in determining the legitimacy of an expression.

[12] UN Human Rights Council (2013). Annual report of the United Nations High Commissioner for Human Rights: Addendum, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4.
[13] Mihkailova, E., Bacovska, J. and Shekerdjiev, T. (2013). Freedom of expression and hate speech. [online] Skopje: OBSE, p. 26. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 01 November 2022].

To provide a real example, we will examine the ECHR judgment in the case of Smajic v. Bosnia and Herzegovina (2018), where Bosnian lawyer Abedin Smajic was arrested "for inciting national, racial and religious hatred, discord or intolerance on the Internet[14]". Here hate speech refers to the applicant's call for action in case of war and subsequent secession of the Republika Srpska, published on the popular website "Bosnahistorija".[15] The applicant used a pseudonym to publish his views. The first instance court held that "such online publications in a publicly accessible place, such as an appropriate website, can impact relations between different ethnic groups, particularly Bosniaks, Croats and Serbs[16]".

In the judgment, the ECHR stated that "freedom of expression is one of the essential foundations of a democratic society and one of the basic conditions for its progress and for everyone's self-expression[17]", and this applies both to expressions that are "well received or regarded as innocuous, but also to those that may offend, shock or disturb[18]". In analysing the content of the statements published by the applicant, the ECHR first noted that the statements addressed "the sensitive issue of ethnic relations in post-conflict Bosnian society[19]". Since the domestic courts had clearly demonstrated a substantial user response to the opinions published on the Internet and had analysed the possible consequences of such statements, the ECHR held that restricting freedom of expression was legitimate. Secondly, the court noted that the statements had been published on a thematic portal. Thus, the audience of such an internet site was people highly involved in the subject matter. If sufficiently reasoned, the applicant's opinion may have substantially influenced the opinion of other users and thus theoretically led to disruptive consequences. Based on these findings, the ECHR found the sanctions imposed on the applicant proportionate to the aim of protecting the reputation and rights of others.

We can see that here the court had a duty to evaluate the geopolitical situation in the country to determine the vulnerability of the audience to whom the statements were directed and whether the criteria for restricting freedom of expression were met. Therefore, this case once again

[14] Smajić v. Bosnia and Herzegovina (2018). Council of Europe: European Court of Human Rights, 48657/16, para. 2 - 12.
[15] Ibid., para. 2.
[16] Ibid., para. 7.
[17] Ibid.
[18] Ibid., para. 3.
[19] Ibid., para. 9.

demonstrates the importance of examining the context when assessing the lawfulness of state action.

3. Consequences of Hate Speech

In addition to insulting the dignity of the person/s to whom they are directed, as part of the decision in Otto-Preminger-Institute v. Austria (1994), the ECHR acknowledged that hate speech in its nature could disrupt public order and incite violence. Society's response to such a statement is both immediate outbreaks of incidents and the incitement of violence between the groups to whom the message was directed. However, the notion of result here encompasses all the socially critical consequences caused by such expression, even if no actual acts are causing more severe consequences. It is not always possible to determine the social reaction to the provocative expression. If the court anticipates the possible weighty consequences for society, such speech will qualify as prohibited hate speech.

Based on the above factors, the elements of the illegality of hate speech can include[20]:

- Hate as a consequence of an intense irrational negative emotion toward an individual or group that takes various forms (written, non-verbal, visual, artistic, and others) and extends to an external audience;

- The presence of defensive provisions for an individual or group that are the target of hate speech (race, gender, religious affiliation, or others);

- The degree of focus on the content and tone of expression, namely the definition of the harmfulness of the statement. "The destructive nature of certain types of freedom of speech and expression is an essential criterion for determining the legitimacy of the restriction of a particular message. A clear distinction must be made as to whether the discriminatory, dehumanising or degrading expression is potentially or actually harmful or whether such expression is an actual harmful consequence expressed in an emotional response[21]";

[20] Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre, p. 28. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].
[21] Ibid.

- The need to establish a causal relationship between the expression and the potential or actual harm: the need for any harm to be probable and imminent;

- The public dissemination of harm associated with targeting a particular audience. In order to exercise a legitimate restriction on expression, such expression must be «exclusively public and affect not a specific person, but a particular protected social group[22]».

Thus, using the factors mentioned above is functional for the courts to determine the direction of the speech analysis when considering cases under the restriction of the right to freedom of expression. Thanks to the above analysis, we have an understanding of which criteria are "benchmarks" for Internet companies in the context of dealing with Hate Speech online. However, another is that human speech or any other mode of expression is a unique, highly flexible and constantly changing tool. The plain distinction between hate speech and non-Hate Speech is insufficient to strike a balance between the fight against hate speech and the right to free expression. For this reason, in the next part of the research, we need to consider the typology of hate speech.

4.3 Hate Speech Typology

When determining the legitimacy of restrictions on freedom of speech within the right to freedom of expression, speech and opinion, it is necessary to consider that not every hostile or aggressive statement should be subject to restriction. Even though the definition of the legitimacy of a particular message is purely case-by-case, it is possible to present a so-called Hate Speech typology. Such a typology is necessary to clarify the different categories of statements, which in one way or another, fall under the concept of hate speech, but are perceived differently by judicial authorities when considering cases due to their specific characteristics. According to the typology presented in the guide to defining and explaining hate speech in the context of Article 19 of the International Covenant on Civil and Political Rights, any expression of hate speech can be divided into three categories. These categories are[23]:

[22] Ibid.
[23] Ibid.

- speech or expression that must be prohibited;
- speech or expression that may be prohibited;
- permitted expressions (free speech).

4.3.1 The first type of hate speech includes those that should be unconditionally prohibited. International human rights law and international criminal law require states to restrict some of the most severe forms of hate speech through criminal, civil and administrative measures[24]. Such prohibitions are designed to prevent the irreversible and exceptional harm that the speaker can or intends to cause.

Direct and public incitement to genocide, although not expressly prohibited in instruments such as the Convention on the Prevention and Punishment of the Crime of Genocide and the Rome Statute of the International Criminal Court, should be considered in the context of international human rights law as speech that should be prohibited. In addition, any advocacy of discriminatory hatred that constitutes incitement to discrimination, hostility and violence is also considered in the context of hate speech, which should be prohibited[25]. It should be noted that at the same time, restrictions on discriminatory hate speech should be used only in cases where the expression jeopardises the preservation of respect for the rights and reputations of others, national security, public order, public health or morals[26].

Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination calls upon States to "condemn all propaganda and all organisations based on ideas of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination[27]".

[24] United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December), Article 20 (2).
[25] UN Human Rights Council (2013). Annual report of the United Nations High Commissioner for Human Rights: Addendum, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4
[26] United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December), Article 19 (3).
[27] United Nations General Assembly (1965). International Convention on the Elimination of All Forms of Racial Discrimination. Treaty Series, vol. 660, Article 5.

In addition, in 2015, the Committee on the Elimination of Racial Discrimination adopted General Recommendation No. 35 on "Combating Racist Hate Speech," where the Committee clarifies the scope of the provisions as mentioned earlier precisely in the context of protecting the right to freedom of expression. This recommendation considers that the Convention carries broader positive obligations of member states to implement the prohibition of specific types of speech. Thus, it is possible to note the vast ground for implementing lawful restrictions on the most severe manifestations of hate speech secured by international treaties.

4.3.2 The next most important category of hate speech is hate speech, which may be prohibited. International human rights law grants states the right to limit freedom of expression in exceptional circumstances and subject to certain conditions. For example, the Human Rights Committee recommends that states comply with the so-called "three-tier test"[28]. This test implies that the following conditions must be met for a State to implement such restrictions:

1. the restrictions must be provided for by law;
2. the restrictions must be implemented for a legitimate purpose (respect for the rights and reputation of others, protection of national security, public order, and others);
Restrictions must be necessary for society.

At the same time, human rights law allows restrictions on freedom of expression to be imposed for any of the following specific purposes[29]:

- in the interest of national security or public safety;
- to prevent disorder or crime (e.g. to prohibit incitement to violence against certain groups);
- to protect health or morals;
- to protect the reputation or rights of others;
- to prevent the disclosure of information received in confidence;
- to maintain the authority and impartiality of the judiciary.

[28] Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre, p. 22. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].
[29] Ibid.

These goals constitute a kind of benchmark for the courts in cases where the admissibility of offensive speech is questioned. However, it is worth noting that even if the restriction meets one of the above objectives, it must also be shown that the limitation is legitimate and necessary in a democratic society[30]. For example, the ECHR has interpreted this to mean that in every case where there is an interference with freedom of expression, a balance must be struck between the individual's right to express himself and the broader public interest justifying the interference (e.g. protection of the rights of other groups). In particular, the restriction must be proportionate to the legitimate aim that the state seeks to uphold through its prohibition. If measures and actions aimed at ensuring respect for the religious beliefs of others serve the purposes of "protecting the rights and freedoms of others" and "protecting public order," they may justify restrictions on the right to freedom of expression[31].

In addition to the above, some forms of hate speech can be understood as explicitly targeting an identifiable victim[32]. This type of hate speech happens when an individual does not seek to incite others to act against a group based on protected characteristics (such as race, gender, religion and others), but his speech is directed solely at another individual. Such utterances include threats of violence, harassment and assault. However, this type of hate speech does not fall within the purview of international legal standards protecting the right to freedom of expression but must be covered by national laws.

Based on the above, it can be concluded that in order to restrict speech which, on the face of it, should certainly be limited by the state, such speech or expression should:

1. Comply with the conditions for restricting freedom of speech, expression and opinion set out in international human rights law;
2. Be directed at a group of persons who share the same characteristics as the target of the speech or at a person belonging to such a group.

[30] Equality and Human Rights Commission (2015). Freedom of Expression. [online] ISBN 978-1-84206-595-2, p.3. Available at: https://www.equalityhumanrights.com/sites/default/files/20150318_foe_legal_framework_guidance_revised_final.pdf [Accessed 20 November 2022].
[31] Ibid., p. 7.
[32] Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre, p. 22. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].

4.3.3 The third type is what is known as legitimate hate speech[33]. When we speak of speech protected by the laws, we refer to those manifestations of freedom of expression, speech and opinion where a such expression may be offensive or provocative but does not meet any of the legitimacy criteria for restricting freedom of expression described above. In most cases, this type of speech is characterised by personal prejudice and supported by expressions of intolerance, but at the same time does not meet the so-called "Severity Threshold"[34]. The most common types of legitimate hate speech include deeply offensive language, blasphemy or "defamation" of religions, denial of historical events and defending the state and public officials.

Deeply offensive expression

It may be surprising, but international standards on freedom of expression protect speech that is offensive, disturbing or shocking and do not allow limitations that are based solely on the speech being addressed to an individual or group[35]. International human rights law does not give individuals the right to be free from such speech, but it certainly protects the right of the people to whom such speech is addressed to oppose it and to speak out against supporters of hate speech. Nevertheless, in practice, states apply sanctions to so-called offensive speech, often varying the degree of offence as the basis for using the restriction[36]. In most cases, state prohibitions on hate speech lack the precision and clarity necessary for the public to regulate their conduct in accordance with the law.

The main difficulty here is that each case of hate speech has to be considered strictly on an individual basis. The criteria for identifying unlawful hate speech mentioned above are often elusive or contradictory, so courts look not only at the expression itself but also at its context. In order to distinguish between unlawful hate speech and profoundly offensive expression, it should be understood that hate speech is inherently any expression of discriminatory hatred towards people that do not necessarily entail inevitable consequences. This most simplistic definition reflects an overly broad range of expressions, including those kinds whose restriction is not legitimate. Not every offensive or deeply offensive speech may qualify as 'incitement to hatred'. The reason for that is the difficulty in drawing the line between the expression of intolerant,

[33] Ibid.
[34] Ibid, p. 67. 1
[35] Handyside v. UK (1976). Council of Europe: European Court of Human Rights, 5493/72, para. 11
[36] Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre, p. 67. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].

offensive or toxic views (which are covered by the right to freedom of expression) and hate speech and other highly offensive communication so severe that it cannot be protected under international human rights law. International human rights law, therefore, considers it necessary to abolish provisions in state law restricting offensive speech, even if such speech is discriminatory.

Blasphemy or "defamation" of religions

Freedom of thought, conscience and religion is one of the foundations of a democratic society according to the content of key instruments of international human rights law. For religious communities, such a right is one of the essential elements that make up the identity and positioning of believers as members of society. This freedom entails, in particular, not only the freedom to practice religion but also the freedom not to adhere to religious beliefs. Such an interpretation of this legal norm is valuable for social groups such as atheists, agnostics and sceptics, as well as for individuals indifferent to religion[37]. Pluralism is inseparable from a democratic society in which several religions or branches of the same religion exist together. For example, according to the ECHR position in Kokkinakis v. Greece (1993), this freedom entails the freedom to have or not have religious beliefs and practice or not practice a religion. Nevertheless, given this coexistence, it is necessary to impose restrictions on this freedom to reconcile the interests of different groups and ensure respect for any of the beliefs. The ECHR in Metropolitan Church of Bessarabia and Others v. Moldova (1999) outlined that states and actors have to remain neutral and impartial in exercising their regulatory powers in this field and their relations with different religions, faiths and beliefs[38].

Speaking of regional laws, many states currently retain laws prohibiting insulting remarks about religion (blasphemy laws), even though such laws contradict international human rights law[39]. Typically, state bans on defamation of religions fit into one or more of the following categories:

[37] European Court of Human Rights (2013). Overview of the Court's case-law on freedom of religion, p. 7.
[38] Metropolitan Church of Bessarabia and Others v. Moldova, Application no. 45701/99, paragraph 115-16, ECHR 2001-XII.
[39] Villa, V. (2022). Four-in-ten countries and territories worldwide had blasphemy laws in 2019. [online] Pew Research Centre. Available at: https://theintercept.com/2022/09/21/facebook-censorship-palestine-israel-algorithm/ [Accessed 12 November 2022].

1. Direct blasphemy, the purpose of which the prohibition most often seeks to protect the state religion, its doctrine, symbols or revered personalities from criticism, stereotyping or defamation[40];

2. Insult to religious feelings, the purpose of the prohibition of which is to seek to protect the feelings of a person or group "offended" or "outraged" by instances of blasphemy against the religion with which they identify themselves[41];

3. Vague, overbroad laws restricting manifestations of freedom of expression relating to religion or belief, which aim to protect public morals or public order, which are used to limit freedom of expression illegally and to halt public debate on faith and beliefs[42].

Despite the presence of such restrictions in national laws, international human rights standards indicate that such prohibitions on blasphemy should be abolished. This recommendation was first highlighted in the Rabat Plan of Action (UN Human Rights Council, 2013) and was strongly endorsed in General Comment 34 (2011) of the Human Rights Committee. To a large extent, the national legislation of European Union countries on the prohibition of defamation of religion has been reflected at the regional level in the CoE, the European Union and the "Inter-American Systems"[43]. The main argument for supporting the repeal of blasphemy laws is their counterproductive nature, both in principle and in practice. In international human rights law, there is a significant distinction between protecting a person's right based on their religion or belief and the protection of ideas and opinions.

Going back to the balance of religiously offensive speech and abusive hate speech, it can be challenging to draw the line between expressing views in an improper manner and hate speech of such a severe degree that it is not protected by the right to freedom of expression[44]. The markers of whether the speech belongs to a particular type of hate speech are only auxiliary tools. When it comes to race, religion or other characteristics, the line must be drawn between speech that may and may not be protected by human rights law. To illustrate an example, we will consider the

[40] Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre, p. 29. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].
[41] Ibid.
[42] Ibid.
[43] Ibid.
[44] Donnelly, J. (2015). Freedom of Religion and Freedom of Expression: Religiously Offensive Speech and International Human Rights. Hum. Rts., 10, 20, pp. 31-32.

statement, "Romanians should not be allowed into our country". While the statement is deeply offensive and discriminatory, it is protected under international human rights law. Since freedom of movement in the EU is a legitimate and contemporary political debate issue, it will be accordingly protected as political speech. However, a Belgian man subsequently convicted of handing out leaflets saying "send non-European jobseekers home" and "oppose the Islamisation of Belgium" could not rely on his right to freedom of expression[45]. A statement that could inflame an already tense situation or provoke conflict would likely be considered hate speech by a court. Therefore, for both examples, restrictions will only be possible if they cause an unwarranted interference with the guaranteed right of another person or group.

It follows from the above that a clear distinction system is used to distinguish between profoundly offensive speech and speech that incites hatred against religious groups or individuals, thus neutralising the issue of a conflict between two fundamental rights - the right to freedom of expression and the right to follow the religion of one's choice. The key to resolving cases where there is an alleged conflict between these rights is the need for the Court to clearly distinguish between hate speech, which needs to be prohibited, and deeply offensive speech, which, despite its negative emotional connotations, cannot be prohibited.

Denial of historical events

Various forms of 'memory laws' exist in many countries and prohibit any expression that denies the occurrence of historical events, in many cases, involve significant violations of criminal law, including periods of severe persecution, genocide and others. Denial of such events under so-called 'memory laws' is perceived as a direct attack on the dignity of the victims of historical events and those associated with them. Belavusau and Gliszczynska-Grabias (2017) noted that such laws have as their stated aim the prevention of the recurrence of adverse historical events[46]. However, it is important that under international human rights law, claims about the truth of historical events are not preserved as such.

Importantly, international standards on freedom of speech, expression and opinion do not permit restrictions on expressing ideas and opinions solely because they are "false" or "wrong", even in

[45] See Féret v. Belgium (2009), Council of Europe: European Court of Human Rights, 15615/07.

cases where such statements are deeply offensive. As the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2012) has noted, "By requiring that writers, journalists and citizens to provide only the version of events that the government has approved, states can subordinate free speech to official versions of events[47]". UN Human Rights Committee (2011) indicates that while there are concerns about the intolerance of speech denying the occurrence of historical events, legitimate restrictions on the expression, as mentioned above, should only be imposed where such expression crosses the line of advocacy of discriminatory hatred that constitutes incitement to violence, hostility or discrimination[48]. The protection of individual rights must be the basis for any restrictions on the right to freedom of expression and opinion.

Protection of the state and state officials

States regularly use the concept of 'hate speech' to discredit or even prohibit speech that is critical of the state, state symbols (flags, emblems, and others) or state officials. While the laws of some states expressly prohibit the insult of an abstract concept such as the state, restrictions on the right to freedom of expression and opinion in the form of the prohibition of 'sedition', expressions contrary to 'national unity' or 'national harmony' are still common. International standards do not allow restrictions on the right to freedom of expression that are implemented to protect the state or its symbols from insult or criticism. The state and its symbols cannot be the object of hate speech because they are not human beings and therefore have no human rights. For individuals associated with the state, such as heads of state and other state officials, state representative status is not a protected characteristic on which claims of hate speech can be based[49]. According to the Johannesburg Principles, "no one shall be punished for criticising or insulting the nation, the state or its symbols, the government, its institutions or public officials unless the criticism or insult was intended and likely to provoke imminent violence[50]".

[47] United Nations General Assembly (2012). The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (on FOE), A/76/357, para. 55.

[49] UN Human Rights Committee (2011). General comment no. 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34, para 38.

[50] Article 19 (1995). The Johannesburg Principles on National Security, Freedom of Expression and Access to Information, principle 7(b)

Furthermore, the Johannesburg Principles stipulate that expression may only be restricted as a threat to national security if the state can demonstrate that:

1. Expression is intended to incite imminent violence;
2. The expression is likely to provoke violence;
3. There is a direct and immediate connection between the expression and the likelihood of violence occurring or occurring[51].

Based on these criteria, the UN Special Rapporteur on Terrorism has proposed a model definition of incitement to terrorist offences: "the intentional and unlawful dissemination of a message to the public to incite the commission of a terrorist offence which causes a risk that one or more such offences may be committed[52]". However, many states justify undue restrictions on freedom of speech and expression to protect national security, including bans on justifying, promoting or glorifying terrorist acts or related extremism and radicalisation. Such broad prohibitions lack justification under international human rights law and can be applied arbitrarily to restrict expressions such as political debate, censorship of minorities or dissenting opinions. International human rights law, therefore, recommends that states should prohibit incitement to terrorist acts but should distinguish such incitement from hate speech and ensure that, where such speech is defined as hate speech, the necessary elements of the prohibition include both the intention to incite terrorist acts directly and the likelihood of a terrorist attack occurring or being committed as a result of such speech. Laws that do not meet the aforementioned requirements should be repealed.

[51] Ibid.
[52] UN Human Rights Council (2010). Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism, A/HRC/16/51, para. 30-31.

Chapter 5: Hate Speech in Cyberspace

5.1 Historical Background

When it comes to protecting the right to freedom of expression, it should be noted that prior to the emergence of the Internet, laws against defamation, insult or incitement to hatred protected only the public sphere[53]. The legal rules governing the right to freedom of expression included those expressions that were directly public in nature - that is, that purportedly affected a wide range of people, had an impact on the public to whom the expression was addressed, and had particular consequences for society. Returning to the criteria for the illegality of hate speech, the public nature of the expression (the public dissemination of harm) is still one of the key criteria for identifying a particular expression. The Internet is indeed an integral part of the public space these days. However, additional problems arise when the public sphere begins to overlap with the private sphere. When a person expresses his or her opinion to another person in a private context in any, even extreme, form, such prohibition cannot legally be enforced by the state due to the lack of an element of publicity. At the same time, publishing one's opinion on the Internet has different consequences.

The words "to publish" and "publication", which are commonly used to refer to any material that exists on the Internet, come from the Latin word "publicare", which means "to make public, to make known". Based on the principles of the existing online platforms, by publishing one's opinion there, one addresses an indefinite audience, which can consist of relatives, friends and acquaintances as well as an indefinite circle of people. To this day, we perceive the Internet as a space free from legal regulation, which gives its users complete freedom of action. Often users perceive their social media accounts as part of themselves, which gives them a sense of freedom to act on their own within their accounts. However, this impression is misleading. Since online platforms are balancing between public and private spheres, the online space is actively regulated

[53] Cammaerts, B. (2009). Radical pluralism and free speech in online public spaces: The case of North Belgian extreme right discourses. International journal of cultural studies, 12(6), pp. 555-575.

by both national and international law in its various fields and branches[54]. International Human Rights law is the branch that directly ensures freedom of expression on the Internet and is therefore responsible for dealing with online Hate Speech. Thus, the joined forces of concerned parties such as international human rights law institutions, representatives of Internet companies and others started to create new instruments and mechanisms to regulate online freedom of expression[55]. In order to analyse the impact of the legal side of human rights on how Big Tech companies work with Hate Speech on internet platforms, we will then discuss the key instruments of international human rights law that regulate freedom of expression and opinion in the online space.

5.2 Legislative Background

One of the first instruments of this kind was the so-called Budapest Convention or Convention on Cybercrime (2001). In fact, it was the first international treaty dealing with crimes committed using or on the Internet. Then came into force the Additional Protocol to the Convention on Cybercrime on the incrimination of racist acts and xenophobic acts committed through information systems (2003). This document became another game-changer with its ruling that the computer system (Internet) should be seen as a special way to disseminate expression, along with oral, written, artistic and other forms of expression[56]. In 2012, the Human Rights Council established that "the same rights that people have offline must also be protected online[57]". It is this thesis that still remains the basis of the struggle for human rights on the Internet. Thus, from a human rights perspective, all of the above documents have the common goal of ensuring that all human rights are respected regardless of online or offline nature.

Lastly, a document explicitly addressing the topic of Hate Speech was adopted in 2016. The adoption of the Code of Conduct on Combating Illegal Hate Speech on the Internet (2016) was a collaboration between the European Commission and major international companies such as Facebook, Microsoft, Twitter and YouTube, with Instagram, Google+, Snapchat, Dailymotion and Jeuxvideo.com joining in 2018. The Code of Conduct implies an obligation on the aforementioned companies to remove hate speech, which relates to hate speech that should be banned, within

[54] Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. Unesco Publishing, p. 14.
[55] Ibid., p. 45.
[56] Council of Europe (2003). Additional Protocol to the Convention on Cybercrime, Incriminating Racist and Xenophobic Acts through Information Systems, European Treaty Series - No. 189.
[57] UN Human Rights Council (2012). Resolution 20/8 on the Internet and Human Rights, A/HRC/RES/20/8.

twenty-four hours of the publication of such speech online. The Code's implementation consists of regularly monitoring the content published on the aforementioned online platforms and its removal.

It is difficult to deny that the theses enshrined in the documents mentioned above are timely and extremely necessary to observe human rights on the Internet. However, what makes them noteworthy is the predominantly lenient nature of the legislation, which means that their provisions can be ignored or circumvented by unscrupulous parties for their own purposes and gain. On the one hand, the involvement of major IT companies in the legal regulation of the online space and their cooperation with international law demonstrates their seriousness in helping the international community secure the right to free expression and combat illegal hate speech on the Internet. On the other hand, the soft, recommendatory nature of human rights regulations creates blind spots that companies can exploit to achieve their goals without directly violating the law. Such blind spots will be discussed below.

International human rights law stipulates that it is the responsibility of the state to ensure that human rights are respected, promoted and protected. At the same time, speaking of online space, the state cannot fully control the online actions of its citizens, as the primary control of a user's online actions is exercised by the online platforms on which the actions take place. The UN Special Rapporteur on Freedom of Opinion and Expression[58] (Special Rapporteur on Freedom of Information), in a 2016 report, instructed states "not to demand or otherwise exert any pressure on the private sector to take measures that unnecessarily or disproportionately interfere with freedom of expression, whether through laws and policies or extrajudicial measures[59]", while pointing out that private intermediaries, namely online platforms, are not well equipped to establish the illegality of the Internet. This statement is somewhat controversial due to the specific nature of the work of online platforms. In the context of hate speech, the most disseminated platforms are social networks. Unlike other resources on the Internet, which are mainly a source of this or that information, social networks were created as a direct way to express one's opinion, discuss and find like-minded people on the Internet. The widespread popularity and large number of users around the world have led in due course to the need to control content (censorship) and user behaviour, to establish certain restrictions on user freedom of action, as well as to create algorithms for the removal of illegal content. Unlike states that are used to fighting hate speech in isolated

[59] UN Human Rights Council (2016). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/32/38, para. 40-44.

manifestations, the major social networks understand the specifics of user behaviour on the Internet and presumably have a better understanding of how to combat hate speech on a large scale. Thus, online platforms constitute a mediator between the individual, who has the right to freedom of speech and opinion, and the state, which has the right to restrict this freedom in cases determined by law.

Nevertheless, in order to combat hate speech effectively, it is necessary to understand the typology of hate speech. The previous chapter of this paper demonstrated how international human rights law distinguishes speech that should be prohibited and speech whose prohibition violates freedom of speech and expression. Based on that, it becomes evident that the cooperation of international law and major online platforms is integral to ensuring the right to free expression on the Internet.

Regarding the criteria for identifying hate speech on the Internet, General Comment 34 established the same restrictions as identified in the previous chapter. The limitations imposed on electronic forms of communication and expression "must be subject to the same criteria used for 'offline' (non-electronic) forms of expression[60]". Accordingly, Internet platforms are obliged to use the criteria defined in international human rights law and the practice of the HRC and the ECtHR to determine the legitimacy of expression and to allow restrictions on freedom of expression only in particular cases. In addition, the Additional Protocol to the Convention on Cybercrime defines as contrary to law 'any writing, any image or any other representation of ideas or theories that advocate or incite hatred, discrimination or violence against any person or group of persons on the basis of race, colour, national or ethnic origin, or religion[61]'. On this basis, it can be established that international human rights law makes no significant distinction in determining the legitimacy of the expression in both online and offline spaces.

When talking about the regulation of published content, it is necessary to refer once again to the typology of hate speech. As defined above, hate speech is divided into speech that should not be prohibited, speech that may be prohibited, and speech that should be prohibited. Concerning online hate speech, the main difference here is the presence of an intermediary (an online platform) between the state and the individuals with the right to free expression. Due to the specific nature of their work, many ways of identifying speech that needs to be removed from the public online

[60] UN Human Rights Committee (2011). General comment no. 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34, para. 12.
[61] Council of Europe (2003). Additional Protocol to the Convention on Cybercrime, Incriminating Racist and Xenophobic Acts through Information Systems, European Treaty Series - No. 189, Article 2.

space have been shaped by Big Tech companies themselves as a response to people's online actions. Before ascertaining whether IT companies follow the standards set by human rights law, it is necessary to examine what identification criteria and ways of identifying hate speech they use.

5.3 Methods of Identification

In terms of methods of identifying online hate speech, it is worthwhile to examine one participant of the Code of Conduct against Hate Speech on the Internet as a case study. As long as Facebook is one of the world's first Internet companies that remains the largest social network with almost three billion users worldwide[62], this company is a great representative to conduct research on how IT companies work with Hate Speech.

Like other prominent social networks, Facebook defines the limits of acceptable content not only by the user agreement but also by having special sections on its website designed to familiarise the average user with the platform's rules of use. Such sections contain a large amount of detailed information on both the tools the platform uses to deal with hate speech and information on the platform's cooperation with states and the international community. By comparison, the social network Instagram, which is also a signatory to the Code of Conduct on Combating Hate Speech online in the Community Guide, does not publish information about what efforts the administration has made to combat hate speech, so it is not easy to find this information. The Community Guidelines, published on the app's official website, explicitly state "no advocacy of violence or attacks against people based on their race, ethnic or national origin, gender, gender identity, religious affiliation, sexual orientation, diseases or disabilities[63]". However, it is clarified that the publication of hate speech is only possible if the publication aims to tell and combat hate speech. In addition, the category of 'hate speech or symbols' has recently been added to one of the categories of unacceptable content on Instagram. Such wording illustrates the fact that, when it comes to the right to freedom of expression, such opinion can be expressed in a variety of forms that are not limited to the oral or written forms that we are accustomed to.

---

[62] Statista (2023). Number of monthly active Facebook users worldwide as of 1st quarter 2023. [online] Statista. Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [Accessed 20 March 2023].

[63] Facebook Transparency Center (n.d.). Hate Speech Policy rationale [online]. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ [Accessed 15 October 2022].

When it comes to hate speech in its classical (offline) manifestations, the state detects hate speech by its own means. However, due to the specific functioning of the online space, online platforms have to develop their own identification methods without direct assistance from states. The methods used by Facebook, Instagram and other social networks to detect hate speech and other dangerous, harmful and illegal content include[64] :

1. Artificial intelligence (AI);
2. Human expertise;
3. User feedback.

Artificial Intelligence is the most high-tech identification method to detect hate speech exclusively on the Internet. The primary technology used is the so-called "language understanding". Using the criteria described in the previous chapter, the statements already detected by the system, which are to be banned, are placed in a special database. Such data is used for future discussions on hate speech and facilitates the search for hate speech by using "matching words and combinations" or codewords and their combinations[65]. In addition, large online platforms now use technology that offers enough ways to assess whether a piece of content might violate human rights[66].

AI is evolving rapidly, allowing internet platforms to automate the process of finding and removing inappropriate content as much as possible. However, a major problem with this technology is the variety of ways it expresses itself. While AI algorithms generally aim to detect hate speech by keywords/word combinations, this technology is not always able to detect illegal hate speech in artistic forms of expression or to distinguish between illegal hate speech and deeply offensive speech and humour, which are mostly legitimate hate speech that falls under international human rights law protection. For this reason, human resources remain the most effective means of identification.

As mentioned in the previous chapter, the context in which an opinion has been expressed is an essential element used by both state and international judicial bodies in cases involving freedom of expression and hate speech. In doing so, the human resource performing the contextual analysis

[64] Meta (2019). Improving Our Detection and Enforcement. [online] Available at: https://about.fb.com/news/2019/09/combating-hate-and-extremism/ [Accessed 10 October 2022].
[65] Ibid.
[66] Ibid.

and assessment of the expression is the judge competent in the related field of law. Internet platforms also have the human resources to identify and assess hate speech.

Experience and perception are the skills that assist a living person in contextual analysis and evaluation of another person's utterance, while AI learns primarily from observing data that it is presented with[67]. At the moment, AI does not have such capabilities. Although such analysis is complex and resource-intensive, a better substitute for human resources in this context does not yet exist. For this reason, major Internet platforms are creating specialised teams whose list of responsibilities includes not only detecting and removing illegal content but also tracking trends and public reactions related to legitimate hate speech[68]. With this information, qualified teams produce reports and surveys that help improve the platform's functioning but are simultaneously used by states, both nationally and internationally, to combat hate speech and ensure our legitimate human rights.

Another feature of online hate speech is the possibility for a user to report detected inappropriate speech to the platform. Such a feature is found in most social networks and other online resources and is designed to facilitate finding inappropriate content. The average user can complain about any content while choosing the reason why they found a particular publication inappropriate. Due to the fact that internet users are often unaware of the existence of different types of hate speech, this function is mainly used to complain about offensive material. As we discussed earlier, offensive content is not always prohibited hate speech, so this method is closely related to the previous one. Expert knowledge is the tool for assessing a large number of user complaints and giving them a peer review. It should be noted here that only some users' complaints about content removal will be enforced. In order to satisfy such a request, the content must meet the criteria for abusive hate speech.

At first glance, considering all tools of Hate Speech identification, we presume that online platforms may be genuinely interested in protecting human rights and make every possible effort to work with the identification and removal of Hate Speech properly. Previously we have briefly described all the ways used by online platforms to track and remove hateful content. While human expertise and feedback are combined with human factors and are more familiar in the context of

[67] See Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st international workshop on gender equality in software engineering, p. 1.
[68] Meta (2019). Improving Our Detection and Enforcement. [online] Available at: https://about.fb.com/news/2019/09/combating-hate-and-extremism/ [Accessed 10 October 2022].

dealing with hate speech (analogous to legal case studies), the use of AI is a fundamentally new practice. Companies argue that this use of technology is highlighted by high levels of efficiency and performance. At the same time, a detailed examination of Facebook policies, numerous journalistic investigations, and a critical analysis of artificial intelligence and algorithms lead us to the main problem - biased algorithms.

Chapter 6: Investigation of Biased Algorithms

To demonstrate the downside of platforms with Hate Speech, we will focus on the most problematic of the three methods mentioned above to identify Hate Speech online: detection by artificial intelligence.

Meta, the mother company of Facebook, claims that AI is an extremely useful, practical and performative tool for identifying Hate Speech on its platforms. Indeed, we cannot deny all the pros of modern technologies. Given the total number of users of Meta's platforms (Facebook, Instagram, WhatsApp and others), numbering millions and billions, the company cannot provide enough human resources and expertise to deal with Hate Speech. Thus, the invention of automotive algorithms made the additional possibility of identifying and removing illegal content and has made the work of experts much easier. At the same time, the statements about the high level of efficiency and productivity of AI raise some questions.

In 2021, the Internet exploded with news that a former Facebook employee had leaked confidential documents showing the actual situation within the company[69]. Former employee Frances Haugen filed a complaint against Facebook with US federal law enforcement agencies, namely the Securities and Exchange Commission. As part of her complaint, she provided original evidence of the company's numerous violations[70]. These documents were first published in the Wall Street Journal and included confirmation that Facebook's internal policies were based on encouraging and spreading Hate Speech, toxic content, disinformation and political unrest[71]. The public spread of this information became the starting point for multiple high-profile articles and investigations of such Internet giants. Among other breaches and controversial actions, details of the automated

[69] Purkayastha, P. (2021). How Facebook's algorithms promote hate and drive toxic content. [online] New Europe. Available at: https://www.neweurope.eu/article/how-facebooks-algorithms-promote-hate-and-drive-toxic-content/ [Accessed 01 December 2022].
[70] Zubrow, K. (2021). Whistleblower's SEC complaint: Facebook knew platform was used to "promote human trafficking and domestic servitude". [online] CBS News. Available at: https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/ [Accessed 20 January 2023].
[71] Pelley, S. (2021). Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation. [online] CBS News. Available at: https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/ [Accessed 20 January 2023].

algorithms of Meta-owned platforms have been exposed. For this reason, the next part of the research will compile a detailed examination of the main issues surrounding AI and its interaction with Hate Speech.

The study of published information and the subsequent comparison of the facts leads us to a disappointing but definite fact: Facebook's internal algorithms are invisibly biased, have weak performance, and the company itself does not follow the identification criteria established by international Human Rights instruments. To justify this hypothesis, the evidence was gathered from various sources.

As a primary source, officially published company documents, such as Community Standards Enforcement Report[72], an update on progress on AI and hate speech detection[73], and the "Improving Our Detection and Enforcement[74]" report were examined to get Facebook's perception of the situation. Digging deeper, such major investigations on biased algorithms as Whistleblower's case (CBS News, 2021), the study on technical failures of Facebook's AI tools (Insider, 2021), hate speech removal problems (TIME, 2019), Facebook's promotion of hate and drive toxic content (NewEurope, 2021) and others were inspected to gather newly discovered facts, carefully hidden by the company. The technical information on automotive systems and hate speech detection mechanisms was gathered from the latest Big Tech-related technical and engineering studies (Tufekci, 2015; Vosoughi, Roy, Aral, 2018; Schwarz, 2019; Ntoutsi, Fafalios, Gadiraju, et al., 2020). Lastly, to explain the grounds for the existence of biased AI, evaluate possible harm for society, and elaborate on the most effective ways of countering hate speech online, some socio-legal and philosophical studies (Chalermsook, Das Sarma, Lall, Nanongkai, 2015, Brown, 2017, Gagliardone, Gal, Alves, Martinez, 2015; Zhang, Luo, 2018; Castaño-Pulgarín, Suárez-Betancur, Vega, López, 2021) were considered.

Comprehensive analysis of these sources helped to outline two main concerns of algorithmic harms:

[72] Facebook Transparency Center (2023). Community Standards Enforcement Report. [online] Available at: https://transparency.fb.com/data/community-standards-enforcement/?source=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement [Accessed 14 April 2023].
[73] Meta (2021). Update on Our Progress on AI and Hate Speech Detection. [online] Available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [Accessed 10 October 2022].
[74] Meta (2019). Improving Our Detection and Enforcement. [online] Available at: https://about.fb.com/news/2019/09/combating-hate-and-extremism/ [Accessed 10 October 2022].

1. Intentionally biased algorithms
2. Weak performance

6.1 Intentionally Biased Algorithms

Numerous journalistic investigations have noted Facebook's deliberate distribution of hateful content. Like the Massachusetts Institute of Technology (MIT) researchers, Facebook quickly discovered that hate posts and fake news contribute to virality, so it has no incentive to curb such posts[75]. For this reason, the company's algorithms, in many cases, not only fail to prevent the spread of hateful content on its platforms but also program algorithms to spread it more aggressively[76]. Here the question arises: What is the goal of companies pursuing by setting up algorithms in this way? Ironically, the primary goal behind such actions is commercial gain.

When we talk about social networks, the original purpose of such online resources was to provide a single platform for user communication and interaction. Subsequently, companies realized that the messaging capability was not enough to retain users and make a profit[77]. For this purpose, the second, currently prevailing function, entertainment, was introduced. Entertainment content and the ability to create and interact with it (likes, comments, reposts, discussions, and much more) not only supported the public's interest in the project but, like any other part of the entertainment industry, helped attract advertisers[78]. Regarding the most popular social networks of our time (Facebook, Twitter, Instagram, YouTube, TikTok and others), registration and their use are usually free, and additional paid features are aimed at a narrow circle of users and do not bring significant profit. Thus, advertising is the primary way to generate income from an audience of millions of users. By placing their advertisements on the Internet, companies strive to ensure that it is seen by as many people as possible. For this reason, we get the following relationship: the larger the audience and its activity, the more companies are willing to pay to place their ads where this

[75] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151, p.3.
[76] Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. Colo. Tech. LJ, p. 205.
[77] Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15, 100058, p.2.
[78] Chalermsook, P., Das Sarma, A., Lall, A., & Nanongkai, D. (2015). Social network monetization via sponsored viral marketing. ACM SIGMETRICS Performance Evaluation Review, 43(1), p. 259.

activity is observed. In this regard, it is beneficial for online platforms not only to encourage user activity but also to create the activity artificially. So, what does this information mean in the context of Hate Speech?

The theory here is that companies see an extreme benefit in intentionally promoting hashtag content to monetise it. One of the key strategies for platforms to maximise the financial gain from the ads they place is to seek to maximise audience engagement[79]. Using modern terminology, the percentage of audience reach, and the degree of user activity is usually referred to as 'internet traffic'. The principle behind Internet traffic is as follows: the more engagement the published content has, the more it is shown to other users, attracting their attention to particular content and encouraging them to participate in the discussion[80]. Traffic is called "low" when user engagement is low and "high" when they are highly engaged. Thus, this mechanism creates a vicious circle of "promotion of content - high traffic - promotion of the content".

At the same time, an inherent element of this scheme is the human factor. Firstly, online space and its members are the projection of the real-life society picture. Social deviance and norm-violating behaviours can be met as often (or more often) as in real life. Hate speech is often rooted in deviant communication standards and habits, specifics of social interaction, and other intentional or unintentional norm violations[81]. Thus, as long as such deviances exist offline, online space will project them.

Secondly, we should also consider the emotional context. The stronger the emotion evoked by an event, statement, or other activity, the more likely the desired outcome. Unfortunately, this technique works for positive emotions and probably even more effectively for negative emotions (anger, aggression, disappointment and others). While the human brain tends to keep its positive emotions, process them and reflect on them from negative emotions, a person seeks to get rid of them in any possible way[82]. Such processing includes either participating in discussions on

[79] Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15, 100058, p.2.
[80] Ibid., p. 5.
[81] See Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. Aggression and Violent Behavior, 58, 101608.
Internet, social media and online hate speech. Systematic review
https://doi.org/10.1016/j.avb.2021.101608
[82] Devlin, H. (2019). Science of Anger: how gender, age, and personality shape this emotion. [online] The Guardian. Available at:
https://www.theguardian.com/lifeandstyle/2019/may/12/science-of-anger-gender-age-personality [Accessed 20 September 2022].

negative topics or actively expressing aggression. This biological perspective demonstrates why harmful online content is characterised by significantly more user engagement than positive content. In this regard, companies see great benefit in promoting Hate Speech, as it attracts users' attention and creates fertile ground for contextual or native advertising. Involving the audience allows for significant financial gain through profits from advertising contracts and attracting new advertisers. Thus, this kind of algorithm setup "kills two birds with one stone" and is incredibly profitable for online platforms. One of the claims supports this theory: "The more negative comments a piece of content generates, the more likely the link is to get more traffic[83]".

To sum up, this theory confirms the theory that Big Tech corporations see extreme benefits in the deliberate promotion of hate content with the aim of its subsequent monetisation. Unfortunately, at the moment, public support for the fight against hate content and active participation in legal initiatives and social projects related to this topic, to a large extent, is only a “play for the public”. Media companies de facto continue to put their commercial goals (profits) higher than human rights and freedoms.

Another point I would like to raise in the context of intentional algorithm bias is political sensitivity. It is no news that many states still seek to control their population and actions as much as possible, including on the Internet. For this reason, they allow Big Tech corporations to enter the market under certain conditions. The conditions often involve heavy censorship and filtering of information. Similar examples can be seen in decisions of the HRC and ECtHR, in which courts have recognised that in some cases, the banning of so-called “offline” hate speech was used by the state not to protect public safety, the public interest and respect for human rights, but to preserve its own profits and interests and achieve certain goals[84]. Today, Big Tech corporations are as serious players in the global economic, political and social arena as states are. The power of today's IT giants is determined not only by the amount of money they make but also by the degree of influence over their audiences[85]. Media monopolies seek to take over the internet space to gain a permanent, loyal audience, and influence the users. As such, states can use these companies as a covert intermediary to achieve their political goals. The company is able to tune its algorithms by

[83] Zubrow, K. (2021). Whistleblower's SEC complaint: Facebook knew platform was used to "promote human trafficking and domestic servitude". [online] CBS News. Available at: https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/ [Accessed 20 January 2023].
[84] Douek, E. (2022). Content moderation as systems thinking. Harv. L. Rev., p. 242.
[85] Schwarz, O. (2019). Facebook rules: structures of governance in digital capitalism and the control of generalized social capital. Theory, Culture & Society, 36(4), p. 6.

linguistic/geographical criteria to specifically filter content based on this criterion - accordingly, such information control allows it to control user's political interests and shift public sentiment in the desired direction[86]. Biased setup of algorithms with political overtones is usually called "whitewashing" or "whitelisting". Next, we will take some of the recent and most prominent cases of Hate Speech and whitelisting for a detailed analysis.

To begin with, the case of the algorithms for the Palestinian-Israeli conflict should be highlighted. In 2022, the independent consultancy Business for Social Responsibility (BSR) was commissioned by Facebook to prepare a report on the company's 2021 activities. The independent report revealed bias in the company's algorithms towards Palestine and its people. The report revealed that Facebook removed publications of Arab origin, identified as Hate Speech, far more often than posts of Israeli origin, confirming long-standing complaints about the company's hateful content within the Palestinian-Israeli conflict[87]. As a second evidence of algorithm bias, BSR noted the presence of so-called "algorithmic verification" for users from Palestine and the absence of the same verification for Israeli users. The algorithm, codenamed "Arabic Hate Speech classifier", is "the use of machine learning to flag potential policy violations" and has no Hebrew equivalent[88].

Secondly, the China whitewashing case should also be taken into consideration. Although access to Facebook and other Meta platforms has been officially banned in China since 2009 and is only available to users using VPNs and other third-party tools, this does not prevent China from using the company to manipulate public opinion and divert attention from human rights abuses by the communist regime in Xinjiang. In 2022, political correspondent Andrew Tillett published an investigation based on a report by the independent Australian Strategic Policy Institute ("ASPI"). The ASPI report accuses the Chinese Communist Party of "using information campaigns to force countries, businesses and civil society not to criticise Beijing for its treatment of the Uyghurs, a Muslim minority[89]". After analysing 613,301 Facebook posts with links to Xinjiang, it was determined that Facebook's role in this information campaign was to deliberately promote

[86] Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15, 100058, p.3.
[87] Biddle, S. (2022). Facebook report concludes company censorship violated Palestinian Human Rights. [online] The Intercept. Available at: https://theintercept.com/2022/09/21/facebook-censorship-palestine-israel-algorithm/ [Accessed 12 November 2022].
[88] Ibid.
[89] Tillett, A. (2022). China using Facebook to whitewash human rights abuse: analysis. [online] Financial Review. Available at: https://www.afr.com/politics/federal/china-using-facebook-to-whitewash-human-rights-abuse-analysis-20220719-p5b2p6 [Accessed 20 January 2023].

disinformation "created by fringe media and conspiracy websites that were often sympathetic to the narrative position of authoritarian regimes[90]".

The examples mentioned above demonstrate that the unpunished bias of Internet companies and allowing prohibited types of Hate Speech inevitably leads to states and certain groups using this for purposes that threaten public security. Despite the pursuit of democratic values and respect for human rights, we still live in a world where politics is messy, and the political and economic gain of big political players is often more important than the right of citizens to free speech and opinion. For this reason, the problem of territorially biased algorithms is the most serious one mentioned in this paper. Undoubtedly, the previously mentioned biased settings are harmful to society. Nevertheless, in the Palestine-Israel and China cases, we see that politicised algorithms can affect not only the lives of individuals or discriminated groups but also the course of history, leading to irreparable consequences for society.

6.2 Weak Performance

While determining the validity/invalidity of speech is difficult even using human resources, there are severe scientific concerns about artificial intelligence's level of accuracy and ability to separate "Hate Speech that should be prohibited and removed" from toxic, offensive and controversial content protected by free speech rights and should be allowed[91]. In various public sources, Meta representatives have emphasised AI's high level of effectiveness in identifying and removing Hate Speech from their online platforms. For example, in its 2020 Facebook report, Meta notes an increased level of performance of its automated systems and steady progress in training them to recognise and remove inappropriate content[92]. Facebook's updates indicate that its automated Hate Speech tracking system is delivering the best results even though "the nature of challenges changes", and people themselves tend to avoid detection of their statements by the AI system. At the same time, there is increasing evidence that the actual level of artificial intelligence

[90] Ibid.
[91] Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15, 100058, p.13.
[92] Facebook Transparency Center (2023). Community Standards Enforcement Report. [online] Available at: https://transparency.fb.com/data/community-standards-enforcement/?source=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement [Accessed 14 April 2023].

performance is much weaker than they are trying to prove. Senior engineers at Facebook and other Meta-owned platforms have claimed in reports that automated systems are only able to remove 2% of Hate Speech[93]. At the same time, other sources within the company stated that "AI recognises and removes between 3% and 5% of illegal hate content and 0.6% of content that violates Facebook's rules on violence[94]". Given that the company's human resources are insufficient to deal with HC within the multi-million user space, we expect artificial intelligence to be the primary and most effective source of identification. So why is a technology with such high expectations showing such poor performance?

Previously, we provided evidence that Facebook has been repeatedly criticised for allowing expressions that should be banned. The paradox is that the company's algorithms not only allow illegal content but also remove (prohibit) allowed content. Facebook admits in its statements that "the same words can often be interpreted as either good or hateful depending on where they are published and who reads them, and training machines to pick up on this nuance is particularly hard[95]". This statement brings us back to the typology of HC, namely the expressions that should be allowed and, accordingly, the content that should not be removed. Analysing the information on automated search and deletion of HCs, we can outline three main problems related to the weak performance of AI:

- weak internal policies;
- weak language detection;
- limited capacity.

6.2.1 Weak Internal Policy

When talking about the weak internal policy of Meta and its platforms, this definition refers primarily to the lack of clearly defined categories and Hate Speech and its criteria. Although the

[93] Walsh, E. (2021). Facebook claims it uses AI to identify and remove posts containing hate speech and violence, but the technology doesn't really work, report says. [online] Insider. Available at: https://www.businessinsider.com/facebook-ai-doesnt-work-to-remove-hate-speech-and-violence-2021-10 [Accessed 20 January 2023].
[94] Ibid.
[95] Meta (2021). Update on Our Progress on AI and Hate Speech Detection. [online] Available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [Accessed 10 October 2022].

company publishes publicly available community rules and guidelines for each of its social networks, numerous studies show that the criteria for identifying Hate Speech are weak, non-transparent and often violate a person's right to free expression and opinion. To confirm this hypothesis, we will look at a real-life example given by Meta on its website. In its 2021 report, Facebook notes the importance of establishing the context of a post to which such a comment responds. "This is great news" can mean entirely different things when it is commended on posts announcing the birth of a child and the death of a loved one[96]. Such an example is quite controversial for demonstrating how artificial intelligence works with the removal of content. In this example, the company is trying to illustrate that the same comment, depending on the context, can either violate the company's Hate Speech policy or comply with it. Accordingly, the company implies that the comment 'this is good news' under news about someone's death is prohibited and should be removed.

Given that Meta actively participates in international legal initiatives to combat hate speech and claims to follow prescribed international standards for identifying hate speech, it is appropriate to analyse this case directly under human rights law. After subjecting the "this is good news" comment under the death announcement to the legally recognised system of identification and qualification of expression, we get the following conclusion: despite the negative, offensive nature of the comment, intended to hurt another person's feelings, we cannot classify the comment as an expression that can or should be prohibited. A user who has left such a comment under the death announcement retains the right to delete such a statement within the powers granted to him/her by the internet platform (personal blog management). At the same time, Facebook cannot remove every comment, statement or publication that offends the feelings of a certain person. A hurt person's feelings often constitute a reaction to dissent or criticism in their direction, to trigger words or topics, to black or cruel humour, or to a direct insult to them. However, none of this should be a reason to ban such speech and remove it from the internet. In offline and online spaces, there should always be room for dissent, which is one of the most important manifestations of freedom of expression. It should not be forgotten that the main marker of illegal Hate Speech is a direct incitement to hatred with a high probability of consequences behind its context. Therefore, offensive content is not automatically recommended for removal and must be subjected to a thorough review of whether it should be removed, both on the severity threshold and on the presence of the following factors:

[96] Meta (2021). Update on Our Progress on AI and Hate Speech Detection. [online] Available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [Accessed 10 October 2022].

- an intense irrational negative emotion;
- the target of the inappropriate expression is a legally protected social group;
- the utterance is destructive, dehumanising and discriminatory towards the aforementioned social group;
- the statement is likely to result in actions or consequences that have a destructive impact on society.

This example demonstrates that the Company's example of identifying and categorising hateful content represents prohibition of speech falls under the protection of the right to freedom of expression and is, therefore, legitimate and authorised. Given the complexity and multiplicity of forms of human expression, permitted Hate Speech is not only the most challenging type in terms of identifying but also the most important in the context of online space. Despite modern society's attempts to address the underlying prejudices from which any expression of hatred originates, the concept of a 'world without hatred' is utopian, as is the 'Internet without hatred'. Considering all the cultural, religious, racial, gender, political and other diversity of the world's population, numerous legal, political, educational and other measures are often insufficient to reduce the percentage of hatred and intolerance in society. For this reason, some examples of hate, despite being controversial, are not always a reason to prohibit such speech.

The difficulty of this debate is that we ourselves often confuse hatred with dissent, both online and in real life. The real ground for the right to freedom of expression implies that offensive, toxic, critical or any other expression of dissent should not be prohibited unless it can be shown to have the intent to incite hatred and the potential for serious consequences of such expression. While freedom of expression may be restricted on the basis of the protection of national security or public order, these provisions cannot be used to suppress dissent in society unless the expression of such dissent meets the criteria for legitimate restrictions on the right to freedom of speech, expression and opinion[97]. Therefore, internet companies' responses to Hate Speech must first and foremost be guided by a commitment to respect and protect the human right to freedom of speech and expression and be based on the norms, standards and recommendations of international human rights law.

[97] UN Human Rights Council (2013). Annual report of the United Nations High Commissioner for Human Rights: Addendum, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4.

6.2.2 Language Detection

A non-obvious point in the context of poor performance of anti-Hate Speech algorithms is weak language detection. Meta claims that its main Facebook platform includes so-called 'language classifiers' in more than 40 languages worldwide[98]. At the same time, the company declined to provide a complete list of languages available to AI for Hate Speech detection[99]. It can be inferred from publicly available information that Meta relies on its users (reports and complaints) and on the company's human resources (mechanical tracking and verification of user complaints) for hateful content in the remaining languages. At the same time, there is evidence that most of the investment in the development of automated Hate Speech tracking systems is directed at improving the English algorithms[100].

Facebook's leaked data demonstrates that Meta deliberately neglects the lack of language classifiers, which contributes to the proliferation of toxic content on platforms[101]. Notably, this problem affects both mainstream and minority languages. For example, data from the SEC filings mentioned earlier indicate that Hindi and Bengali, which are respectively the third and sixth most widely spoken languages in the world, have serious problems with Hate Speech processing. At the same time, minority languages are even more affected by such language bias.

The main "underwater stone" of the uneven distribution of Hate Speech detection efforts in a language context is the following: the lack of language-specific algorithms reduces the chances of Hate Speech detection and simultaneously increases the chance of its rapid and uncontrolled distribution. A human resource represented by company employees is not sufficient to manually process the enormous flow of content published daily on social networks. Meta responds to such criticism by saying that it does not currently have "a dataset large enough to train an AI

[98] Perrigo, B (2019). Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. [online] Time. Available at: https://time.com/5739688/facebook-hate-speech-languages [Accessed 20 February 2023].
[99] Ibid.
[100] Perrigo, B. (2021). Facebook Let an Islamophobic Conspiracy Theory Flourish in India Despite Employees' Warnings'. [online] Time. Available at: https://time.com/6112549/facebook-india-islamophobia-love-jihad/ [Accessed 05 October 2023].
[101] Purkayastha, P. (2021). How Facebook's algorithms promote hate and drive toxic content. [online] New Europe. Available at: https://www.neweurope.eu/article/how-facebooks-algorithms-promote-hate-and-drive-toxic-content/ [Accessed 01 December 2022].

programme[102]". At the same time, Facebook itself, in an interview with TIME, reported that it has a list of countries that are given priority in the context of language algorithms. Such bias is related to the aforementioned Hate Speech criterion set by the Rabat Plan of Action - possible consequences. Because countries such as Myanmar, Sri Lanka, India, Libya, Ethiopia, Syria, Cameroon, the Democratic Republic of Congo, and Venezuela are extremely likely to move hate speech from the online environment to the real world, Facebook claims to focus on algorithms in the language of those states[103]. This proves the fact that the company is capable of training its artificial intelligence tools in Hate Speech language recognition, but only in particular circumstances. While Facebook's intention to prevent real-life hate speech is commendable, this does not negate the fact that such segregation discriminates against linguistic minorities and deliberately spreads hateful content.

The case of hate speech against the Rohingya minority in Myanmar should be highlighted as tangible evidence of the discriminatory nature of language algorithms. In 2018, the population of Rohingya living in Myanmar were mass attacked, forcing them to flee the country. Ethnic cleansing in the country has been triggered by ethnic disinformation campaigns and direct incitement to violence, spread precisely on Facebook, Myanmar's most popular means of communication. Banned hateful content in Burmese (spoken by some 42 million people) was actively circulated on Facebook due to the lack of an HC identification algorithm in Burmese, and moderators only responded to content that had been complained about by a user[104]. As a result, UN Fact Finders said that disinformation campaigns orchestrated by Facebook played a "defining role in provoking the latest episode of violence against Rohingya"[105].

Based on the above, the problem of linguistic bias in algorithms is more severe than it first appears. The principle of non-discrimination is a key one in the context of human rights. Accordingly, such behaviour by the company constitutes direct discrimination on the basis of language. Even considering the company's claims that its financial, technical and human resources are limited, this

[102] Perrigo, B (2019). Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. [online] Time. Available at: https://time.com/5739688/facebook-hate-speech-languages [Accessed 20 February 2023].
[103] Ibid.
[104] Barron, L. (2018). Facebook Is Failing to Control Hate Speech Against the Rohingya in Myanmar, Report Finds. [online] Time. Available at: https://time.com/5368709/facebook-hate-speech-myanmar-report-rohingya [Accessed 14 January 2023].
[105] Meixler, E. (2018). U.N. Fact Finders Say Facebook Played a 'Determining' Role in Violence Against the Rohingya', [online] Time. Available at: https://time.com/5197039/un-facebook-myanmar-rohingya-violence/ [Accessed 20 December 2022].

is not a valid reason for such a human rights violation. Although we do not have open and precise information about how the algorithms are programmed, we can assume that their language customisation is based on teaching artificial intelligence to recognise certain combinations of words and phrases - Hate Speech markers. Given that the company's Hate Speech policy is common to all users regardless of geographical, linguistic or any other affiliation, it is only a matter of translating those ubiquitous Hate Speech markers into the correct language for the algorithms. Even taking into account the wide variety of world languages and dialects, the implementation of Hate Speech language algorithms in a particular language is more a matter of company effort and willingness than a large number of resources.

6.2.3 Capacity Limit

A final issue for the algorithms that needs to be addressed is the capability limit. Either Facebook or numerous external sources reveal that automated systems and their algorithms often have problems determining whether the content is prohibited or allowed. The reason for this is that the nature of expression problems is volatile and constantly evolving.

First of all, as mentioned in the theoretical part of this paper, when we talk about the right to freedom of expression today, we mean not only freedom of speech but also other contemporary forms of expression (written, non-verbal, visual, artistic and others). As such, it falls on the shoulders of artificial intelligence to monitor and evaluate non-verbal forms of potential Hate Speech. Due to the high complexity of the task, which involves searching and evaluating a variety of forms of expression and their context, there is a high probability that the algorithms may overlook inappropriate content.

Second, it should be taken into account that Internet and social media users are able to avoid detection by automated systems deliberately. On the one hand, people's eagerness to avoid detection by Hate Speech recognition systems suggests that users are beginning to "filter" their online speech by adjusting to the users' guidelines and community rules, indicating more conscious action and awareness of possible consequences within the online space. On the other hand, despite the lack of open and transparent information about how the algorithms work, the average user can consider the basic principles of how they work and thus seek ways to express hate that will not be noticed/qualified as prohibited hate.

Thirdly, context is still one of the most challenging elements for automated analysis. In its updated report on the progress of AI in working with Hate Speech, Facebook claims that "the algorithms can be easily bypassed as long as the means of human expression are too versatile, volatile, creative, and most of the time somewhat controversial depending on the angle of a sight[106]", and working with evaluating the context of a statement "has historically been a challenge for AI, because determining whether a comment violates our policies often depends on the context of the post It is replying to[107]".

Summarising the previous points, the question of the limits of the algorithms' capacity brings us one more time to the problem of the application of identification criteria. As mentioned earlier, international human rights law criteria for identifying hate speech are auxiliary rather than mandatory. The human element in human rights law makes it impossible to limit human expression to a fixed framework. For example, courts can shape and modify the criteria for identifying Hate Speech based on factors such as the current political environment, religious influences, attitudes towards minorities and public opinion while respecting standards for dealing with Hate Speech and not violating the right to freedom of expression. At the same time, artificial intelligence aimed at working with hate speech is not critical thinking and cannot assess context and other factors as effectively as humans. Consequently, the lack of such flexibility in automated systems and low level of accuracy leads to inevitable disruptions, whether deleting allowed content or allowing banned content. Therefore, AI systems need to evolve with these issues on time, which means being constantly updated by companies.

[106] Meta (2021). Update on Our Progress on AI and Hate Speech Detection. [online] Available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [Accessed 10 October 2022].
[107] Ibid.

Chapter 7: Combating the Blind Spots and Finding a Balance

The analysis of the automated algorithms performed in the previous chapter supports the hypothesis that Facebook's Hate Speech detection and removal algorithms are invisibly biased and do not comply with the international legal standards of Hate Speech identification. This theory begs the answer to the further question: What are the ways to combat these blind spots?

Today Facebook's infringement of the right to free speech, expression and opinion is primarily based on a deliberate disregard for international standards and regulations on handling Hate Speech. The problems, loopholes and weaknesses in algorithms are not only a consequence of limited algorithmic capabilities and lack of artificial intelligence but rather internal company policy. Big Tech corporations depend on those who can provide them with tangible (commercial) and intangible (degree of influence, political power) support. Although the users bring money and popularity to businesses, financial gain is not the only type of gain companies pursue. The power of today's IT giants is not only determined by the amount of money earned but also by the degree of influence. This theory explains the biased algorithms that allow companies to get the most out of their work. Some might argue and suggest that companies are autonomous entities, free to create their own community rules and thus identify Hate Speech without using international human rights standards. Indeed, media moguls and Big Tech corporations are not states and, accordingly, are not subject to such strict human rights obligations. In order to answer the question of the need for international legal standards, we should return to the basis of Hate Speech - the right to free expression.

Like other fundamental human rights, the right to freedom of expression is protected by national and international law. As such, online resources interacting closely with this freedom cannot independently provide a legally justifiable assessment of a particular expression. As early as June 2018, the UN Special Rapporteur on the Promotion and Protection of Free Expression, David Kaye, called on social networks to adopt international human rights law as the authoritative standard for moderating content on their platforms in a Human Rights Council thematic report on

Internet content regulation[108]. This view was immediately supported publicly by the two largest social networks, Facebook and Twitter. For example, Twitter CEO Jack Dorsey responded to a similar statement by David Kaye, published not only in an official report but also on his personal Twitter account, by saying that Twitter's rules on freedom of speech and expression should be based exclusively on international human rights law[109]. Meta itself has officially stated that decisions of the competent courts and international recommendations within the framework of the right to freedom of expression will be the basis for content regulation on the platform[110]. The conclusion drawn from these statements is that cooperation between internet platforms and the international community lies in the platforms' willingness to follow international recommendations for identifying and combating illegal hate speech. In October 2019, a new supplemental report was released, where David Kaye reiterated the significant problems in regulating online hate speech and the need for an accountability mechanism for internet companies. The main challenge identified in the report is the application of standards for identifying and combating hate speech, initially developed for states, by internet platforms.

Hate speech is the most challenging area of content moderation due to the lack of unique, very specific features[111] and the need to define the category of speech for each specific case. To this end, Kay first recommends that social networks, along with states, use the previously mentioned elements outlined in the Rabat Plan of Action, namely the context of the speech, the status of the speaker, the intention, the form and content of the speech, the scope and harmfulness of the speech. International human rights law recommends that content should only be removed where unlawful hate speech is identified solely through the use of these elements. At the same time, states also retain the possibility of requiring the removal of unlawful hate speech on the Internet. Regarding the removal of hateful content, Kay clarifies in the 2018 report that when a state intends to oblige a platform to remove certain content, states must comply with the usual requirements for restricting freedom of expression secured by international human rights law[112]. Such requirements include

[108] UN Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35.
[109] jake. (2018, August 10). Agree w all of this. Our early values informed our rules. We likely over-rotated on one value, & then let the rules react to rapidly changing circumstances (some we helped create). We need to root these values in human rights law. A starting consideration: [Tweet]. Available at: https://twitter.com/jack/status/1027962500438843397?lang=en
[110] Douek, E. (2019). Why Facebook's 'Values' Update Matters. Lawfare.
[111] Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, 10(5), pp. 5-6.
[112] UN Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35.

legality, the need to protect national security, public order, public health or morals, and respect for the rights or reputations of others.

At the moment, international human rights law still needs to develop ways to increase the accountability of the IT-companies. The cooperation between the actors is the key to developing an effective response to the problem[113]. In terms of specific recommendations that IT companies can and do apply at this point, the following steps should be made:

1. Steps towards transparency

While Facebook has publicly available information about what the platform perceives as hate speech and how it combats it, these findings are non-exhaustive and need more transparency. The speed and scope of technical development of online platforms only allow for a partial assessment of their actual impact on human rights and freedoms, particularly on the right to freedom of expression. For this reason, the provision of transparent periodic reviews and reports by IT companies in the context of activities related to identifying and combating abusive hate speech can be adequate.

Considering all difficulties related to the use of AI in hate speech detection, we do not need companies to provide fake numbers and statistics on the effectiveness of its mechanisms. For the technical side, all we need here is to understand to what extent we can actually rely on AI tools, what are the recent weaknesses and what steps we should jointly take to increase the performance without harming human rights. From the prism of social sciences, these reports will also allow safeguard mechanisms of human rights to track trends and developments in online hate speech, analyse human behaviour online and make recommendations based on the actual information companies provide. Lastly, the transparency of such reports will help human rights instruments to identify legal gaps and provide a framework of operation in accordance with international human rights law for each specific company.

[113] Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. Unesco Publishing, p. 53.

2. Development of precedent

This recommendation is probably the most innovative concept. In this context, the notion of 'precedent' is considered an example of content deleted by a platform with a legal justification for the deletion. As for an average user, it is often hard to understand hate speech and its categories without seeing an actual example. Seeing simple examples of different hate speech categories, forms of expression and contexts from real-life or hypothetical cases would constate a significant contribution to informal education of the public in the framework of the right to freedom of expression. Simultaneously, a publicly available 'precedent library' would not only have an informative function for users but allow for monitoring compliance of precedents with legal requirements. As the concept of hate speech is perceived from different perspectives in jurisprudence, media, sociology, and everyday life, having such real but practical examples of various hate speech cases would create a better understanding of all fields of study.

3. Cross-sectional studies initiatives

In this paper, all AI-related issues were examined from the views of socio-legal sciences. However, analysing the potential and actual effects of automotive algorithms on human rights requires a comprehensive understanding of their technical side. Thus, the proposal here is to find ways to further the researchers' interdisciplinary education. Such legislative initiatives as the Draft Ethics Guidelines for trustworthy AI from the European Commission's high-level Expert Group on AI (2018) or the European Parliament resolution containing recommendations to the Commission on Civil Law Rules on Robotics (2017) became the first steps towards general regulation of artificial intelligence [114]. Notwithstanding, we need some further human rights-related actions.

From the view of human rights-related sciences, the collaboration of Big Data experts and socio-legal study representatives would increase the effectiveness of future human rights regulations and mechanisms. As for Big Tech companies such as Facebook, they can organise training, coaching sessions, seminars and other educational visits of socio-legal scholars to provide updates on the newly introduced AI tools and their features. On the other side, international human rights bodies

[114] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356, p. 10.

can also arrange educational sessions for data scientists and representatives of media corporations to explain the specifics of hate speech online regulations.

As a result, we would get highly-qualified specialists trained to work specifically with AI and human rights who would significantly contribute to today's discussion's main goal – combating online hate speech.

All in all, when imposing restrictions on speech that qualifies as hate speech, Internet companies must follow the same standards as their counterparts in cases involving restrictions on freedom of expression. International human rights law permits restrictions on freedom of expression only when necessary to protect the rights or reputations of others, national security, public order or public health or morals[115]. As an intermediary between international human rights law, the state and the individual, today's online platforms are in dire need of implementing international standards on hate speech identification to ensure the sanctity of the right to freedom of expression online except where a restriction of such freedom is legitimate under the letter of the law. At the same time, the use of the experience of online hate speech platforms by international law is worthwhile to improve the tools for ensuring the right to freedom of expression.

[115] United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December): p. 171, Article 19 (3).

Chapter 8. Conclusions

To this day, the concept of hate speech remains subtle and vague. Its changing nature, directly depending on factors such as public opinion and permissiveness, educational and cultural processes, and regional and global politics, together with others, influence our perception of hate speech at a specific time. Modern society has become more tolerant of things that used to be considered immoral or taboo. At the same time, the global spread of information and the ability to exchange views, launched by modern technologies, results in the active expression of dissent as an integral part of human communication. Answering the question "What is hate speech?" we are generally guided by general notions of morality. It is human nature to distinguish between good and evil, truth and lies, reality and fiction. The catch is that all these categories depend solely on the point of view. Evil is the second alternative norm, while good cannot be universal and always follows someone's interests. Thus, we are still dependent on all the attempts to make the universal definition, typology and criteria of hate speech, whether proposed by human rights instruments or means of artificial intelligence.

Artificial intelligence and algorithms are not evil. It is a vital tool, without which we would easily drown in an excessive flow of information that is difficult to filter with human resources alone. All the algorithms and automotive technologies is a new practice that still has severe shortcomings despite continuous improvement and performance enhancement. However, while some algorithm failures can be attributed to a lack of technology and its limitations, most of the problems and blind spots described in this study are the results of internal company policies. The analysis shows that the company benefits from certain violations of the right to freedom of expression and opinion. To achieve truly successful and effective AI tools, companies like Facebook should first be obliged to comply with international standards for Hate Speech. Creating a perfect AI mechanism to detect Hate Speech is a utopian idea. Just as we cannot replace judges in courts with artificial intelligence, we cannot put the responsibility of deciding whether the content is malicious or not entirely in the hands of automated systems. In this vein, we can put our maximum efforts towards a comprehensive approach that includes automated search and removal processes for the more straightforward cases, using human resources (individual experts, commissions, translators, and

others) to deal with more complex, ambiguous cases, and educating users about free speech and online rules of conduct.

We should bear in mind that despite the publicly shown respect for human rights and freedoms and the desire to protect them, such companies always put their own gain first. With the enormous level of influence over their users, Big Tech platforms can impose on us their understanding of ethics and morality to benefit their goals or those with whom they cooperate. The immeasurable flow of information that has gripped us in recent years does not always allow us to distinguish an enforced opinion from our own. Human rights have become the political and ideological dominant of our time. Thus, it is vital to ensure that there is a mechanism capable of creating unbiased and uniform norms that can protect the individual from biased and unscrupulous actors and make them accountable for their illegitimate actions.

Internet platforms and social media represent such a global mass phenomenon that they are hard to control. Their operation is highly based on computer and engineering sciences, creating difficulties for social and legal sciences scholars to get the full picture of how to protect human rights from their harmful impact. From the social sciences side, we cannot yet be accountable for the technical side of hate speech identification algorithms. However, we should find ways to encourage IT companies to keep the maximum transparency level on their hate speech detection mechanism so we can supervise the use of automated decision-making and its compliance with human rights regulations.[116] International human rights law should remain the primary safeguard mechanism for protecting the right to freedom of speech and opinion against unscrupulous companies and holding them accountable. Still, the lack of awareness of the technical processes from legal mechanisms demonstrates the need for cross-sectional research to understand the roots of the bias and prevent irrevocable harm to human rights. We must emphasise that the key to combating online hate speech and protecting our right to freedom of expression is actors' cooperation, conscientious approach, and constant response to the incoming challenges.

[116] Page 5, Ntoutsi, E, Fafalios, P, Gadiraju, U, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining Knowl Discov. 2020; 10:e1356. https://doi.org/10.1002/widm.1356

Bibliography

Legal sources

Article 19 (1995). The Johannesburg Principles on National Security, Freedom of Expression and Access to Information.

Committee on Legal Affairs (2017). Report with recommendations to the Commission on Civil Law Rules on Robotics, European Parliament, 2015/2103.

Council of Europe (2003). Additional Protocol to the Convention on Cybercrime, Incriminating Racist and Xenophobic Acts through Information Systems, European Treaty Series - No. 189.

Council of Europe (2001). Convention on Cybercrime. European Treaty Series - No. 185.

Council of Europe (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, ETS 5.

Equality and Human Rights Commission (2015). Freedom of Expression. [online] ISBN 978-1-84206-595-2. Available at: https://www.equalityhumanrights.com/sites/default/files/20150318_foe_legal_framework_guidance_revised_final.pdf [Accessed 20 November 2022]

European Court of Human Rights (2013). Overview of the Court's case-law on freedom of religion.

European Commission (2016). The EU Code of conduct on countering illegal hate speech online.

European Commission's High-Level Expert Group on Artificial Intelligence (2018). A definition of AI: Main capabilities and scientific disciplines.

Féret v. Belgium (2009), Council of Europe: European Court of Human Rights, 15615/07

Jersild v. Denmark (1994). Council of Europe: European Court of Human Rights, 36/1993/431/510.

Handyside v. UK (1976). Council of Europe: European Court of Human Rights, 5493/72.

Metropolitan Church of Bessarabia and Others v. Moldova (1999). Council of Europe: European Court of Human Rights, 45701/99, ECHR 2001-XII.

Otto-Preminger-Institut v. Austria (1994). Council of Europe: European Court of Human Rights, 13470/87.

Smajić v. Bosnia and Herzegovina (2018). Council of Europe: European Court of Human Rights, 48657/16.

United Nations General Assembly (1965). International Convention on the Elimination of All Forms of Racial Discrimination. Treaty Series, vol. 660, p. 195.

United Nations General Assembly (1966). International Covenant on Civil and Political Rights. Treaty Series 999 (December): p. 171.

United Nations General Assembly (2010). Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. Frank La Rue: addendum, A/HRC/14/23.

UN Human Rights Council (2010). Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism, A/HRC/16/51

United Nations General Assembly (2012). The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (on FOE), A/76/357.

UN Human Rights Committee (2011). General comment no. 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34.

UN Human Rights Council (2012). Resolution 20/8 on the Internet and Human Rights, A/HRC/RES/20/8.

UN Human Rights Council (2013). Annual report of the United Nations High Commissioner for Human Rights: Addendum, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4.

UN Human Rights Council (2017). Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism, A/HRC/6/17.

UN Human Rights Council (2016). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/32/38.

UN Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35.

UN Human Rights Council (2012). The promotion, protection and enjoyment of human rights on the Internet, 20/8.

UN General Assembly (1948). Universal Declaration of Human Rights, 217 A (III).

Doctrinal sources

Article 19 (2015). Hate Speech Explained Toolkit. [online] Free World Centre. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 20 November 2022].

Barendt, E. (2005). Freedom of speech. OUP Oxford

Belavusau, U., & Gliszczynska-Grabias, A. (2017). Law and Memory: Towards Legal Govenrance of History. Cambridge University Press, ISBN-13: 978-1107188754.

Brown, A. (2017). What is hate speech? Part 1: The Myth of Hate. Law and Philosophy Vol. 36 No. 4, pp. 419–468. Springer.

Brown, A., & Sinclair, A. (2019). The politics of hate speech laws. Routledge.

Cammaerts, B. (2009). Radical pluralism and free speech in online public spaces: The case of North Belgian extreme right discourses. International journal of cultural studies, 12(6), pp. 555-575.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. Aggression and Violent Behavior, 58, 101608.

Chalermsook, P., Das Sarma, A., Lall, A., & Nanongkai, D. (2015). Social network monetization via sponsored viral marketing. ACM SIGMETRICS Performance Evaluation Review, 43(1), 259-270.

Douek, E. (2019). Why Facebook's 'Values' Update Matters. Lawfare.

Douek, E. (2022). Content moderation as systems thinking. Harv. L. Rev., 136, 526.

Donnelly, J. (2015). Freedom of Relgion and Freedom of Expression: Religiously Offensive Speech and International Human Rights. Hum. Rts., 10, 20.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. Unesco Publishing.

Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st international workshop on gender equality in software engineering (pp. 14-16).

Mihkailova, E., Bacovska, J. and Shekerdjiev, T. (2013). Freedom of expression and hate speech. [online] Skopje: OBSE. Available at: https://www.osce.org/files/f/documents/1/e/116608.pdf [Accessed 01 November 2022].

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15, 100058.

Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: a comparative analysis. Cardozo L. Rev., 24, 1523.

Schwarz, O. (2019). Facebook rules: structures of governance in digital capitalism and the control of generalized social capital. Theory, Culture & Society, 36(4), 117-141.

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. Colo. Tech. LJ, 13, 203.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Walker, S. (1994). Hate speech: The history of an American controversy. U of Nebraska Press.

Weinstein, J., & Hare, I. (2009). General Introduction: Free Speech, Democracy, and the Suppression of Extreme Speech Past and Present in Extreme Speech and Democracy. OUP Oxford.

Weber, A. (2009). Manual on hate speech. [online] Strasbourg: Council of Europe, Cop. Available at: http://icm.sk/subory/Manual_on_hate_speech.pdf [Accessed 29 September 2022].

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, 10(5), 925-945.

Websites and articles:

Barron, L. (2018). Facebook Is Failing to Control Hate Speech Against the Rohingya in Myanmar, Report Finds. [online] Time. Available at: https://time.com/5368709/facebook-hate-speech-myanmar-report-rohingya [Accessed 14 January 2023].

Biddle, S. (2022). Facebook report concludes company censorship violated Palestinian Human Rights. [online] The Intercept. Available at: https://theintercept.com/2022/09/21/facebook-censorship-palestine-israel-algorithm/ [Accessed 12 November 2022].

Devlin, H. (2019). Science of Anger: how gender, age, and personality shape this emotion. [online] The Guardian. Available at: https://www.theguardian.com/lifeandstyle/2019/may/12/science-of-anger-gender-age-personality [Accessed 20 September 2022].

Facebook Transparency Center (2023). Community Standards Enforcement Report. [online] Available at: https://transparency.fb.com/data/community-standards-enforcement/?source=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement [Accessed 14 April 2023].

Facebook Transparency Center (n.d.). Hate Speech Policy rationale [online]. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ [Accessed 15 October 2022].

jake. (2018). Agree w all of this. Our early values informed our rules. We likely over-rotated on one value, & then let the rules react to rapidly changing circumstances (some we helped create). We need to root these values in human rights law. A starting consideration: [Tweet]. Available at:

https://twitter.com/jack/status/1027962500438843397?lang=en [Accessed 20 December 2022].

Meixler, E. (2018). U.N. Fact Finders Say Facebook Played a 'Determining' Role in Violence Against the Rohingya', [online] Time. Available at: https://time.com/5197039/un-facebook-myanmar-rohingya-violence/ [Accessed 20 December 2022].

Meta (2019). Improving Our Detection and Enforcement. [online] Available at: https://about.fb.com/news/2019/09/combating-hate-and-extremism/ [Accessed 10 October 2022].

Meta (2021). Update on Our Progress on AI and Hate Speech Detection. [online] Available at: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [Accessed 10 October 2022].

Perrigo, B (2019). Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. [online] Time. Available at: https://time.com/5739688/facebook-hate-speech-languages [Accessed 20 February 2023].

Perrigo, B. (2021). Facebook Let an Islamophobic Conspiracy Theory Flourish in India Despite Employees' Warnings'. [online] Time. Available at: https://time.com/6112549/facebook-india-islamophobia-love-jihad/ [Accessed 05 October 2023].

Purkayastha, P. (2021). How Facebook's algorithms promote hate and drive toxic content. [online] New Europe. Available at: https://www.neweurope.eu/article/how-facebooks-algorithms-promote-hate-and-drive-toxic-content/ [Accessed 01 December 2022].

Statista (2023). Number of monthly active Facebook users worldwide as of 1st quarter 2023. [online] Statista. Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [Accessed 20 March 2023].

Tillett, A. (2022). China using Facebook to whitewash human rights abuse: analysis. [online] Financial Review. Available at:

https://www.afr.com/politics/federal/china-using-facebook-to-whitewash-human-rights-abuse-analysis-20220719-p5b2p6 [Accessed 20 January 2023].

United Nations. What is hate speech? [online]. Available at: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech?gclid=Cj0KCQjw2v-gBhC1ARIsAOQdKY1-lyuALindV77gVDFVTCAoHjG8Z1Ft-a3TXmyMD-didrWl4cyUAnYaAuS4EALw_wcB [Accessed 13 September 2022].

Walsh, E. (2021). Facebook claims it uses AI to identify and remove posts containing hate speech and violence, but the technology doesn't really work, report says. [online] Insider. Available at: https://www.businessinsider.com/facebook-ai-doesnt-work-to-remove-hate-speech-and-violence-2021-10 [Accessed 20 January 2023].

Villa, V. (2022). Four-in-ten countries and territories worldwide had blasphemy laws in 2019. [online] Pew Research Centre. Available at: https://theintercept.com/2022/09/21/facebook-censorship-palestine-israel-algorithm/ [Accessed 12 November 2022].

Zubrow, K. (2021). Whistleblower's SEC complaint: Facebook knew platform was used to "promote human trafficking and domestic servitude". [online] CBS News. Available at: https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/ [Accessed 20 January 2023].

Pelley, S. (2021). Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation. [online] CBS News. Available at: https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/ [Accessed 20 January 2023].