Digital Gatekeepers: Addressing Fake News, 'Deepfakes', and Hate Speech While Safeguarding Free Speech

Stella Mala[1]

Abstract

Fake news that distorts elections, deepfakes that impersonate identities, and hate speech that fuels intolerance are increasingly converging in the digital sphere, creating unprecedented risks. The most dangerous threat emerges where deepfakes carry elements of hatred, amplifying both disinformation and incitement. This article examines how far online intermediaries can go in filtering such content without eroding fundamental rights. Anchored in Article 10 of the European Convention on Human Rights and the Rabat Plan of Action, it surveys the full body of relevant case law while situating the analysis within the framework of the Digital Services Act and evolving standards of intermediary liability. Comparative examples highlight the limits of algorithmic moderation and the quasi-judicial role of human operators. The central claim is urgent: Without clear safeguards, the intertwined challenges of fake news and hate speech risk undermining democracy itself.

Keywords: freedom of expression; hate speech; fake news; deepfakes; intermediary liability

1. Introduction

Algorithms geared towards keeping users safe online are prone to making errors, often removing legitimate content such as nude artworks or advertisements featuring innocuous subjects. For instance, in 2020, Facebook's algorithm erroneously removed an advertisement depicting onions because it perceived them as nudity.[2] In another notable case, the algorithm removed the renowned artwork *The Origin of the World*, misinterpreting it as pornographic material.[3] In this instance, the algorithm

[1] Adjunct, Department of Law, University of Nicosia, Cyprus. Email: mala.ste@unic.ac.cy

[2] Adam Satariano, 'Facebook Can Be Forced to Delete Content Worldwide, E.U.'s Top Court Rules' (*The New York Times*, 3 October 2019), available at https://www.nytimes.com/2019/10/03/technology/facebook-europe.html (last accessed 15 June 2025).

[3] Lorena Muñoz-Alonso, 'Parisian Court Rules It Has Jurisdiction in "L'Origine du Monde" vs Facebook

failed to assess the legitimacy of the speech effectively. Furthermore, incidents such as the Imane Khelif case illustrate the enduring difficulty of effectively addressing hate speech in online environments. An Algerian boxer who competed in the 2024 Paris Olympics, Imane Khelif was the target of false claims alleging she was transgender, which led to a surge of online hate speech and discrimination against her.[4] This unfounded news spread rapidly, driven by prejudice, and caused significant damage to her reputation. The incident underscores the harmful effects of fake news and online hate speech, particularly regarding issues of gender and sex identity. Human content moderators bear the responsibility of demoting, removing, or eliminating content and may further this action when the content is deemed illegitimate or, more problematically, when it is mistakenly perceived as such despite being legitimate. Addressing the liability of algorithms and human content operators in the context of hate speech or of fake news is a particularly pressing and complex issue.

This article adopts a doctrinal and normative legal methodology, interrogating the evolving interface between human rights law and digital speech regulation. It critically engages with the legal architecture surrounding freedom of expression, hate speech, and fake news, particularly where these phenomena overlap or are indistinctly classified. The analysis is anchored in the three-part test under Article 10(2) of the European Convention on Human Rights (ECHR), serving as a framework to assess the legality, legitimacy, and proportionality of speech restrictions in the digital sphere. To further nuance the treatment of hate speech, the Rabat Plan of Action is employed as a threshold matrix for evaluating incitement. The inquiry is informed by comparative perspectives and adopts a critical lens on the normative role of digital intermediaries, whose algorithmic governance increasingly mediates the boundaries of permissible, lawful expression.

Case' (*artnet*, 9 March 2015), available at https://news.artnet.com/art-world/parisian-court-rules-it-has-jurisdiction-in-lorigine-du-monde-vs-facebook-case-275117 (last accessed 13 June 2025).

[4] Rachel Baig, 'Boxer Imane Khelif targeted by hate speech, disinformation' (*DW*, 08 June 2024), available at https://www.dw.com/en/paris-olympics-boxer-imane-khelif-battles-hate-speech/a-69863650 (last accessed 14 June 2025).

2. Freedom of Expression, Fake News, and Hate Speech: Exploring Boundaries in a Connected World

'Expression' is the internal state of mind or intellect externalised through speech, writing, symbols, and actions.[5] It is a fundamental human freedom to express what one feels, believes, experiences, or wishes to share. Expression takes numerous forms, and there may be new forms of expression in the future that we have yet to discover. The answer to whether this right should be protected is affirmative, especially when considering that significant battles were fought and many lives were sacrificed to secure the right for individuals to express themselves. It is crucial to protect the right to freedom of expression, which should only be limited in exceptional circumstances.

Freedom of expression is a fundamental right and should be seen as a collective one, affecting both the sender and receiver of speech, as well as society as a whole.[6] Many studies characterise freedom of expression as a universal and natural right inherent to every human being, contributing to the establishment and smooth functioning of a democratic society.[7] Freedom of expression is a universal and foundational human right, recognised not only in international human rights law but also as a constitutional right in many democratic systems. It lies at the very core of the ECHR and serves as a cornerstone of democratic society, intimately connected to the purposes and spirit of the Convention itself. Its protection is presumed from the outset, forming the backbone of pluralism, open debate, and public accountability. Article 10 of the ECHR is structured into two paragraphs: the first defines the freedoms protected, namely the freedom to hold opinions, and to receive and impart information and ideas, without interference by public authorities and regardless of frontiers.[8] The second paragraph outlines the three conditions under which a State may legitimately restrict these freedoms.[9]

Article 10(2) of the ECHR makes clear that freedom of expression protects not only neutral or agreeable ideas, but also those that may offend, shock, or disturb. Through its interpretation, the European Court of Human Rights (ECtHR) emphasises that such freedom is essential to individual autonomy and democratic plural-

5 Alexander Brown, 'What Is Hate Speech? Part 1: The Myth of Hate' (2017) 36 *Law and Philosophy*.

6 Jonathan Seglow, 'Hate Speech, Dignity and Self-Respect' (2016) 19 *Ethical Theory and Moral Practice*.

7 Şener v. Turkey, App no 26680/95 (ECtHR, 18 July 2000).

8 The only exemption derives from par. 2.

9 *Aleksey Ovchinnikov v. Russia* Appl. no. 24061/04 (ECtHR, 16 December 2010).

ism. At the same time, the Court defines the freedom's limits, reminding us that even fundamental rights are not absolute and must be balanced against the rights of others and the interests of a democratic society.[10] In the light of the above, a very fine line separates the one from the other, often making the distinction unclear. At the same time, this paradox makes freedom of expression more appealing, as it does not imply that any provocative, offensive, or shocking expression should be automatically prohibited, but rather that it should be heard as it is intrinsically linked to a democratic society, individual freedoms, pluralism, tolerance, and open-mindedness.

The three-part test demands that any limitation to free speech must: a) be prescribed by law; b) pursue a legitimate aim; and c) be necessary in a democratic society, with the latter element requiring a nuanced proportionality assessment. Any limitation upon this right must be regarded as exceptional, justified only under strictly defined conditions, and applied with the utmost restraint.[11] On this basis, one could reasonably argue that it was appropriate to publicly address during the Olympics Khelif's gender identification or sexual identity, as there was a legitimate public interest in discussing whether, given her gender characteristics, she could compete fairly and equally against cisgender women. The exercise of the right to freedom of expression is far more complex than it may seem. Nevertheless, internet users often hastily form opinions and express themselves without fully processing the information or considering the legitimacy of the speech involved. This tendency underscores the challenges inherent in balancing the right to express oneself with the responsibility to do so thoughtfully and responsibly.

The right to freedom of expression has expanded into new media platforms in the wake of the technological revolution, transforming public discourse. As a result, discussions surrounding the legality of expression are no longer solely governed by human rights; they now also incorporate elements from new technologies and social networking. This intersection of legal frameworks reflects the evolving nature of communication in the digital age, where the parameters of expression are increasingly shaped by the complexities of modern media.[12] The right includes the freedom

[10] *Handyside v. the United Kingdom* Appl. No. 5493/72 (ECtHR, 7 December 1976).

[11] *Aleksey Ovchinnikov v. Russia* Appl. no. 24061/04 (ECtHR, 16 December 2010).

[12] Stella Mala*, The Legal Framework of Online Hate Speech* (*Το Νομικό πλαίσιο του Διαδικτυακού Μισαλλόδοξου Λόγου*) (Nicosia, Hippasus, 2023) (in Greek).

of speech and the formation and expression of opinions and ideas,[13] including online expression.[14]

Case law shows that forms of expression protected by the ECHR include documents,[15] radio broadcasts,[16] paintings,[17] films,[18] poetry,[19] artistic work,[20] novels,[21] electronic information systems,[22] and satirical expression.[23] The freedom to share information and ideas is inherently linked to the freedom to receive them, whether in print or broadcast media. Public information should be disseminated to foster dialogue that promotes research, questioning, and development. Restrictions on disseminating information should be proportional and justified,[24] aiming not to discourage the right itself, as such a result would be detrimental to States and the participatory interests of their citizens.

3. Bridging EU Law with International Obligations and the Convergence of Hate Speech and Fake News

The EU's 2008 Framework Decision defines hate speech as the intentional public incitement to violence or hatred directed against a group of persons or a member of a group identified based on race, colour, religion, descent, or national or ethnic origin, with the aforementioned offense being committed through the dissemination, by any means, of written material, images, or other elements.[25] This Framework Decision is binding on EU Member States (MS) and harmonises national laws.[26] It is recognised

[13] *Handyside v. the United Kingdom* Appl. No. 5493/72 (ECtHR, 7 December 1976) par. 49; *Erbakan v. Turkey* Appl. No 59405/00 (ECtHR, 6 July 2006) par. 56

[14] *Delfi AS v. Estonia* Appl. No. 64569/09 (ECtHR, 16 June 2015).

[15] *Handyside v. the United Kingdom* Appl. No. 5493/72 (ECtHR, 7 December 1976).

[16] *Groppera Radio AG and Others v. Switzerland*, Appl. No. 10890/94 (ECtHR, 28 December 1990).

[17] *Müller and Others v. Switzerland* Appl. No. 10737/84 (ECtHR, 24 May 1988).

[18] *Otto-Preminger-Institut v. Austria* Appl. No. 13470/87 (ECtHR, 20 September 1994).

[19] *Karataş v. Turkey* Appl. No. 23168/94 (ECtHR 8 July 1999).

[20] *Müller and Others v. Switzerland* Appl. No. 10737/84 (ECtHR 24 May 1988).

[21] *Akdaş v. Turkey* Appl. No. 41056/04 (ECtHR, 16 February 2010).

[22] *Eon v. France* Appl. No. 26118/10 (ECtHR 14 March 2013); *Kuliś and Różycki v. Poland* Appl. No 27209/03 (ECtHR 6 October 2009); *Alves da Silva v. Portugal* Appl. No. 41665/07(ECtHR, 20 October 2009).

[23] *Vereinigung Bildender* Künstler v. Austria Appl. No. (ECtCR, 25 January 2007).

[24] *Mouvement raëlien Suisse v. Switzerland* Appl. No. 16354/06 (ECtHR, 13 July 2012) par. 75.

[25] Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L328/55.

[26] See Cyprus: The Law on Combating Certain Forms and Expressions of Racism and Xenophobia through Criminal Law of 2011 (134(I)/2011), Criminal Code of Cyprus article 99A; see France: Law of 29

as a secondary source of European law and sets the minimum standards for criminalising hate speech, racism, and xenophobia across the EU. Following the ECHR and the 1997 Recommendation, [27] MS have also criminalised hate speech based on sexual orientation and gender identity.[28]

Its chronological precedence is precisely what makes the 1997 Recommendation one of the most well-known and widely cited attempts to define hate speech. Further, its broad definition that protects groups of people identified by sexual orientation and gender identity is well-known to EU MS, as they are parties to the ECHR. It has thus motivated these States to extend protections to groups identified by their sexual or gender identity. The ECtHR has frequently referenced the 1997 Recommendation in its decisions, underscoring its significance in shaping the legal understanding of hate speech.[29] On several occasions, the Court has drawn upon the 1997 Recommendation, further solidifying its role in guiding MS on the protection of groups based on sexual or gender identity.[30]

Complementing the definitional guidance offered in the 1997 Recommendation and the 2008 Framework Decision, the Rabat Plan of Action offers an internationally recognised framework that assists in assessing whether an expression qualifies as unlawful incitement to hatred, while safeguarding the right to freedom of expression.[31] The Rabat Plan was developed under the auspices of the United Nations (UN) and grounded in Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), which provides that: 'Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law'.[32] The Rabat Plan defines 'hatred' and 'hostility' as intense, irrational feelings

July 1881 on freedom of the press Article 32 and Law of 29 July 1881 on freedom of the press, Article 24; Germany: German Criminal Code (Strafgesetzbuch – StGB article 130.

[27] Recommendation (EC) R (97) 20 Recommendation of the Committee of Ministers to Member States on "Hate Speech", [1997] Committee of Ministers.

[28] See the French Law of 29 July 1881 on Freedom of the Press, Article 24 and Cyprus Criminal Code, Article 99A.

[29] Federica Casarosa, 'The European Regulatory Approach toward Hate Speech Online: The balance between efficient and effective protection' (2019) 55 *Gonzaga Journal of International Law*.

[30] *Carl Jóhann Lilliendahl v. Iceland*, Appl. no. 29297/18, (ECtHR, 12 May 2020); *Beizaras and Levickas v. Lithuania*, Appl.no. 412888/15, (ECtHR, 14 January 2020).

[31] United Nations High Commissioner for Human Rights, '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' (United Nations High Commissioner for Human Rights 2013) available at https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf.

[32] International Covenant on Civil and Political Rights (ICCPR), adopted 16 December 1966, UNGA Res

of opprobrium and aversion toward a targeted group; 'advocacy' implies an intention to publicly promote such attitudes; and 'incitement' refers to speech that creates an imminent risk of discrimination, hostility, or violence against individuals belonging to protected groups.

At the core of this framework lies a threshold test designed to determine when speech may lawfully be subject to restriction or criminal sanction. All six elements must be satisfied: a) the broader social and political environment is one in which the speech is likely to exacerbate discrimination, hostility, or violence; b) the speaker occupies a position of status, influence, or authority such that their words are capable of amplifying harm; c) the speech is made with the deliberate aim of inciting harm, rather than through negligence or recklessness; d) the language, tone, and structure of the message are inflammatory or provocative, making incitement more likely; e) the reach, audience, and means of communication are sufficiently extensive to increase the potential impact of the speech; and f) there is a real and imminent risk that the speech will directly result in discriminatory or violent action.

By introducing a rigorous and context-sensitive framework, the Rabat Plan ensures that limitations on speech are narrowly defined, necessary, and proportionate. This approach helps prevent arbitrary or overly broad restrictions while safeguarding both human dignity and freedom of expression. The Rabat test underscores the non-automatable nature of such assessments: While algorithms may support the detection of certain indicators such as speaker identity and dissemination metrics, the full application of the Rabat criteria, specifically those involving intent, context, and likelihood of harm, requires human, legal judgment. Thus, the Plan reinforces the importance of careful, reasoned adjudication in balancing the right to free expression with the need to combat hate speech. Defining hate speech in legal terms and applying the Rabat Plan are already complex tasks, and the rise of fake news further complicates the landscape, blurring the already fragile lines between expression and incitement, information and propaganda.

In recent years, online hate speech has grown,[33] occasionally taking the form of fake news (disinformation or misinformation).[34] This is not a new phenomenon—in

2200A (XXI), entered into force 23 March 1976, 999 UNTS 171.

[33] D. Madrid-Morales & H. Wasserman, 'Research Methods in Comparative Disinformation Studies' in Wasserman H and Madrid-Morales D (eds), *Disinformation in the Global South* (Wiley Blackwell, 2022) 41–57.

[34] Photios Spyropoulos, 'The Spread of False News in the Age of "Fake News"' (2019) 8 *Epistimonika Apotipomata*.

the 2012 *Raëlien Suisse* case,[35] the ECtHR addressed the issue of digital disinformation in relation to a poster campaign designed to convince people of the existence of extraterrestrials. The distribution of fake news might be a criminal offence under national laws.[36] It is not explicitly an offense under EU law but when a false information incites hatred, violence, xenophobia, and racism it can be prosecuted under existing laws.

A critical challenge in regulating harmful speech lies in the overgeneralisation of 'fake news', often commingling misinformation, disinformation, and malinformation without adequate legal precision. This lack of distinction risks both unjustified censorship and ineffective protection against genuinely harmful content. To enhance legal clarity, it is necessary to differentiate these categories based on intent, harm, and legal status even though certain categories of fake news generally fall outside the scope of legal scrutiny due to their inherently innocent character and absence of intent.

Misinformation refers to false or misleading information shared without malicious intent, such as erroneous reporting or satire.[37] For example, a satirical depiction of political figures, such as an image portraying the US president dancing with the Russian president, may be created as humorous commentary on current events rather than an attempt to deceive.[38] Such content, while factually inaccurate, lacks the deliberate intent to mislead or cause harm, distinguishing it from disinformation or incitement to hatred under legal frameworks. Such speech typically remains protected under freedom of expression as enshrined in Article 19 of the ICCPR. In contrast, disinformation involves deliberately false information intended to deceive and cause harm, for example, misleading propaganda that incites violence or discrimination. This type of expression may lawfully be restricted under Article 20(2) ICCPR. Malinformation, meanwhile, describes the use of truthful information out of context to harm individuals, such as the non-consensual disclosure of private data; these instances may fall under privacy or defamation laws.

---

[35] *Raëlien Suisse v. Switzerland*, Appl.no. 16354/06, (ECtHR, 13 July 2012).

[36] See for example the Criminal Code of Cyprus, Article 50.

[37] Jonathan Greenberg, *The Cambridge Introduction to Satire* (Cambridge University Press, 2019) 7 ('They all shape their judgments into an artistic form and blend attack with entertainment').

[38] *Müller and Others v. Switzerland* (Application No 10737/84) (ECtHR, 24 May 1988) (recognising that Article 10 protects artistic expression); see also *StraußKarikatur* (1 BvR 313/85) BVerfGE 75, 369 (Order of the First Senate, German Constitutional Court, 3 June 1987) ('satire can be art, but not all satire is art').

The legal landscape surrounding synthetic media has grown increasingly complex, particularly with the emergence of deepfakes—digitally fabricated or manipulated content, those present unique regulatory challenges. According to the Artificial Intelligence Act, 'deep fake' means AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places, entities, or events and would falsely appear to a person to be authentic or truthful.[39] For example, non-consensual intimate deepfakes are broadly condemned and typically fall under existing privacy and sexual offence laws in many jurisdictions. On the other hand, political deepfakes, especially those intended as parody or satire often remain protected by free speech guarantees, though they may be scrutinised under evolving rules aimed at safeguarding electoral integrity and media transparency.

This intricate regulatory terrain is reflected in international developments such as the Council of Europe's Recommendation CM/Rec (2022)16 on combating hate speech and the EU's Digital Services Act.[40] This is complemented by the leading international authorities, including UN Special Rapporteurs on freedom of expression and hate speech, who have underscored the need for a carefully balanced legal approach.[41] They call for frameworks that uphold fundamental rights while effectively addressing real harms. These experts stress the importance of applying clear legal definitions and maintaining high thresholds before imposing any restrictions on speech, particularly in politically or socially volatile contexts.

Ultimately, a clear and coherent legal framework that distinguishes between misinformation, disinformation, and malinformation is essential. Grounding these categories in international human rights principles and the latest normative guidance allows policymakers and courts to strike a fair balance: preserving freedom of expression while ensuring accountability where speech crosses the line into incitement, discrimination, or violence. Such clarity helps guard against arbitrary or disproportionate censorship, while promoting responsible and rights-respecting regulation.

---

[39] Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), [2024] OJ L ..., Art 3(60).

[40] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1.

[41] See, for example, UN Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, A/HRC/47/25 (13 April 2021); and UN General Assembly, *Report of the Special Rapporteur on the rights to freedom of peaceful assembly and of association*, A/74/486 (2019).

Despite the growing urgency to regulate harmful online content, a persistent difficulty lies in defining key terms such as *'misinformation'* and *'disinformation'*. The conceptual ambiguity surrounding these notions continues to challenge both lawmakers and courts. Nevertheless, certain institutional and legislative efforts have been undertaken to provide at least a functional framework or working definitions. For instance, there are some European policies, such as the *Code of Practice on Disinformation*[42] and the *Digital Services* Act,[43] which focus primarily on platform transparency and accountability rather than penalising fake news or attempting to offer a precise definition of it. In 2019, the ECtHR introduced the term *'fake news'* in *Brzezinski v. Poland*,[44] providing a broadly accepted, general definition encompassing both disinformation and misinformation. Current policy initiatives, while significant, remain insufficient, although the European Commission has made an effort to define the term *'disinformation':*

> [V]erifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. Public harm comprises threats to democratic political and policy-making processes as well as public goods such as the protection of EU citizens' health, the environment or security.[45]

Furthermore, the Commission clarifies that disinformation does not include reporting errors, satire, and parody, or clearly identified partisan news and commentary. In contrast to disinformation, misinformation[46] consists of false or misleading information that is shared without the intent to deceive or cause harm, and the person spreading it is not the originator of the content. Consequently, if someone mistakenly spreads fake news without intending to deceive or cause harm, this would be regarded as misinformation.[47] However, if the claim was made with malicious intent, it could be considered disinformation. Manipulation, rumours based on falsehoods,

[42] European Commission, 'Code of Practice on Disinformation' (Digital Strategy, 26 May 2021) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation accessed accessed 3 June 2025.

[43] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) [2022] OJ L277/1.

[44] *Brzeziński v. Poland*, Appl. no. 47542/07, (ECtHR, 25 July 2019).

[45] European Commission, 'Code of Practice on Disinformation' (2018) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation accessed on 10 June 2025.

[46] Philippe Jougleux, *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework* (Springer, 2022).

[47] European Parliament, *The Legal Framework to Address 'Fake News': Possible Policy Actions at the EU Level* (Policy Department for Economic, Scientific and Quality of Life Policies, 2018).

conspiracy theories, and misinformation are characteristics frequently found in both hate speech and fake news.[48] Fake news is closely linked to the offense of hate speech; both directly impact the right to information and both constitute a danger to democratic society.[49]

Malicious AI-generated content, such as deepfakes, presents a clear threat to the smooth functioning of democratic States.[50] Strengthening independent and reliable media outlets is a crucial safeguard against such threats. Free and pluralistic media are a key pillar of democracy and essential for a healthy market economy. The EU adopted the European Media Freedom Act (EMFA) (Regulation (EU) 2024/1083),[51] which entered into force on 7 May 2024. Its major provisions are already in effect, with full application across the EU commencing on 8 August 2025. The Act aims to protect media freedom and pluralism, ensure the cross-border operation of both public and private outlets without undue pressure, and address the challenges posed by the digital transformation of the media sector.

The EMFA seeks to standardise national laws across the EU regarding editorial freedom, media pluralism, and independence, addressing the challenges posed by digital transformation. It is widely accepted that the media must have a strong voice to inform citizens with integrity about current affairs. This 'asylum' status ensures journalists' voices are not weakened.[52] In the same vein, journalism should be recognised as a crucial profession that contributes to the establishment, development, and smooth functioning of a democracy.[53] The main goals of the EMFA are to ensure the sustainability of media outlets, strengthen democratic engagement, combat disinformation, and protect media freedom and pluralism. The Act also addresses concerns

[48] Bente Kalsnes, 'Fake News' (2018) *Oxford Research Encyclopedia of Communication*.

[49] Philippe Jougleux, *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework* (Springer, 2022).

[50] S Rayhan & S Rayhan, 'The Role of AI in Democratic Systems: Implications for Privacy, Security, and Political Manipulation' (MSC thesis, 2023); J. Twomey & al., 'Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine', (2023) 18(10) *Plos One*.

[51] European Media Freedom Act, Regulation (EU) 2024/1083 of the European Parliament and of the Council of 11 April 2024 on safeguarding media freedom in the European Union, OJ L, 7 May 2024, p. 1, available at https://eur-lex.europa.eu/eli/reg/2024/1083/oj (last accessed 28 August 2025).

[52] http://ec.europa.eu/information_society/newsroom/image/document/2016-50/2016-fundamental-colloquium-conclusions_40602.pdf last accessed on 28 August 2024.

[53] *Editorial Board of Pravoye Delo and Shtekel v. Ukraine* Appl. No. 33014/05 (ECtHR, 5 May 2011); *Times Newspapers Ltd v. the United Kingdom,* Appli. No. 3002/03 Appl. No 23676/03 (ECtHR, 10 March 2009).

about political bias, transparency in media ownership, and the allocation of State advertising, with the aim of preventing political interference in the media and safeguarding the rights of journalists and their sources.

Speech is not acceptable by default—hate speech is a clear legal limit on the right to freedom of expression. It is important to differentiate between cases where hate speech targets a protected group or an individual from that group and instances where hate speech is driven by false information or where fake news is deliberately created to incite prejudice against a group or an individual of the protected group. Although it may initially seem that hate speech and fake news are closely linked, as both can disseminate falsehoods about an individual or protected group, this is not always the case. There are situations where hate speech is propagated through lies, misinformation, and intentional distortion of the truth to provoke hatred against a protected group. It is crucial to establish whether the fake news was disseminated with intent, which would categorise it as disinformation, or without intent or unintentional inaccuracies, which would classify it as misinformation. Disinformation requires proof of deliberate intention by the sender, whereas misinformation does not.

According to the theoretical framework of disinformation, if someone intentionally spreads a fake story with hate speech, this act qualifies as both disinformation and an offense of hate speech.[54] Conversely, if an individual genuinely believes a false news story to be true and spreads it with hate speech, this situation is classified as misinformation, though it still constitutes hate speech. A separate scenario occurs when someone disseminates fake news, intentionally or unintentionally, without incorporating hate speech but for financial gain. In this case, the fake news itself does not necessarily incite hate or violence. If this news is later republished with hate speech commentary, the situation becomes more complex, as the original disseminator is not accountable for the subsequent actions of others. Therefore, it is crucial to distinguish between disinformation and hate speech to accurately evaluate the nature of the offense, if any. Fake news creates considerable confusion and challenges for end users.[55] Each individual is responsible for their own actions and intentions, and it is important to recognise the distinction between the acts of the sender and those of the receiver or subsequent disseminators of information. While the sender of

[54] E. Humprecht, F. Esser, F & P. Van Aelst, 'Resilience to Online Disinformation: A Framework for Cross-National Comparative Research' (2020) 25(3) *International Journal of Press/Politics*, 493–516.

[55] European Parliament, Regulation of the European Parliament and of the Council on the European Media Freedom Act (2018), available at https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/619013/IPOL_IDA(2018)619013_EN.pdf (last accessed 29 May 2025).

information may be held accountable for their intentions and the content they create or spread, the receiver or further disseminator can also be responsible for how they handle and share that information. This distinction ensures that each party's role in the dissemination of information is evaluated appropriately, taking into account their specific actions and intentions.

4. The ECHR's Stance on Fake News and Hate Speech

Fake news that includes hate speech does not enjoy the protection offered by Article 10 of the ECHR, which safeguards freedom of expression.[56] There are limitations and responsibilities associated with the rights of others, national security, public safety, prevention of disorder or crime, or the protection of health or morals. In the case of fake news, especially if it leads to harm, incites violence or hatred, or spreads disinformation that could cause significant public harm, authorities may justify restrictions under Article 10(2). The ECtHR has consistently upheld that while free expression is crucial, it does not extend to protecting false information that can cause significant harm or threaten public order. Unlike other forms of expression, hate speech is subject to stricter limitations under Article 10 of the ECHR. While the right to free expression is fundamental, the ECHR recognises that this right does not extend to speech that incites hatred, violence, or discrimination, and allows States to impose restrictions on such speech to protect the rights and safety of others.

Therefore, while some forms of expression, even controversial or offensive ones, are protected, fake news that crosses certain thresholds such as incitement to violence, hate speech, or causing harm may not be protected under Article 10. The judicial approach to disinformation includes recognising the high level of protection afforded to value judgments and personal opinions under freedom of expression. According to ECtHR case law, such opinions are less susceptible to proof and should be protected more robustly than false factual assertions.[57] This distinction between facts and value judgments is particularly relevant in cases of misinformation, where false information is shared unknowingly, as opposed to disinformation, which involves

[56] Article 11(1) of the Charter of Fundamental Rights of the European Union (2000/C 364/01) recognises the freedom of expression and information: 'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers'.

[57] *Lingens v. Austria*, Appl. no. 9815/82, (ECtHR 8 July 1986).

intentional deceit.[58] The protection of opinions, therefore, becomes crucial when assessing the culpability of individuals sharing misinformation.

The ECtHR has emphasised that governments cannot suppress speech simply because it challenges official views;[59] minority opinions should be protected particularly in ongoing debates on unresolved public issues. The ECtHR has also held that ECHR Article 10 does not forbid the discussion of information even if its truthfulness is dubious. To balance freedom of speech with the right to accurate information, the focus should be on promoting responsible communication, encouraging media pluralism, and discouraging users from sharing unverified content.[60]

The ECtHR determines on a case-by-case basis whether there is a legitimate reason for restricting speech, assessing the potential harm. There have been instances where the Court has dismissed hate speech cases as unfounded, invoking Article 17 (the abuse clause) to emphasise that the appeal itself undermines the principles of the ECHR. However, the abuse clause should not be overused to combat disinformation, as this can erode fundamental speech protections. It is also important to recognise that not all disinformation is illegal under domestic or EU law. Therefore, policymakers must carefully balance the restriction of disinformation with the right to freedom of expression.

On 27 August 2024, the ECtHR issued a landmark decision regarding fake news.[61] The Court held that Article 10 of the ECHR, which protects the right to freedom of expression, does not extend to the dissemination of scientifically unfounded opinions about coronavirus vaccines.[62] Klaus Bielau, an Austrian physician, was disciplined by the Austrian Medical Association for promoting baseless anti-vaccine claims, such as denying the existence of pathogens and the effectiveness of vaccines. After unsuccessful cases before the national courts, Bielau brought his case to the ECtHR, alleging a violation of his freedom of expression.

After weighing the various rights involved, the ECtHR held that while doctors have the right to participate in public health debates, including expressing critical and minority views, this freedom is not unlimited. The Court emphasised that restric-

[58] European Council, Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making (2017) https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c accessed 30 August 2024

[59] *Sergey Petrovich Salov v. Ukraine,* Appl. no. 65518/01, (ECtHR, 27 April 2004).

[60] Registrar of the European Court of Human Rights (ECtHR), Press Release: ECHR 370 (2013), EUR. COURT OF HUMAN RIGHTS (December 17, 2013).

[61] *Bielau v. Austria*, Appl. no. 20007/22, (ECtHR, 27 August 2024).

[62] Votes 6/7.

tions on freedom of expression may be necessary when false and categorical public statements are made about medical matters, particularly when such statements are published online. The decision further noted that the €2,000 fine imposed on Bielau was modest, considering the potential harm caused by his statements. The Court's decision reinforced that free speech has limits, especially when public health is at stake and when the issues are serious enough to affect human life. Six out of seven judges voted against finding a violation of Article 10.

In a dissenting opinion, one judge held that there was no violation of Article 10, which protects freedom of expression. The judge argued that the applicant's published article, which encouraged readers to question conventional medical practices, should not be dismissed as unreasonable. The article was intended for a specific audience already open to alternative medicine and did not have a significant public reach. The judge also highlighted a similar case, *Stambuk v. Germany*, where the Court found a violation of Article 10 in a situation involving advertising by a medical practitioner.[63] The judge concluded that the restrictions imposed on the applicant were disproportionate and constituted a form of censorship, which could deter future expression of opinions, thereby threatening democratic society.

5. The Dual Threat: Why Fake News Coupled with Hate Speech Matters

False and misleading information, in its myriad forms, is profoundly disturbing and carries far-reaching consequences.[64] It can foster widespread misconceptions, incite mass unrest, and fuel resistance to accurate knowledge, which collectively pose significant threats to individual well-being, societal cohesion, and the integrity of democratic processes.[65] Furthermore, the proliferation of fake news can engender bigoted rhetoric and generate fear, potentially resulting in hate crimes due to the disarray and confusion it causes. Of particular concern is the dissemination of false information concerning critical matters such as public health, evident during the coronavirus pandemic, which can undermine trust in health authorities and exacerbate public health crises.[66] Similarly, the deliberate distortion of facts to manipulate election out-

[63] *Stambuk v. Germany*, App no 37928/97 (ECtHR, 17 October 2002, Third Section),

[64] Rayhan & Rayhan (no 50).

[65] S. Rosenfeld, *Democracy and Truth: A Short History* (University of Pennsylvania Press, 2019); J. Strömbäck & al., *Knowledge Resistance in High-Choice Information Environments* (Routledge, 2022).

[66] A. Lazić & I. Žeželj, 'A Systematic Review of Narrative Interventions: Lessons for Countering Anti-Vac-

comes, overthrow governments and create disorder,[67] shape public attitudes,[68] justify wars,[69] or threaten environmental stability represents an even more egregious use of misinformation, with potentially devastating consequences for democratic institutions and global stability.

6. Algorithmic Transparency and Accountability

The rapid growth of social media and digital platforms has led to a surge in user-generated content (UGC), including illegal material. Intermediary liability regimes differ significantly across jurisdictions, revealing deep normative tensions. Through Section 230 of the 1960 Communications Decency Act, the US provides extensive immunity to platforms for third-party content,[70] prioritising free expression and limiting State interference. This section, referred to as 'the Good Samaritan' protection, remains a cornerstone of internet regulation in the US, granting broad immunity to online platforms for UGC. While it has played a critical role in enabling the internet's rapid expansion, legal scholars increasingly point to its inadequacies in addressing contemporary harms such as disinformation and hate speech. As Dickinson notes, courts have constructed an expansive immunity doctrine that has struggled to adapt to evolving technologies and societal challenges, thereby protecting even bad-faith actors and impeding meaningful regulatory reform. With the US Supreme Court recently considering *Gonzalez v. Google LLC*, its interpretation of Section 230 may provide an opportunity to align legal protections with modern expectations of accountability and platform responsibility.[71]

---

cination Conspiracy Theories and Misinformation' (2021) 30(6) *Public Understanding of Science* 644–670, available at https://doi.org/10.1177/09636625211011881.

[67] M. Spring, 'Sadiq Khan says fake AI audio of him nearly led to serious disorder' (*BBC News*, 14 February 2024), available at https://www.bbc.com/news/uk-68146053, last accessed 13 June 2025.

[68] Oxford Internet Institute, 'Social Media Manipulation by Political Actors: An Industrial Scale Problem' (University of Oxford, 2021) available at https://www.oii.ox.ac.uk/publications/social-media-manipulation-by-political-actors-an-industrial-scale-problem/ last accessed 30 August 2024.

[69] E. Smalley, 'Russia's False Claims About Biological Weapons in Ukraine Demonstrate the Dangers of Disinformation and How Hard It Is to Counter – 4 Essential Reads' *The Conversation* (2022)

[70] 1 47 USC § 230 (Communications Decency Act); see also Gregory M Dickinson, 'Section 230: A Juridical History' (2025) 28 *Stan Tech L Rev* 1.

[71] *Gonzalez v. Google LLC*, 598 U.S. ___ (2023), available at https://www.supremecourt.gov/opinions/22pdf/21-1333_6jfm.pdf

Unlike the US, which tends to take a hands-off approach, the EU follows a much more regulated path. Under the Directive on electronic commerce,[72] platforms are exempt from liability if they can show that they were genuinely unaware of the illegal content disseminated through their services, and if they act expeditiously to remove or disable access once they become aware of it. With the introduction of the Digital Services Act (DSA),[73] however, the EU now places stronger responsibilities on major platforms, including conducting risk assessments and being more transparent about how they operate. The approach has also been shaped by the case law of the European Court of Human Rights, most notably in *Delfi AS v. Estonia*, where the Court upheld liability for a news portal over user-generated comments, emphasizing the balance between freedom of expression and the protection of rights from harmful online content.[74]

Meanwhile, the UN is pushing for a human rights-centred approach through its Special Rapporteurs and Guiding Principles on Business and Human Rights.[75] This means encouraging platforms to respect freedom of expression, while also protecting users from harm. These different approaches—the US's more permissive model, the EU's regulatory oversight, and the UN's emphasis on rights—can sometimes clash. This makes it harder to create clear, global standards for holding platforms accountable. The DSA does not replace the old Directive on electronic commerce but instead builds on it to tackle today's digital challenges. It reflects growing pressure on online platforms to do more when it comes to illegal content. And increasingly, those efforts rely on AI tools to help monitor and manage what is happening online.[76]

---

[72] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178/1, available at https://eur-lex.europa.eu/eli/dir/2000/31/oj (last accessed 28 August 2025). *Delfi AS v. Estonia* App no 64569/09 (ECtHR, 16 June 2015); *Sánchez v. Spain* App no 45532/20 (ECtHR, 6 February 2024).

[73] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1, available at https://eur-lex.europa.eu/eli/reg/2022/2065/oj (last accessed 29 August 2025).

[74] *Delfi AS v. Estonia* App no 64569/09 (ECtHR, 16 June 2015); *Sánchez v. Spain* App no 45532/20 (ECtHR, 6 February 2024).

[75] United Nations Human Rights Council, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (UN Doc HR/PUB/11/04, 2011) available at https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (last accessed 30 May 2025).

[76] Rayhan & Rayhan (no 50).

Understanding how the algorithm functions is crucial for establishing the legal framework concerning the social media platform's liability.[77] Algorithms are often perceived as 'black boxes' because their opaque decision-making processes regarding content legitimacy.[78] Consequently, it is often unclear why certain content is (not) flagged as hate speech. Furthermore, audit trails are considered critical, as regular audits of algorithms ensure they are functioning as intended. Establishing a framework for these regular audits is essential. The algorithm functions like a 'surgeon' for the internet, receiving notices or identifying illegal content in their platform and removing illegal content, much like a 'carcinoma'.[79] Such content threatens to spread and 'infect' the entire online environment. With hate speech, time exacerbates the damage it inflicts on the individual.

Furthermore, the internet, as a global village, allows information to spread in all directions, leading to the bubble phenomenon. The time required for a human content operator to perceive, judge, and act must be immeasurably swift.[80] In recent years, there has been a noticeable coexistence of strict hard law and soft law. So, while the regulation of hate speech on social media is governed by European legal frameworks, there is a simultaneous growing trend toward developing non-binding, 'soft' rules for the operation of social networking services. The soft law frameworks governing social media platforms focus on reducing the spread of false information by limiting its visibility and promoting accurate sources. At the same time, these platforms are legally obligated to remove content that crosses the line into illegality, such as when false information also constitutes hate speech. In such cases, the content is not only misleading but also harmful, necessitating its removal under both voluntary guidelines and binding legal obligations.

[77] FRA, 'Bias in Algorithms –Artificial Intelligence and Discrimination' (2023) available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf (last accessed 18 August 2024).

[78] Erwan Le Merrer & Gilles Trédan, 'What is a black box algorithm?: Tractatus of algorithmic decision-making' (2023) *HAL* ffhal-03940259f.

[79] The Digital Services Act (DSA) establishes an EU-wide framework for detecting, flagging, and removing illegal content, along with new risk assessment obligations for large online platforms and search engines to identify how such content spreads. What qualifies as illegal content is not determined by the DSA, but by existing EU or national laws—for instance, terrorist content, child sexual abuse material, or illegal hate speech are defined at the EU level. If content is illegal only in a specific MS, it should generally be removed only within that State's territory.

[80] Digital Services Act 2022 (EU) Regulation 2022/2065, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065 (last accessed 26 May 2025).

Within the broader discourse on algorithmic transparency and accountability, the Rabat Plan of Action highlights the indispensable role of human judgment in assessing harmful speech. While algorithms may detect surface-level indicators, such as speaker identity or dissemination reach, they cannot adequately evaluate nuanced elements like intent, context, and likelihood of harm. These inherently legal and interpretive assessments require human oversight. As such, the Plan reinforces the need for transparent, accountable systems that preserve fundamental rights. In an environment increasingly shaped by AI-driven content and fake news, such safeguards are more critical than ever to prevent the erosion of legal standards and protect both freedom of expression and public safety.

The European Commission has collaborated with social media companies to address hate speech online, treating them like other media channels, through the Code of Conduct.[81] Accordingly, IT companies should review most valid notifications for the removal of illegal hate speech within 24 hours and, if necessary, remove or disable access to such content. The Directive on electronic commerce prompted the creation of take-down procedures, though it does not provide detailed regulation.[82] These 'notice and action' procedures start when someone alerts a hosting service provider, like a social network or e-commerce platform, about illegal content such as racist or abusive material. The process concludes when the provider takes action regarding the content. For this reason, online human content operators may struggle to keep pace with the rapid technological advancements of the internet. The responsibility of the provider within the global village is immense.

Imagine a deep fake combined with hate speech which is one of the most dangerous forms of disinformation where advanced technology is used to impersonate someone and spread false, hateful views in their name. This misuse of identity can have serious repercussions, not only damaging the individual's reputation but also posing broader risks to society. Such actions can contribute to the spread of harmful fake news, incite violence, or fuel discrimination. The role of algorithms in this

[81] European Commission, 'Code of conduct for combating the online illegal hate speech', available at http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300; K. Podstawa, 'Hybrid Governance or… Nothing? The EU Code of Conduct on Combatting Illegal Hate Speech Online', *Use and Misuse of New Technologies* (Springer, 2019).

[82] Article 14 of the European Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on Electronic Commerce').

context is crucial, as they often dictate how widely and quickly this illegal content is disseminated.

Effective control over algorithmic systems requires more than proper design, and monitoring demands a multilayered approach involving both technical and regulatory oversight. Technical controls such as code audits, testing frameworks, and formal verification methods help ensure that algorithms function as intended. At the same time, regulatory mechanisms, including government oversight, industry standards, and impact assessments, provide external accountability. Yet, the inherent complexity of modern algorithms presents persistent challenges: they can produce emergent behaviours, operate across legal jurisdictions, and evolve more rapidly than oversight frameworks can adapt. Additionally, commercial confidentiality often restricts independent review, further complicating efforts to establish comprehensive governance. While no control mechanism is foolproof, combining robust technical safeguards with adaptive regulatory strategies remains essential in managing the risks associated with complex algorithmic systems.

The UK has proposed legislation to criminalise the creation and distribution of deepfakes, especially those involving sexually manipulated images, as part of a broader effort to combat harmful online content.[83] There is a big difference between fake news that involves the spread of false information affecting public opinion, fake news that involves spreading of hate speech, and deepfakes, which use AI-generated media to manipulate individuals' images, often with prejudice for malicious purposes like privacy violations and spreading of hate speech. The proposed penalties reflect the serious threat these forms of disinformation pose to individuals and society. Deepfake content and hate speech warrant zero tolerance and should be removed immediately from the digital space. As the ECtHR argued:

> The risk of harm posed by content and communications on the Internet to the exercise and enjoyment of human rights and freedoms, particularly the right to respect private life, is certainly higher than that posed by the press. Therefore, the policies governing reproduction of material from the printed media and the Internet may differ. The latter, undeniably, have to be adjusted according to the technology's specific features in order to secure the protection and promotion of the rights and freedoms concerned.[84]

[83] UK Government, 'Government Cracks Down on Deepfakes Creation' (GOV.UK, 30 November 2022) available at https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation (last accessed 12 September 2024).

[84] *Editorial Board of PravoyeDelo and Shtekel v. Ukraine*, Application no. 33014/05, (ECtHR, 5 May

7. The Jurisprudential Development of the Responsibility of the Internet Provider

Communication via the internet has evolved significantly over the past two decades. Initially used mainly by researchers to share messages and information, the internet eventually became accessible to the public, leading to a proliferation of websites hosted by internet service providers (ISPs). Some of these websites contained unlawful content, prompting rightsholders to seek liability from ISPs. ISPs claimed they were mere intermediaries and could not control their subscribers' content. Rightsholders argued that ISPs often benefited from infringing activities and should be held accountable. To address these issues, 'safe harbour' protections were introduced, shielding ISPs from liability for their subscribers' illegal actions. In Europe, the Directive on electronic commerce provides broad immunity for online service providers, although rightsholders can still seek judicial relief to stop unlawful behaviour or gather information on infringers.

Shtekel had previously established the principle of immunity status for ISPs: 'No one can be held responsible for online content that they did not author, unless they either accepted it as their own or refused to comply with a court order for its removal'.[85] This principle is no longer applicable today, as it would imply ISPs are covered by a regime of immunity and limited liability.

Delfi began by highlighting the need to establish a lack of preventive liability and direct violation in order to hold the intermediary accountable.[86] It involved an online newspaper with significant reach and visibility in Estonia and Russian-speaking countries, publishing hundreds of articles daily. Readers could comment on these articles anonymously, leading to the publication of extreme, intolerant, threatening, abusive, and defamatory comments. The newspaper allowed readers to express themselves directly through options like 'Add comment', 'Post comment', and 'Read comment'. It made it clear in its operating rules that the responsibility for comments lay with their creators, not the provider. Readers could report comments using a notification button, and the provider could then delete the content if deemed illegal. This approach was fully compliant with both European and national laws. A shipping company requested the removal of extreme comments against it and sought monetary compensation. While the provider removed the comments, the request for com-

2011).

[85] Ibid.

[86] *Delfi AS v. Estonia*, Appl. No.64569/09, (ECtHR 10 October 2013)

pensation was denied. A prolonged legal dispute culminated in the ECtHR, which was tasked with determining whether the right to freedom of expression had been violated and assessing the responsibility of the internet provider involved. The ECtHR examined a potential violation of Article 10 of the ECHR by applying a three-stage test to assess the lawfulness of the restriction on freedom of expression. The court concluded that there was no violation.

The *Delfi* judgment established new jurisprudence by confirming that an ISP could, in certain circumstances, be held liable for illegal content on its service, despite not being the original creator. This liability is based on the provider's exclusive control over its service, the inability of the victim to prevent harm, and the financial benefits the provider accrued from the content until its removal was mandated. The ECtHR's case-by-case approach evaluates the reason for the restriction, the conduct of the provider, the financial benefit derived, the harm inflicted, and the disadvantage suffered by the complainant. This rationale was later affirmed by the ECtHR Plenary, which ruled that speech containing illegal content is not protected under Article 10 of the ECHR.[87] Consequently, an ISP may be liable for such content on its service, even if it did not create the content, if the provider could have controlled the service but failed to do so, thereby establishing culpability and a causal link between the provider's inaction and the resultant harm.

In its decision, the ECtHR established four criteria for evaluating a platform's liability for user comments: a) the context of the comments; b) the steps taken to prevent or remove unlawful comments; c) the feasibility of holding the actual authors accountable; and d) the consequences of the domestic ruling for the company. The court found Delfi AS had ultimate control over the comments and profited from them, and its measures to delete hateful comments were insufficient. Anonymity on the platform made it impossible to hold actual authors accountable. The fine imposed on Delfi AS was minimal and not considered disproportionate and thus did not violate freedom of expression. The ECtHR underlined that platforms like Delfi AS, which exercise control over and derive profit from user-generated comments, can be held liable for such content.[88]

The *Magyar* case further illuminated the reasoning in *Delfi* and pointed out the circumstances in which a departure is warranted.[89] In the present case, although the

[87] *Delfi v. Estonia*, Appl. No.64569/09, (ECtHR, Plenary session 16 June 2015).

[88] Ibid. paras 142–143.

[89] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*, Appl. No. 22947/13, (ECtHR, 2 May 2016).

online comments contained illegal content, they did not amount to intolerant speech such as hate speech or incitement to violence. Instead, the Court examined the responsibility of the ISP in relation to defamatory remarks. The applicants, who were self-regulated ISPs operating a news portal, were required to remove readers' comments after domestic courts judged them to be defamatory and unlawful. This case is important for the responsibility of the ISP but also for the protection of the right to freedom of expression. The case ruled that the ISP is not liable if it hosts vulgar or offensive speech on its service. Furthermore, the responsibility of the ISP could not be established due to the lack of economic interest factors, which justified the divergence from *Delfi*'s reasoning.

The ECtHR has consistently expressed scepticism towards the notice-and-action principle. In the 2013 *Delfi* case, reaffirmed by the Grand Chamber in 2015, the Court ruled that holding a platform liable for third-party comments does not necessarily violate freedom of expression under the ECHR, even if the platform has a notice-and-action system in place. The Court emphasised that States have broad discretion in balancing privacy rights (ECHR Article 8) and freedom of expression (ECHR Article 10) and would require strong reasons to override national courts' decisions. The ECtHR stated that countries could impose liability on platforms that fail to promptly remove clearly unlawful comments, such as hate speech and direct threats, even without notification from victims or third parties.

Eight years later, the Court has strengthened its stance, suggesting that States have a positive obligation to penalise platforms that do not proactively remove hate speech. In *Zöchling*, an Austrian news portal published an article leading to death threats and insults against journalist Christa Zöchling.[90] Although the platform quickly deleted the comments upon request and blocked the users, the ECtHR criticised the lack of a notice-and-action system and the absence of automatic filtering measures. The Court argued that the platform could have foreseen the offensive comments, given past experiences, and found that the lack of balancing of competing interests violated the State's procedural obligations under ECHR Article 8.

In the case of *Sanchez v. France*,[91] the Grand Chamber of the ECtHR rejected Julien Sanchez's claim against France, where he argued that his criminal conviction for not removing hateful comments from his Facebook page violated Article 10 of

[90] *Zöchling v. Austria*, Appl. no. 4222/18, (ECtHR, 5 September 2023).

[91] *Sanchez v. France (Grand Chamber)*, App no 45581/15 (ECtHR, 15 May 2023), available at https://hudoc.echr.coe.int/eng?i=003-7648098-10537594 (last accessed 29 August 2025).

the ECHR. Sanchez, a French politician, was fined for failing to delete third-party comments that were discriminatory and incited hatred against Muslims. He contended that this fine unfairly burdened his freedom of expression by requiring constant monitoring of his public Facebook page. The ECtHR's Fifth Section ruled that his conviction did not violate Article 10, as the comments were clearly unlawful and his inaction towards them warranted the penalty. The Grand Chamber upheld this decision, stating that the interference with Sanchez's freedom of expression was lawful, necessary in a democratic society, and pursued a legitimate aim, emphasising his greater duty to manage hateful comments as a politician.

The Court suggested that a minimum degree of moderation or automatic filtering is desirable to quickly identify unlawful comments, a stance reiterated in *Zöchling*. This position indicates a lack of awareness of the controversies surrounding filter systems and an uncritical view of AI in handling complex human issues. Additionally, the ECtHR's decision may imply that States have a positive obligation to require platforms to monitor for unlawful content, conflicting with the EU's legal framework, as reinforced by the DSA. This issue is not about what types of content should be allowed online, but rather the timing and automation of content management.

In *Pătraşcu v. Romania*, the applicant was held liable by domestic courts for offensive third-party comments posted under his Facebook post, which criticised the Bucharest National Opera. The ECtHR found this to be a violation of his right to freedom of expression under ECHR Article 10. Unlike previous rulings such as *Delfi* and *Sánchez*, where the Court accepted liability under specific conditions, here it emphasised that holding a private individual responsible for third-party content—without clear legal standards—was disproportionate. The decision reflects a notable shift in the Court's reasoning, acknowledging that ordinary users do not have the same editorial capacity as media outlets or public figures. Imposing liability in such cases risks encouraging self-censorship and undermining meaningful public debate, particularly on matters of public interest.

8. Elevating Accountability: The Expanded Responsibilities of ISPs

Under the Directive on electronic commerce,[92] human content operators are not liable for information transmitted or stored when performing specific activities such

[92] Directive on electronic commerce, supra note 77.

as mere conduit, caching, and hosting. Mere conduit refers to providing unfiltered internet access, caching involves temporarily storing information to improve transmission efficiency, and hosting pertains to storing information like websites on ISP servers. The directive does not cover hyperlinking, leading to initial national court exemptions, which were later considered potentially infringing. The Court of Justice of the European Union (CJEU) ultimately ruled that hyperlinking is not infringing unless it links to infringing material with actual or constructive knowledge of its illegality. Human content operators have no obligation to monitor stored information or seek out illegal activities. However, hosting human content operators must not have actual or apparent knowledge of illegal activity and must act promptly to remove such content upon gaining knowledge. Unlike US law, European legislation does not require a human content operator to control unlawful activities or financially benefit from them to claim immunity, though case law has affirmed these principles.

The role of the human content operator has evolved with technological advancements, expanding from simple hosting to complex platforms like social media, requiring different regulations for liability and content management. The online service provider now carries an increased level of responsibility for their platform. Under the DSA,[93] when a complaint (notice) is submitted, the provider is made aware that illegal content exists on their service; they must assess it and, if necessary, remove it expeditiously. Notably, liability is not triggered by the mere submission of a complaint, but arises once the provider has actual knowledge of illegal content and fails to act—a standard that has been criticised as stringent.

Platforms are required to act 'expeditiously' to remove or disable access to illegal content once they become aware of it, as outlined in Article 16 of the DSA. This provision establishes the notice-and-action mechanism, mandating that platforms put in place procedures for handling notifications and act without undue delay. While certain EU laws impose specific deadlines such as the one-hour removal requirement for terrorist content under the 2021 Terrorist Content Online Regulation,[94] the DSA and the Directive on electronic commerce rely more generally on the standard of acting swiftly, without prescribing strict time limits for hate speech or disinformation. In this respect, the EU Code of Conduct on Countering Illegal Hate Speech Online plays a complementary role: though not legally binding, it recommends a 24-hour time-

[93] Regulation (EU) 2022/2065, supra note 78

[94] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L 172/79, article 3(3).

frame and reflects the EU's expectations for rapid responses by social media companies. Since its adoption, successive evaluations show that major IT firms, including TikTok, have significantly improved their removal times for racist and xenophobic content.

9. Combating Digital Deception and Harmful Content by Penalising Deepfakes and the Chilling Effect on Freedom of Expression

There are good reasons to worry about fake news—producing and spreading disinformation online is much cheaper and easier due to the availability of a platformised, digital, end-to-end infrastructure for information exchange. What does this mean? It means that anyone, without cost, immediately, and under the cloak of anonymity, can easily take actions that range from something as simple as defrauding a person, to causing much more terrifying outcomes. These include inciting hatred and hate crimes, threatening peace, public order, public health, and democracy, spreading false propaganda, and using sexually explicit deepfakes. As mentioned, a deepfake can be an AI-generated video, audio, or image that mimics real individuals or events with striking realism, using techniques like deep learning and generative adversarial networks. Deepfakes can pose serious risks, such as spreading fake news and/or hate speech, enabling identity theft, and creating non-consensual explicit content, making it increasingly difficult to distinguish between authentic and manipulated media.

A critical question arises as to whether fake news should carry a criminal penalty.[95] In Cyprus, fake news is criminalised with a penalty of up to three years in prison and this provision of Cypriot criminal law should be promptly revised to reflect contemporary formats of disinformation. The provision states:

> Publication of fake news, etc.
>
> 50.-(1) Whoever publishes, in any form, false news or information that may undermine public order or public confidence in the state or its institutions, cause fear or anxiety in the public, or violate public peace and order in any way, is guilty of a misdemeanor and is punishable by imprisonment not exceeding two years, a fine not exceeding one thousand five hundred pounds, or both

[95] Philenews, 'September to the House of Representatives: Amendment for Criminalization of Dissemination of False News (26 June 2024) available at https://www.philenews.com/kipros/koinonia/article/1486457/septemvrio-sti-bouli-tropologia-gia-pinikopiisi-exivrisis-psevdon-idiseon/ (last accessed 2 September 2024).

> penalties.... Provided that it is a defense to the accused if they can prove to the satisfaction of the Court that the publication was made in good faith and was based on facts justifying such publication.[96]

Meanwhile, the UK has proposed legislation to criminalise the creation and distribution of deepfakes, especially those involving sexually manipulated images, as part of a broader effort to combat harmful online content.[97] The criminalisation of fake news, which may be justified in extreme cases such as fraudulent deepfakes, sexually explicit deepfakes, or content threatening democracy, public safety, and health, also carries a chilling effect on freedom of expression. Recognising the chilling effect as an independent theory is important in light of the ECtHR's growing emphasis on factors that undermine free speech. Even when restrictions are deemed legitimate after passing the three-part test, there remains an inherent suspicion of censorship. This underscores the need to treat the chilling effect as an additional control mechanism, since even minimal censorship can deter lawful expression.

Not all these issues warrant censorship or efforts to block information. Requiring internet intermediaries to filter out non-mainstream or non-fact-based opinions could significantly impoverish our democracy and society. The key to balancing freedom of expression with the right to accurate information lies in: a) promoting responsible information-sharing practices; b) implementing proactive media policies that encourage pluralism and diversify content exposure; c) enhancing media literacy and supporting user behaviour through educational initiatives; and d) addressing extreme and dangerous online falsehoods under a special regulatory regime to mitigate their potential extremely serious harm. The mere existence of a criminal penalty undoubtedly exerts a deterrent effect on the right to freedom of expression,[98] which is a valid concern for advocates of this right. Nonetheless, the gravitas of certain forms of online fake news justifies imposing substantial financial penalties or even imprisonment, should the court find it appropriate and reasonable.

More and more national legal systems, like Cyprus and the UK, are turning to criminal law to tackle disinformation, including deepfakes and fake news. But relying

[96] Cyprus, Criminal Code, Cap. 154, s 50 (as amended), available at https://www.cylaw.org/nomoi/enop/non-ind/0_154/full.html (last accessed 29 August 2025).

[97] UK Government, 'Government Cracks Down on Deepfakes Creation' (GOV.UK, 30 November 2022) https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation (last accessed 12 September 2024).

[98] Natalie Alkiviadou, 'Prison for Fake News?' (*Verfassungsblog*, 19 June 2024) available at https://verfassungsblog.de/prison-for-fake-news/ (last accessed 2 December 2024).

too heavily on criminal penalties raises serious concerns for freedom of expression. According to Article 19(3) of the ICCPR and the UN's Human Rights Committee, any restriction on speech must be necessary and proportionate and criminal sanctions should only be used as a last resort. Heavy dependence on these measures can create a chilling effect, especially for journalists, activists, and vulnerable groups. Instead, other solutions, like civil penalties, working with independent fact-checkers, or co-regulation through tools like the EU Code of Practice on Disinformation can offer a more balanced and effective response, without undermining free speech.

10. Conclusion

Determining whether fake news qualifies for protection under freedom of expression due to its severity is a complex task for human content operators. This challenge is heightened when distinguishing between contentious but permissible opinions and harmful disinformation or deepfakes that could incite violence or panic. While serious fake news might affect public opinion or democratic processes without crossing legal boundaries, extremely serious deepfakes—particularly those posing threats to public safety or inciting hate crimes—necessitate stricter regulation. The situation becomes even more intricate when the content also involves hate speech, as it not only misrepresents facts but also promotes discrimination or violence against certain groups. Human content operators must carefully balance safeguarding freedom of expression with addressing dangerous content. This is complicated by the constantly evolving tactics of disinformation and the overlap with hate speech, necessitating clear and effective guidelines to ensure that measures target harmful content without unduly suppressing legitimate discourse. To conclude, the role of online intermediaries is becoming increasingly challenging with the advancement of technology. By using algorithms and their workforce, these intermediaries must make critical decisions that often place them in a quasi-judicial role. They need to balance protecting freedom of expression, safeguarding their platforms from illegal content, and considering the principle of deterrence. This complex responsibility requires careful judgment to navigate the nuances of permissible speech and harmful content, ensuring that their actions do not inadvertently infringe on fundamental rights or contribute to an overly restrictive environment.

References

Books and Articles

Alkiviadou Natalie, 'Prison for Fake News?' (*Verfassungsblog*, 19 June 2024) available at https://verfassungsblog.de/prison-for-fake-news/ (last accessed 2 December 2024).

Brown A., 'What Is Hate Speech? Part 1: The Myth of Hate' (2017) 36 *Law and Philosophy* 123–164.

Casarosa F., 'The European Regulatory Approach toward Hate Speech Online: The Balance between Efficient and Effective Protection' (2019) 55 *Gonzaga Journal of International Law* 89–128.

Dickinson G.M., 'Section 230: A Juridical History' (2025) 28 *Stanford Technology Law Review* 1–35.

Greenberg J., *The Cambridge Introduction to Satire* (Cambridge University Press, 2019).

Humprecht E., Esser F. & Van Aelst P., 'Resilience to Online Disinformation: A Framework for Cross-National Comparative Research' (2020) 25(3) *International Journal of Press/Politics* 493–516.

Jougleux P., *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework* (Springer, 2022).

Kalsnes B., 'Fake News' (2018) *Oxford Research Encyclopedia of Communication*.

Lazić A. & Žeželj I., 'A Systematic Review of Narrative Interventions: Lessons for Countering Anti-Vaccination Conspiracy Theories and Misinformation' (2021) 30(6) *Public Understanding of Science* 644–670.

Le Merrer E. & Trédan G., *What is a Black Box Algorithm? Tractatus of Algorithmic Decision-Making* (2023) ffhal-03940259f.

Madrid-Morales D. & Wasserman H., 'Research Methods in Comparative Disinformation Studies' in Wasserman H. & Madrid-Morales D. (eds), *Disinformation in the Global South* (Wiley Blackwell 2022) 41–57.

Mala S., *The Legal Framework of Online Hate Speech (Το Νομικό πλαίσιο του Διαδικτυακού Μισαλλόδοξου Λόγου)* (Nicosia, Hippasus 2023) (in Greek).

Podstawa K., 'Hybrid Governance or… Nothing? The EU Code of Conduct on Combatting Illegal Hate Speech Online' in *Use and Misuse of New Technologies* (Springer, 2019).

Rayhan S. & Rayhan S., 'The Role of AI in Democratic Systems: Implications for Privacy, Security, and Political Manipulation' (2023) DOI: 10.13140/RG.2.2.31121.61281.

Rosenfeld S., *Democracy and Truth: A Short History* (University of Pennsylvania Press, 2019).

Seglow J., 'Hate Speech, Dignity and Self-Respect' (2016) 19 *Ethical Theory and Moral Practice* 1103–1116.

Smalley E., 'Russia's False Claims About Biological Weapons in Ukraine Demonstrate the Dangers of Disinformation and How Hard It Is to Counter – 4 Essential Reads' *The Conversation* (2022).

Spring M., 'Sadiq Khan Says Fake AI Audio of Him Nearly Led to Serious Disorder' *BBC News* (2024).

Spyropoulos P., 'The Spread of False News in the Age of "Fake News"' (2019) 8 *Epistimonika Apotipomata*.

Strömbäck J., Wikforss Å., Glüer K., Lindholm T. & Oscarsson H., *Knowledge Resistance in High-Choice Information Environments* (Routledge 2022).

Twomey J., Ching D., Aylett M.P., Quayle M., Linehan C. & Murphy G., 'Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets that Discuss Deepfakes in the Russian Invasion of Ukraine' (2023) 18(10) *Plos One*.

Legislation & International Instruments

Charter of Fundamental Rights of the European Union [2000] OJ C 364/1, art 11(1).

Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L 328/55.

Council of Europe, Committee of Ministers, Recommendation No R (97) 20 on 'Hate Speech' (1997).

Cyprus, Criminal Code, Cap 154, Article 50.

Cyprus, Criminal Code, Cap 154, Article 99A.

Cyprus, Law on Combating Certain Forms and Expressions of Racism and Xenophobia through Criminal Law 2011 (134(I)/2011).

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178/1, available at https://eur-lex.europa.eu/eli/dir/2000/31/oj (last accessed 28 August 2025).

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178/1, art 14.

European Commission, *2016 Fundamental Rights Colloquium: Conclusions* (2016) http://ec.europa.eu/information_society/newsroom/image/document/2016-50/2016-fundamental-colloquium-conclusions_40602.pdf (last accessed 28 August 2024).

European Commission, *Code of Conduct on Combating Illegal Hate Speech Online* http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300.

European Commission, *Code of Practice on Disinformation* (2018) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation (last accessed 10 June 2025).

European Commission, *Code of Practice on Disinformation* (Digital Strategy, 26 May 2021) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation (last accessed 3 June 2025).

European Council, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* (2017) https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (last accessed 30 August 2024).

European Media Freedom Act, Regulation (EU) 2024/1083 of the European Parliament and of the Council of 11 April 2024 on safeguarding media freedom in the European Union [2024] OJ L 7 May 2024, 1, available at https://eur-lex.europa.eu/eli/reg/2024/1083/oj (last accessed 28 August 2025).

European Parliament, 'The Legal Framework to Address "Fake News": Possible Policy Actions at the EU Level' (Policy Department for Economic, Scientific and Quality of Life Policies 2018).

Germany, Criminal Code (Strafgesetzbuch – StGB), Article 130.

International Covenant on Civil and Political Rights (ICCPR), adopted 16 December 1966, UNGA Res 2200A (XXI), entered into force 23 March 1976, 999 UNTS 171.

Oxford Internet Institute, *Social Media Manipulation by Political Actors: An Industrial Scale Problem* (University of Oxford 2021) https://www.oii.ox.ac.uk/publications/social-media-manipulation-by-political-actors-an-industrial-scale-problem/ (last accessed 30 August 2024).

Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online [2021] OJ L 172/79, art 3(3).

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1, available at https://eur-lex.europa.eu/eli/reg/2022/2065/oj (last accessed 29 August 2025).

Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L ..., art 3(60).

UN General Assembly, Report of the Special Rapporteur on the rights to freedom of peaceful assembly and of association, A/74/486 (2019).

UN Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/47/25 (13 April 2021).

United Nations High Commissioner for Human Rights, 'Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incite-

ment to National, Racial or Religious Hatred' (2013) https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf (last accessed 29 August 2025).

United Nations Human Rights Council, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (UN Doc HR/PUB/11/04, 2011) https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (last accessed 30 May 2025).

United States, *Communications Decency Act*, 47 USC § 230 (enacted 1996).

Cases

Akdaş v Turkey, App no 41056/04 (ECtHR, 16 February 2010).

Alves da Silva v Portugal, App no 41665/07 (ECtHR, 20 October 2009).

Beizaras and Levickas v Lithuania, App no 41288/15 (ECtHR, 14 January 2020).

Bielau v Austria, App no 20007/22 (ECtHR, 27 August 2024).

Brzeziński v Poland, App no 47542/07 (ECtHR, 25 July 2019).

Delfi AS v Estonia, App no 64569/09 (ECtHR, 10 October 2013).

Delfi AS v Estonia, App no 64569/09 (ECtHR, 16 June 2015).

Delfi AS v Estonia, App no 64569/09 (ECtHR, Grand Chamber, 16 June 2015).

Editorial Board of Pravoye Delo and Shtekel v Ukraine, App no 33014/05 (ECtHR, 5 May 2011).

Eon v France, App no 26118/10 (ECtHR, 14 March 2013).

Erbakan v Turkey, App no 59405/00 (ECtHR, 6 July 2006).

Gonzalez v Google LLC, 598 U.S. ____ (2023).

Groppera Radio AG and Others v Switzerland, App no 10890/94 (ECtHR, 28 December 1990).

Handyside v United Kingdom, App no 5493/72 (ECtHR, 7 December 1976).

Karataş v Turkey, App no 23168/94 (ECtHR, 8 July 1999).

Kuliś and Różycki v Poland, App no 27209/03 (ECtHR, 6 October 2009).

Lingens v Austria, App no 9815/82 (ECtHR, 8 July 1986).

Lilliendahl C.J. v Iceland, App no 29297/18 (ECtHR, 12 May 2020).

Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary, App no 22947/13 (ECtHR, 2 May 2016).

Mouvement Raëlien Suisse v Switzerland, App no 16354/06 (ECtHR, 13 July 2012).

Müller and Others v Switzerland, App no 10737/84 (ECtHR, 24 May 1988).

Otto-Preminger-Institut v Austria, App no 13470/87 (ECtHR, 20 September 1994).

Ovchinnikov v Russia, App no 24061/04 (ECtHR, 16 December 2010).

Salov v Ukraine, App no 65518/01 (ECtHR, 27 April 2004).

Sanchez v France, App no 45581/15 (ECtHR, Grand Chamber, 15 May 2023).

Sánchez v Spain, App no 45532/20 (ECtHR, 6 February 2024).

Şener v Turkey, App no 26680/95 (ECtHR, 18 July 2000).

Stambuk v Germany, App no 37928/97 (ECtHR, 17 October 2002, Third Section).

Strauß Karikatur, 1 BvR 313/85, BVerfGE 75, 369 (German Constitutional Court, 3 June 1987).

Times Newspapers Ltd v United Kingdom, App nos 3002/03 and 23676/03 (ECtHR, 10 March 2009).

Vereinigung Bildender Künstler v Austria, App no 68354/01 (ECtHR, 25 January 2007).

Zöchling v Austria, App no 4222/18 (ECtHR, 5 September 2023).

Online News & Media Sources

Deutsche Welle, 'Paris Olympics Boxer Imane Khelif Battles Hate Speech' (DW, 14 June 2025) https://www.dw.com/en/paris-olympics-boxer-imane-khelif-battles-hate-speech/a-69863650 (last accessed 14 June 2025).

Kinsella E., 'Parisian Court Rules It Has Jurisdiction in L'Origine du Monde vs. Facebook Case' (Artnet News, 5 March 2015) https://news.artnet.com/art-world/parisian-court-rules-it-has-jurisdiction-in-lorigine-du-monde-vs-facebook-case-275117 (last accessed 13 June 2025).

Philenews, 'September to the House of Representatives: Amendment for Criminalization of Dissemination of False News' (Philenews, 26 June 2024) https://www.philenews.com/kipros/koinonia/article/1486457/septemvrio-sti-bouli-tropologia-gia-pinikopiisi-exivrisis-psevdon-idiseon/ (last accessed 2 September 2024).

Satariano A., 'Facebook Can Be Forced to Delete Content Worldwide, E.U.'s Top Court Rules' (The New York Times, 3 October 2019) https://www.nytimes.com/2019/10/03/technology/facebook-europe.html (last accessed 15 June 2025).

Institutional & Reports

European Union Agency for Fundamental Rights (FRA), *Bias in Algorithms – Artificial Intelligence and Discrimination* (2023) https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf (last accessed 18 August 2024).

UK Government, 'Government Cracks Down on Deepfakes Creation' (GOV.UK, 30 November 2022) https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation (last accessed 12 September 2024).