Detection of disinformation and content filtering using machine learning: implications to human rights and freedom of speech

Lumbardha Hasimi[1,†], Aneta Poniszewska-Marańda[2,**,†]

[1] *Institute of Information Technology, Lodz University of Technology, Lodz, Poland*
[2] *Institute of Information Technology, Lodz University of Technology, Lodz, Poland*

Abstract

The spreading of fake news and disinformation has become a pressing and widely debated issue in recent times, with far-reaching implications for society. While efforts to combat fake news and disinformation have gained momentum, there is a need to consider the implications for human rights in the context of disinformation detection and content filtering. The paper investigates the implications of fake news and disinformation and various aspects of human rights and freedom of speech. It examines the use and the impact of Artificial Intelligence (AI), while highlighting the potential risks to the right of freedom of expression when responses to these issues lead to censorship and the suppression of critical thinking. The paper further emphasizes the need for a balanced approach that safeguards freedom of expression and human rights while addressing the negative impacts of misinformation and biased algorithms.

Keywords

Disinformation and fake news detection, machine learning, data collection, content filtering, human rights

1. Introduction

The proliferation of fake news and disinformation has become a pressing and widely debated issue in recent times, with far-reaching implications for society. Fake news refers to intentionally false news items that aim to deceive readers, and its impact extends beyond misinformation to interference in political affairs and the incitement of hate crimes [20]. While efforts to combat fake news and disinformation have gained momentum, there is a need to consider the implications for human rights in the context of disinformation detection and content filtering.

Recognizing the significance of the issue, the "Joint Declaration on Freedom of Expression and Fake News, Disinformation, and Propaganda" emphasized that general prohibitions on the dissemination of "false news" are incompatible with international standards for restrictions on freedom of expression [8]. The same highlighted the responsibility of state actors to disseminate reliable and trustworthy information, refraining from knowingly or recklessly spreading false statements. State actors should, by their domestic and international legal obligations and their public duties, ensure that they disseminate reliable and trustworthy information, including matters of public interest, such as the economy, public health, security and the environment [21].

However, balancing the need to control fake news with the preservation of freedom of expression presents challenges, as criminalizing fake news could lead to censorship and the suppression of critical thinking, especially when recognizing the AI approaches as effective tools towards massive control. In this context, Machine Learning (ML) as a subset of AI, has emerged as a crucial tool in the battle against fake news [22]. Nevertheless, the use of algorithms for content filtering and moderation is known to pose risks to the right to freedom of expression. Human biases and subjective judgments can influence

ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2024: the 46th European Conference on Information Retrieval), March 24, 2024, Glasgow, UK

[**] Corresponding author.

[†] These authors contributed equally.

lumbardha.hasimi@dokt.p.lodz.pl (L. Hasimi); aneta.poniszewska-maranda@p.lodz.pl (A. Poniszewska-Marańda)
0000-0001-5774-3356 (L. Hasimi); 0000-0001-7596-0813 (A. Poniszewska-Marańda)

the outcomes of automated filtering systems, leading to potential discrimination and infringement of privacy and data protection rights [13]. Furthermore, the presence of biases within ML models, such as sample bias, exclusion bias, observer bias, and racial bias, can perpetuate unfair discrimination and reinforce cultural biases.

Legal frameworks and international declarations emphasize the need for transparency, accountability, and proportionality in implementing content filtering measures. Blocking and filtering systems should be limited to cases that serve a pressing social need, be proportionate to the legitimate aims pursued, and not substitute for law enforcement [23]. Over-blocking or false positives and under-blocking or false negatives present challenges, as legitimate content may be wrongfully restricted or illegal content may go undetected.

Freedom of expression is a fundamental right, but it is not absolute and may be subject to limitations to protect the rights of others, morality, public order, and general welfare. However, any restrictions must be justified within the framework of human rights law and should not jeopardize the right itself. While the fight against fake news and disinformation is essential, it is crucial to consider the implications for human rights, including freedom of expression, privacy, and non-discrimination. Balancing the need for effective disinformation detection and content filtering with the preservation of fundamental rights is a complex task that requires careful consideration of legal frameworks, transparency, and accountability in implementing such measures.

This paper investigates the implications of fake news and disinformation and various aspects of human rights and freedom of speech. It discusses the Joint Declaration on Freedom of Expression and "Fake News", emphasizing the need for reliable and trustworthy information. The paper examines the use and the impact of Artificial Intelligence (AI), while highlighting the potential risks to the right of freedom of expression when responses to these issues lead to censorship and the suppression of critical thinking. The study delves into different types of biases, while discussing the legal basis and documents relevant to content filtering and blocking, considering the importance of proportionality and necessity in restricting freedom of expression. The paper further emphasizes the need for a balanced approach that safeguards freedom of expression and human rights while addressing the negative impacts of misinformation and biased algorithms.

The paper is structured as follows: Section 2 presents the problem of implications to human rights in the area of disinformation detection and content filtering. Section 3 overviews the concept of use of machine learning for the fake news. Section 4 describes the types of Biasis present in ML and its implications while section 5 deals with legal basis, documents and articles implying the issue.

2. Disinformation detection and content filtering: implications to human rights

Fake news and disinformation is one of the most urgent and debated issues recently due to various interference, polarizing societies and inspiring hate crimes. "Fake news" refers to news items that are intentionally and verifiably false, and seek to mislead readers [24]. Unfortunately, not only is this type of information useful for aspiring autocrats, extremists and other bad actors, but it is becoming a harsh business for many entities as it affects various aspects of human activity.

Considering the appealing situation regarding the fake news and disinformation, many movements have started aiming to tackle many of the issues that lead and impact its wider scope of concerns. In March 2017, the Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda [1], stressed that general prohibitions on the dissemination of information based on vague and ambiguous ideas, such as "false news", are incompatible with international standards for restrictions on freedom of expression, and furthermore does not justify the dissemination of knowingly or recklessly false statements by official or state actors. The Joint Declaration urged state actors to exercise caution in disseminating information, emphasizing the importance of reliability and trustworthiness. It specifically called on ensuring against the creation, endorsement, promotion, or dissemination of statements that are known or reasonably should be known to be false, or demonstrate a careless disregard for verifiable information [2].

In the struggle trying to control or monitor fake news through Machine Learning (ML) and policies, responses to certain issues pose a risk to the right to freedom of expression. Furthermore, there are widespread concerns that criminalising fake news could lead to censorship and the suppression of critical thinking and dissenting voices [3]. Another implication regarding fake news and its detection relates to another issue, the impact of fake news and its detection on human rights.

3. Battle against the fake news using Machine Learning

When assessing the human rights impacts of algorithms, it further must be considered that designers of algorithmic systems have varying levels of discretion when deciding, for instance, what training data to use or how to respond to false positives, and that the power of the operator of the algorithm may lie in and the knowledge of the structure of the dataset, rather than in insight into the exact workings of the algorithms [4].

For instance, content filters which can make mistake in case of unwanted materials that can be accepted by the filter (false negatives), or acceptable materials can be rejected (false positives). Mistakes can take place in all filtering/classification models, this when the filtering is based on human intervention, on a fully automated process or a combination of the two. Given the fallibility of filtering, every socio-technical system for moderation should include measures to identify mistakes and react to them.

Moreover, in machine learning domain the term "ground truth" is used to refer to the correct outcome, as identified through standards external to the system, as opposed to the outcome that is proposed by the system. This expression is often used in machine learning apparently derives from cartography, and opposes the representation on a geographical map to the real situation on the ground, which provides the undisputable standard to determine whether the map is correct or wrong [5].

In online filtering, however, the ground truth is not provided by a physical reality but rather by human assessments on whether a certain item falls or not into a category of unwanted content, according to laws, guidelines and social norms, as interpreted by the individual assessors. Human assessments constitute indeed the training set of automated classifiers and provide the basis for evaluating and correcting the outcomes of such classifiers [8]. The labelled data is used to train the classifier, allowing it to learn to recognize and classify similar patterns in new, unseen data. Although, in best cases human assessors compare the classifier's predictions with the ground truth labels to determine its effectiveness and identify any errors or areas for improvement, the model will ultimately be limited to the ground truth. Thus, the working of an automated filtering system will reflect the attitudes, and possibly the biases of the humans whose behaviour the system is meant to emulate [9].

The standard for making content accessible to the public also varies in different social and cultural contexts: language and content that is acceptable in certain communities may be unacceptable to others. In large communities, different views may exist concerning what is appropriate or inappropriate to the community concerned. For instance, there has been a discussion concerning the removal of certain images in social networks, in general, or in connection with special contexts. In some cases, filtering may raise some issues concerning unfair discrimination, to the extent that the content delivered by or concerning certain groups may be excluded or deprioritised.

However, removing certain content from being available could align with certain ethical principles such as protecting individuals from harm and upholding community standards. On the other hand, the right to equal inclusion regardless of the content's ethical nature raises issues related to freedom of expression, individual autonomy, and the need for diverse representation. Upholding this principle may be seen as promoting inclusivity, supporting marginalized voices, and challenging stigmatization. Determining which approach is "more ethical" depends on various factors including cultural context, legal frameworks, societal values, and individual perspectives. Although all mechanisms carefully consider the potential impacts and ethical implications of each course of action, balancing competing values and interests while striving for fairness, justice, and respect for human rights presents many social dilemmas. In table 1, two common elements emerge among all the mechanisms presented regarding the ground limitation of a right: morality and determined by law. These elements serve as foundational criteria for justifying the restriction or limitation of certain rights within legal and ethical frameworks.

In this context, content filtering can lead to not only the identification of unwanted material but also to the identification of the individuals that have published the materials being filtered, which in return will affect the privacy and data protection rights of these individuals [5, 10]. In certain cases, to an extent that is disproportionate relative to the benefit that filtering out unwanted content may provide. This issue has emerged in the EU law, in connection to automated filtering for detecting copyright violations [11]. Content filtering may also be used to identify and target political oppositions, as has happened in China and Egypt [19].

4. Types of Biasis present in ML and its implications

There are a few basic types of Biasis present in ML and its implications defined in the literature Fig. 1).

Sample bias. Sample bias occurs when a dataset does not reflect the realities of the environment in which a model will run [5]. The differences between the dataset and real-world environment can lead to inaccurate or biased predictions/decisions. Sample bias can arise for various reasons, such as uneven sampling, selection bias, or the presence of confounding factors that are not adequately represented in the dataset. An example of this is certain facial recognition systems trained primarily on images of white men [5]. These models have considerably lower levels of accuracy with women and people of different ethnicities [11]. Another name for this bias is selection bias.

Exclusion bias. Exclusion bias is most common at the data preprocessing stage due to various factors. In [12] it is stated as most often a case of deleting valuable data thought to be unimportant or due to the systematic exclusion of certain information. It happens for instance when dealing with datasets that contain the absolute majority of one kind of group, and the given minority is not significant in comparison [15]. Additionally, exclusion bias can arise from unintentional omission of specific information that may be relevant to the problem being addressed.

Observer bias. Known also as the confirmation bias; observer bias is the effect of seeing what you expect to see or want to see in data [11]. This can happen when researchers go into a project with subjective thoughts about their study, either conscious or unconscious [13]. It is also seen that when labellers let their subjective thoughts control their labelling habits, resulting in inaccurate data, which in certain cases can lead to discrimination certain group or area.

Racial bias. Racial bias occurs when data skews in favor of particular demographics [15]. Unfortunately, even the most developed models seem to inherit the issue of demographics. This has been seen in facial recognition and automatic speech recognition technologies. There are various factors responsible for this, such as the imbalanced representation of diverse racial or ethnic groups in the training data, inherent biases in the algorithms themselves, or environmental factors that influence the data collection process.

Association bias. This bias occurs when the data for a machine learning model reinforces and/or multiplies a cultural bias [5]. The dataset may have a collection of jobs in which all men are doctors and all women are nurses. Considering the context of gender bias, association bias manifests when the dataset reflects stereotypical links between certain attributes and genders, leading the model to biased predictions. Association bias is best known for creating gender bias [4, 11].

Recent findings indicate that the presence of diversity in training data significantly impacts a neural network's ability to mitigate bias. However, it is worth noting that dataset diversity can also have a negative impact on the overall performance. Additionally, studies highlight the significance of training methods and the types of neurons that develop during the process, as they can greatly influence the capacity to address bias within a given dataset [11].

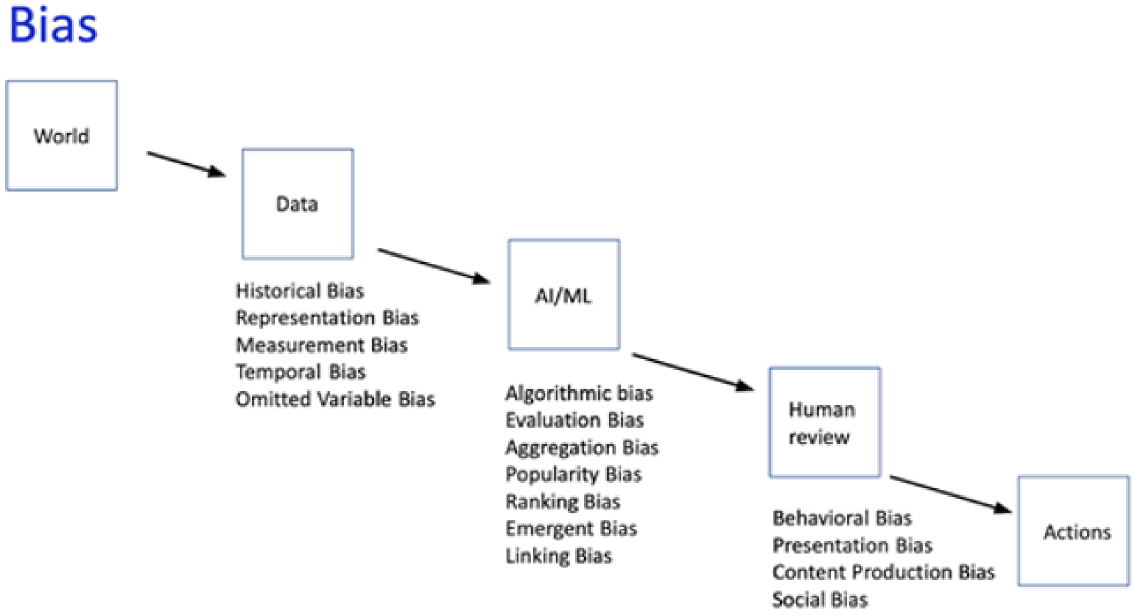

Figure 1: Bias in AI, retrieved from [15].

5. Legal basis, documents and articles implying the issue

Internet service providers should treat all data that travels over their networks fairly, without improper discrimination in favour of a particular application, website or service. Filtering is commonly associated with the use of technology that blocks pages by reference to certain characteristics, such as traffic patterns, protocols or keywords, or based on the perceived connection to content deemed inappropriate or unlawful [19].

Blocking, by contrast, usually refers to preventing access to specific websites, domains, IP addresses, protocols or services included on a blacklist. Blocking targets specific online content or services deemed objectionable or harmful. Whereas internet shutdowns entail a broader disruption of internet or electronic communications. An intentional interruption of internet or electronic communications, made inaccessible or effectively unusable, for a specific population or within a location, often to exert control over the flow of information [13].

5.1. Joint Declaration on Freedom of Expression and the Internet

According to the 2011 Joint Declaration on Freedom of Expression and the Internet [12]:

- "Mandatory blocking of entire websites, IP addresses, ports, network protocols or types of uses (such as social networking) is an extreme measure – analogous to banning a newspaper or broadcaster – which can only be justified by international standards, for example, when necessary to protect children."
- "Content filtering systems which are imposed by a government or commercial service provider and which are not end-user controlled are a form of prior censorship and are not justifiable as a restriction on freedom of expression."
- "Products designed to facilitate end-user filtering should be required to be accompanied by clear information to end-users about how they work and their potential pitfalls in terms of over-inclusive filtering."

5.2. Digital Rights

More than half the worlds population is now online, consequently several hundreds of millions of posts, photos and hundreds of thousands of hours of video is uploaded daily. Behaviours and opinions which may once have been fringe or stayed in the private sphere can go viral. Content which travels seamlessly across border, may be legal in one country, but prohibited in another. People's different social, religious, cultural references also create varying level of sensitivities to the content [10].

The Internet has created an unprecedented ability to communicate across borders to connect communities and create virtual ones. Yet the vastness and immediacy of online platforms have created new tools for online harassment, hate speech. Access to the internet provides access to a universe of information and knowledge, yet it also is full of disinformation and misinformation. While it might facilitate public debate, it also contributes to a more polarized world and potentially skews the truth. People use it organize and mobilise, in some instances in ways that governments find threatening. The internet is an important tool for individuals to document and report on violence and crimes, yet it is also mechanism for bad actors to organize and perpetrate those crimes. As a right of every individual, as articulated in Article 19 of the Universal Declaration of Human Rights, it applies regardless of frontiers and encompasses all forms of media, including digital platforms.

Connecting it with Article 19's prescient wording, the right to freedom of expression applies regardless of frontiers and through any media of one's choice [8]:

- That means that Digital rights comprise the rights which are implicated in our access to and use of these technologies.
- The UN has stated the rights people have offline, extends to the online sphere. Limitations are subject to the same tests and balancing.

General Comment 34 echoes that but specifically puts a positive obligation on states to foster the independence of these new media and to ensure access of individuals thereto [8].

5.3. Key issues in the filtering/classification of data

The operation of algorithms and data processing techniques has tremendous effects on the right to freedom of expression, which includes the right to receive and impart information [23].

According to Article 10 of the ECHR [23], any measure that blocks access to content through filtering or removal of content must be prescribed by law, pursue one of the legitimate aims foreseen in Article 10.2, and must be necessary in a democratic society. In line with the jurisprudence of the European Court of Human Rights, any restriction of the freedom of expression must correspond to a "pressing social need" and be proportionate to the legitimate aim(s) pursued.

Over-blocking or 'false positives', no system can ensure that legitimate content is not wrongfully restricted. In particular, legitimate sites may be blocked because they use the same IP address. Furthermore, according to [12] the following can apply:

- Under-blocking' or 'false negatives': conversely, sites containing illegal or targeted content might not be caught by the blocking/filtering system. This is particularly problematic in the case of online child protection as parents derive a false sense of security from the knowledge that web-blocking measures are in place.
- Failure to address the root causes: blocking/filtering do not address the root causes of the particular problem at issue and are no substitute for law enforcement and the prosecution of serious crimes committed over the Internet.
- Possibility of circumvention: blocks/filters are generally relatively easy to circumvent both by sufficiently tech-savvy end-users and "criminals" when they detect that they have been added to a blocking list.
- Violation of human rights: granular blocking/filtering strategies are deeply intrusive of users' right to privacy and freedom of expression as they analyse the content of the material exchanged between users.

Table 1
Ground for right limitations in different mechanisms adapted from [19].

Grounds for the limitation of a right	UDHR	ICCPR	ACHPR	ACHR	ECHR
Determined by law	X	X	X	X	X
Necessary				X	X
Rights of others	X	X	X	X	
Morality	X	X	X	X	X
General welfare	X				
Public order	X	X		X	X
National / collective security		X	X	X	X
Common interest			X		
Territorial Integrity					X
Prevent disclosure of Confidential Information					X
Maintain Authority/Impartiality of Judiciary					X

Given the above, blocking and filtering mechanisms pose significant threats to freedom of expression and human rights while often proving ineffective in addressing the issues they aim to solve. Consequently, they are considered disproportionate measures and may not be warrant for implementation [13].

5.4. General limitations of freedom of expression

Freedom of expression is not an absolute right. Striking the appropriate balance between competing rights and interests lies at the heart of many disputes concerning rights [13]. Some of the mechanisms to consider:

- It is a general principle of human rights law, found in the UN instruments, the ECHR (Article 17), the ACHR (Article 29) and the ACHPR (Article 27(2)) that "human rights may not be exercised in a manner that violates the rights of others".
- Article 5(1) of the ICCPR, which provides that "nothing in the present Covenant may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms recognized herein or at their limitation to a greater extent than is provided for in the present Covenant".
- Some rights, such as Freedom of Expression, may be subject to internal limitations within the right itself as in paragraph 3 of the ICCPR seen above, or as part of the general limitations clause of the treaty. In principle, any restriction of a right must be capable of being justified both in terms of any internal limitation, as well as the general limitations clause in the treaty [13].

Article 29 of the Universal Declaration of Human Rights (UDHR), on the other hand, contains a general limitations clause, which provides as follows:

- "In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society."

It should be noted that any restrictions on the right to freedom of expression may not put the right itself in jeopardy (Table 1).

Freedom of expression is not an absolute right It is a general principle of human rights law, found in the UN instruments, the ECHR (Article 17), the ACHR (Article 29) and the ACHPR (Article 27(2)) that human rights may not be exercised in a manner that violates the rights of others. For instance, article 5(1) of the ICCPR, provides that "nothing in the present Covenant may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms recognized herein or at their limitation to a greater extent than is provided for in the present Covenant. This means it would prevent a government from passing laws, actions of which would constitute the destruction of a fundamental right guaranteed by the ICCPR. Similarly, it would prohibit a group or individual from engaging in activities that systematically suppress or curtail the exercise of rights beyond what is permitted under the Covenant.

5.5. General Comment No. 34 at para 21.

Striking the appropriate balance between competing rights and interests lies at the heart of many disputes concerning rights. Some rights, such as Freedom of Expression, may be subject to internal limitations within the right itself as in paragraph 3 of the ICCPR as seen above, or as part of the general limitations clause. Any restriction of a right must be capable of being justified both in terms of any internal limitation, as well as the general limitations clause in the treaty.

Another relevant tool with implications for the topic in the EU law is the "Directive on Copyright in the Digital Single Market" known as the "Copyright Directive". This directive was adopted in April 2019 and is composed of several articles, including Article 17 (formerly Article 13), which deals specifically with online content filtering and classification [14]. Article 17 requires certain online platforms to take measures to prevent the unauthorized uploading of copyrighted material, including implementing content recognition technologies. It also requires these platforms to enter into licensing agreements with rights holders in order to ensure that their content is properly licensed and compensated.

5.6. EU Code of Practice on Disinformation

The Code of Practice on Disinformation, announced by the Commission on September 26, 2018, marked the world's first government-encouraged self-regulatory initiative of its kind. The Code was the result of extensive deliberation over four months among a working group consisting of major online platforms and advertisers [16]. It also involved input from a "Sounding Board" comprising stakeholders such as media, civil society, fact-checkers, and academia. The initial signatories included Facebook (including Instagram), Google (including YouTube), Mozilla, Twitter, and four key advertising associations. Microsoft later joined in May 2019 [16].

In contrast to the EU Code of Conduct on Countering Illegal Hate Speech Online, which was established in May 2016 and explicitly negotiated with the Commission, the Commission's association with the Code of Practice is not as apparent [16]. The Code draws inspiration and guidance from statements in the Commission's April Communication. However, it acknowledges that the signatories operate differently and have distinct approaches to addressing non-illegal content. Consequently, not all obligations outlined in the Code apply equally to all signatories.

6. Discussion and Conclusions

This paper discussed the issue of fake news and disinformation, highlighting their impact on society, freedom of expression, and human rights. It emphasizes the need for efforts to address these issues through machine learning (ML) and policies. However, it also raises concerns about the potential risks to freedom of expression and the possibility of censorship.

While efforts to combat disinformation are crucial, it is essential to strike a balance between addressing the negative impacts of misinformation and biased algorithms while safeguarding fundamental rights. The Joint Declaration on Freedom of Expression emphasizes the need for reliable and trustworthy information and cautions against general prohibitions on the dissemination of information that could infringe upon freedom of expression. It highlights the responsibility of state actors to disseminate accurate information while refraining from spreading knowingly or recklessly false statements.

The use of Artificial Intelligence (AI) and machine learning algorithms in content filtering and moderation introduces potential risks to freedom of expression. Biases inherent in the training data and the subjective judgments of human operators can lead to discrimination and privacy infringements. The presence of biases, such as sample bias, exclusion bias, observer bias, racial bias, and association bias, can perpetuate unfair discrimination and reinforce cultural biases. Furthermore, legal frameworks and international declarations stress the importance of transparency, accountability, and proportionality in implementing content filtering measures. Blocking and filtering systems should serve a pressing social need, be proportionate to the legitimate aims pursued, and not substitute for law enforcement. Striking the right balance is challenging, as over-blocking or false positives may restrict legitimate content, while under-blocking or false negatives may allow illegal content to go undetected.

While freedom of expression is not an absolute right and may be subject to limitations, any restrictions must be justified within the framework of human rights law and should not undermine the right itself. It is crucial to consider the implications for human rights, including privacy and non-discrimination, when implementing measures to combat fake news and disinformation.

The battle against fake news and disinformation requires a comprehensive approach that involves technological solutions, legal frameworks, transparency, and accountability. Diversity in training data and careful consideration of the biases present in machine learning algorithms are essential. Furthermore, clear information about the functioning and potential pitfalls of content filtering systems should be provided to end-users.

The paper explores different types of biases present in ML and their implications, such as sample bias, exclusion bias, observer bias, racial bias, and association bias. It further examines the legal basis, documents, and articles related to the issue, including the Joint Declaration on Freedom of Expression and the Internet. The limitations of freedom of expression and the general principles of human rights law are also discussed, followed by a conclusion highlighting the need to strike a balance between competing rights and interests while ensuring the protection of human rights.

References

[1] Huff, M.: Joint Declaration on Freedom of Expression and "Fake News", Disinformation, and Propaganda. In: Secrecy and Society, Vol. 1(2) (2018)

[2] Grassegger, H. and Krogerus, M.: The Data That Turned the World Upside Down. In: MediaWall, [Online]. Available: https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win (2017)

[3] Goodman, E.: Editors vs algorithms: who do you want choosing your news?. Reuters institute. [Online]. Available: http://reutersinstitute.politics.ox.ac.uk/news/editors-vs-algorithms-who-do-you-wantchoosing-your-news

[4] Article 19: Algorithms and automated decision-making in the context of crime prevention. [Online]. Available: https://www.article19.org/resources.php/resource/38579/en/algorithms-and-automated-decisionmaking-in-the-context-of-crime-prevention (2016)

[5] Roselli, D., Matthews, J., and Talagala, N.: Managing Bias in AI. In: Companion Proceedings of 2019 World Wide Web Conference (WWW'19), USA, pp. 539–544 (2019)

[6] Rainie, L. and Anderson, J.: Code-Dependent: Pros and Cons of the Algorithm Age'. In: Pew Research Center (2017)

[7] Kitchin, R.: Thinking Critically About and Researching Algorithms. In: Information, Communication & Society, Vol. 20(1) (2017)

[8] OSCE.: Joint declaration on freedom of expression and "fake news", disinformation and propaganda. In: OSCE. [Online]. Available: http://www.osce.org/fom/302796?download=true (2017)

[9] UNHCR, UN Refugee agency: General Comment No.34: Article 19: Freedoms of opinion and expression. In: General Comment No.34: Article 19. [Online]. Available: https://www.refworld.org/legal/general/hrc/2011/en/79729 (2011)

[10] UN Human Rights Council: Report of The Special Representative of The Secretary-General on The Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, on Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework'. UN Doc A/HRC/17/31, Principles 1–10 (2011)

[11] Bostrom, N., and Yudkowsky, E.: The Ethics of Artificial Intelligence. In: Frankish, K. and Ramsey,

W. (Eds), Cambridge Handbook of Artificial Intelligence, Cambridge University Press, Vol. 316, pp. 316–317 (2014)

[12] UN Human Rights Council: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression on Freedom of Expression, States and the Private Sector in the Digital Age. UN Doc A/HRC/32/38, paras 35–37 (2016)

[13] Article 19: Joint declaration on freedom of expression and the internet. [Online]. Available: https://www.article19.org/resources/joint-declaration-freedom-expression-internet/

[14] UN Human Rights Committee: General Comment No. 34, Article 19: Freedoms of Opinion and Expression. UN Doc CCPR/C/GC/34, paras 21–22, 24–30, 33–35; UN Human Rights Council, 'Report of the UN High Commissioner for Human Rights on The Right to Privacy in the Digital Age' (n 16) para 10; Zakharov v Russia (n 104) para 230; Khan v The United Kingdom App No 35394/97 (ECtHR, 12 May 2000) para 26; Kroon and Others v The Netherlands App No 18535/91 (ECtHR, 27 October 1994) para 31 (2011)

[15] Curto, N. E.: EU Directive on Copyright in the Digital Single Market and ISP Liability: What's Next at International Level?. Case W. Res. J.L. Tech. & Internet, Vol. 11, p. 84 (2020)

[16] Reagan, M.: Understanding Bias and Fairness in AI Systems. In: Medium. [Online]. Available: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3 (2021)

[17] Chase, P. H.: The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem. In: Transatlantic Working Group on Content Moderation Online and Freedom of Expression. [Online]. Available: https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf (2019)

[18] Juusonen, H.: Global Freedom of Expression – ppt presentation'. [Online]. Available: https://slideplayer.com/slide/17554256/ (2022)

[19] Sartor, G.: The impact of algorithms for online content filtering or moderation. Upload filters'.

[20] Vasist, P.N., Chatterjee, D., and Krishnan, S.: The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. In: Information Systems Front, pp. 1–26 (2023)

[21] Rikhter, A.: Policy brief paper on international law and policy on disinformation in the context of freedom of the media. [Online]. Available: https://www.osce.org/representative-on-freedom-of-media/485606 (2023)

[22] Northwood, A.: Unmasking the Truth: How AI is Taking on Fake News and Winning. In: Medium. [Online]. Available: https://medium.com/@alexnorthwood/unmasking-the-truth-how-ai-is-taking-on-fake-news-and-winning-d4211067713f (2023)

[23] EHCR.: Guide on Article 10 of the European Convention on Human Rights. [Online]. Available: https://rm.coe.int/guide-on-article-10-freedom-of-expression-eng/native/1680ad61d6 (2023)

[24] Allcott, H., and Gentzkow, M.: Social Media and Fake News in the 2016 Election. In: Journal of Economic Perspectives, Vol. 31, No. 2, pp. 211–236 (2017)