European Journal of Law and Political Science
Vol 4 | Issue 1 | February 2025
ISSN 2796-1176

REVIEW

Theory and Practice of Social Media's Content Moderation by Artificial Intelligence in Light of European Union's AI Act and Digital Services Act

Gergely Gosztonyi[1,*], Dorina Gyetván[2], and Andrea Kovács[2]

ABSTRACT

After a brief, general introduction to AI, the present article will discuss whether AI itself has freedom of expression or whether it only entitles to tech companies that own AI. All this is relevant in the context of whether we can consider AI a separate legal actor in the context of content creation on social media, which is the second main issue of this article. Here, negative and positive cases of content moderation will be presented, i.e., whether the full spectrum of content moderation in social media can be entrusted to AI as a separate actor with respect to liability, or whether moderation requires some form of human control, direction or supervision.

Keywords: AI Act, content moderation, Digital Services Act, transparency.

Submitted: December 06, 2024

Published: February 09, 2025

10.24018/ejpolitics.2025.4.1.165

[1]Habil. Associate Professor, Eötvös Loránd University (ELTE), Faculty of Law, Hungary.
[2]PhD Candidate, Eötvös Loránd University (ELTE), Doctoral School of Law, Hungary.

Corresponding Author:
e-mail: gosztonyi@ajk.elte.hu

1. Introduction

Although Gartner's so-called Hype Cycle curve exploring the rise of technology shows that generative AI has already entered the disappointing phase in August 2024 (Gartner, 2024), we are seeing more and more software with some form of AI. This trend cannot escape social media platforms, for example, Meta, the company behind Facebook and Instagram, is developing its own AI model called Llama, which users can feel free to tailor to their preferences (Meta, 2021). In addition, Meta is openly committed to using AI to filter content (Meta, 2021), as are X/Twitter and TikTok.

The rise of artificial intelligence, maybe seeming sudden to the public, has not escaped the attention of legislators. The problem, however, is that–in the words of Kinga Pázmándi– "on one plate of the sliding scale are the protection goals (...) that the law is supposed to safeguard, while on the other end is the socially beneficial innovation, and the »adjustment of the values of the sliding scale« is accompanied by constant (...) uncertainty" (Pázmándi, 2018, p. 11). At the time of writing the present study, within the last year, the European Union has adopted the EU AI Act[1] (hereinafter: AI Act) and the accompanying Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence[2] and the President of the United States has issued an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (hereinafter: Executive Order 14110).[3]

Although there is no generally accepted definition of artificial intelligence in the current literature (Nikolinakos, 2023), both the EU and the US define what is meant by artificial intelligence. According to the AI Act, means "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content,

[1] Regulation (EU) No 2024/1689 of the European Parliament and of the Council of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139, (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (the Artificial Intelligence Regulation), PE/24/2024/REV/1, OJ L 2024/1689, 12.7.2024, 12.7.2024.

[2] Directive of the European Parliament and of the Council on the adaptation of the rules on non-contractual civil liability to artificial intelligence (Directive on liability in the field of artificial intelligence) COM(2022) 496 final 2022/0303 (COD).

[3] Executive Order (E.O.) 14110 on Safe, Secure and Trustworthy Development and Use of Artificial Intelligence, 88 FR 75191.

recommendations, or decisions that can influence physical or virtual environments."[4] And according to Executive Order 14110, artificial intelligence is a "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action."[5]

In contrast to legislators, the literature is more sentimental, focusing on similarities or differences with human thinking.[6] When the literature examines the freedom of expression of AI–and the related matter of personality or the lack thereof–it tends to focus on the less technical, more human aspects, which is the focus of the first part of this article. The article then examines the general issues and regulation of content moderation in the EU, and the role and potential–positive and negative–of AI in connection with content moderation practices.

2. The Freedom of Expression of Artificial Intelligence

In case of some artificial intelligences, we have seen examples of them being given some kind of right separate from their author, as for example in the case of Sophia (Fuchs, 2024)[7] and Mirai (Alqodsi & Gura, 2023).[8] However, due to the increasing and widespread use of AI, liability issues are in the forefront of relevant literature,[9] while, if we leave aside the physical dimension, robot law has long been concerned with the possible rights of man-made artificially intelligent machines, including the right to communicate (Vadymovych, 2017).

When it comes to artificial intelligence and freedom of expression it is worth examining who is the subject to freedom of expression. The subject could be artificial intelligence, or an underlying subject, namely (a) the programmer (author), (b) the owner/operator, or (c) the user.

From the possible subjects, the personality of AI itself is the most controversial. The possibility of recognising AI as a subject of freedom of expression seems to be a valid possibility in the academic literature (Koops *et al*., 2010), for which different criteria are defined by different authors, ranging from quite technical, relatively simple to abstract conditions. One simple criterion, for example, is that the recognition of AI's own freedom of expression is conditional on the outcome being indeterminate and unpredictable (Świerczyński & Więckowski, 2023). A similar position is taken by Tim Wu, who argues that one can speak of a distinct personality if (1) one is capable of conceptual thinking and (2) one is able to express one's formed opinions (Wu, 2013).[10] Moving on to more abstract criteria, Vadymovych (2017) argues that in order to recognise AI's freedom of expression, it is necessary to consider whether the subject of the inquiry has a socially complex functionality that pushes the boundaries of human autonomy, and Schwitzgebel and Garza (2015) argue that adequate social status may be sufficient. Noteworthy, however, that AI may also be given legal personality–and thus also certain rights–and that the granting of legal personality is a discretionary decision of the legislator and, as such, the decision may also follow considerations of practicality (Ribeiro *et al*., 2024).

From another perspective, however, AI is not intrinsic, some kind of human activity or input is always required it (Kaminski & Jones, 2024), so if we do not attribute freedom of expression to AI, we need to look at the person behind it (Hines, 2019; Pizzetti, 2021). In case of the programmer, the operator and the user, the roles are intersecting and there is no sharp boundary between them. With regard to moderation, the following examples can illustrate possible scenarios:

a) The programmer, the operator and the user are the same. In this case, the same company programs, operates (troubleshoots, provides support) and uses (prompts, parameterises, makes decisions based on output),
b) The company outsources the development of artificial intelligence to another company, but it operates and uses it itself,
c) The company just acquires and uses the AI, while the programming and operation remains with someone else.

Whereas in case a) the situation is clear, the rights and responsibilities are all being linked to the company under scrutiny, in case c) the question is not only how much of the software is compact, so-called commercial off-the-shelf software, but how much of the software is custom-developed. With regards to the role of the operator, the AI Act can be followed as a guideline, which considers the service

[4] AI Act Article 3(1).
[5] E.O. 14110 Section 3 (b).
[6] For a summary of the characteristics of strong and weak artificial intelligences, see Bory *et al*. (2024).
[7] For example, Sophia has been granted citizenship in Saudi Arabia.
[8] Mirai was granted a permanent establishment permit independently of the operating company.
[9] E.g., see Montagnani *et al*. (2024).
[10] See also in this context *Autronic AG v Switzerland* App no. 12726/87 (ECtHR, 22 May 1990).

provider, the user, the authorised representative, the importer and the distributor as falling within this role, if the AI is considered to be hazardous operation.[11]

However, from these actors, we will presumably consider the one who set the parameters necessary to generate (or in this case moderate) the content as the speaker (Hines, 2019). What the roles of the programmer, operator and user have in common, however, is that they all have a natural and/or legal personality, which gives them freedom of expression. With regard to social media providers, due to the proprietary development, the activity of AI will presumably be attributed to the providers, especially given the fact that the policies and guidelines that need to be implemented are also drafted by the provider (Robison, 2024).

Other authors, however, do not address the question of legal personality, but opt for a different justification for defending the output of AI, partly because in case of AI, precisely because of its unpredictability, its activity may not be attributable to humans, it may not be easily traceable back to human activity (Higby, 2021). These arguments include:

1. There is no doctrine that would explicitly exclude artificial intelligence from the scope of the First Amendment (Koops *et al*., 2010).
2. The output of AI may contain elements that are of interest or importance to people, or it may contain communications that are otherwise protected by freedom of expression (Koops *et al*., 2010). In turn, they participate in democratic debates and information dissemination (Kaminski & Jones, 2024). As part of this argument, the right of the audience to access information as part of freedom of expression should be highlighted.[12]

In case of AI-generated messages, there are consequences both for accepting that AI has freedom of expression and for not accepting as well. If AI does not fall within the scope of freedom of expression simply because it does not come directly from a human being, then it can be used to censor otherwise protected speech (Higby, 2021). If, on the other hand, it is not excluded from the scope of freedom of expression, then there may be a problem in judging speech that is not otherwise protected, depending on the speaker's intent, and there may be a question of liability (Higby, 2021). In case of social media, where proprietary artificial intelligence is actually used for content moderation, the situation is clear: both the development and the parameterisation (the drafting of the policies to be implemented and the way they are implemented) are united under one hand (Robison, 2024).

3. General Issues and Regulation of Content Moderation in the EU

3.1. Content Moderation and DSA, Particularly the Relevant Transparency Requirements

In March 2021, Mark Zuckerberg told during his congressional hearing that 95% of hate speech content and, to the best of his knowledge, 98%–99% of terrorist content is now being identified by artificial intelligence, instead of humans (House Energy and Commerce Subcommittee on Communications & Technology, 2021). Regulating the moderation practices and policies of Facebook, X/Twitter and other social media platforms around the world is a major challenge. 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 (hereinafter: the DSA)[13] primarily regulates intermediary service providers, including online platforms, complementing other EU efforts to regulate AI through provisions on content moderation and algorithmic transparency and accountability requirements (Brown, 2023).

In this context, the DSA also seeks to address some of the problematic issues of content moderation, with the objectives of ensuring transparency and accountability[14] and the effective remedy of content moderation decisions by online platform providers.[15]

Although the DSA defines the concept of content moderation, it has not previously had a–and still does not have in literature universally accepted–definition, but it is rather considered an umbrella-term (Gosztonyi, 2023). From a scientific standpoint, content moderation can be grasped as encompassing all actions of social media service providers that raise fundamental rights issues, in all cases involving user content by restricting access to it by human resources or artificial intelligence, but which do not presuppose the illegality of the content concerned, since their primary "legal basis" is to be found in the contractual terms of the platforms rather than in legislation and which is usually used by the platform to pursue its own economic or social agenda (Koltay, 2024) or to avoid liability,

[11] AI Act Article 3 (3–7).
[12] European Convention on Human Rights (ECHR) Article 10.
[13] 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), PE/30/2022/REV/1, OJ L 277, 27.10.2022, p. 1–102.
[14] DSA Recital 49.
[15] DSA Recitals 44, 109.

thus lacking transparency and accountability when it comes to decision-making. Moderating due to the platform's own economic interest (in order to avoid users encountering offensive, outrageous or disturbing content) is more predominant the more frequent the exchanges of views between users and the more intense the communication is (Koltay, 2024). By contrast, content moderation under the Article 3 of the DSA means:

> "the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account."[16]

Comparing the above two definitions, we can clearly see that there are substantial differences between the two definitions, the DSA does not make the fundamental rights aspect a conceptual element, although these decisions always involve–at least–a restriction of the users' freedom of expression. The other key difference between the definitions is that, while content moderation under the DSA also includes the process of detecting content affected by moderation, detection can hardly be considered as moderation in the sense that it does not affect fundamental rights in the absence of actual interference with the content. For the purposes of this study, given its fundamental rights approach, the act of non-interference detection, despite the definition of the DSA, will not be considered as moderation.

To counteract the moderation decisions of platforms based on largely automated decision-making and to safeguard fundamental rights restrictions,[17] the DSA contains five main tools: (i) comprehensible report on content moderation–terms and conditions; (ii) transparency reports; (iii) statement of reasons; (iv) internal complaint-handling system with the supervision of appropriately qualified staff; and (v) out-of-court dispute settlement.

Ad (i) Article 14 obliges intermediary service providers to include in their contractual terms and conditions on content moderation in a "clear, plain, intelligible, user-friendly and unambiguous language", and "shall be publicly available in an easily accessible and machine-readable format", including information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint handling system (Husovec, 2023).

Ad (ii) Among the fundamental characteristics of AI are its complexity, its dependence on data,[18] its inherent opacity (Vasiliki, 2022), its unpredictability and its unknowability (Yampolskiy, 2024), from which it logically follows that, in order to counterbalance these factors (Vasiliki, 2022), one of the main moral and fundamental rights objectives when it comes to the regulation of AI, is transparency (Stahl *et al*., 2023). The opacity of AI is also likely to hinder, for example, the exercise of the fundamental procedural rights of social media users, in particular the right to an effective remedy and to fair trial (Kattnig, 2024).

At the same time, autonomous behaviour is another fundamental characteristic of AI,[19] which makes ensuring transparency challenging, for example, especially with regard to machine learning. This problem is also reinforced by the so-called "black box effect", which essentially refers to the characteristic of AI systems that autonomous AI systems operate in a way that is inherently unintelligible to humans (users) (Wulf & Seizov, 2020). Moving along this line of thought, we are faced with the paradoxical situation (information overload or transparency paradox (Vasiliki, 2022) that full transparency, i.e., the disclosure of the exact programming code or algorithm, is not able to achieve the desired objective of transparency, since users are unable to understand the systems' operation due to a lack of knowledge, or only with a disproportionate expenditure of time and energy, which has the exact opposite effect: it potentially misleads users rather than inform them about the reliability of the system. Furthermore, transparency in the ordinary sense could also mean that knowledge about AI is systematically transferred from one stakeholder to another (Blackman & Ammanath, 2022), which could, however, result in imposing transparency requirements on AI systems to an inappropriate extent

[16] DSA Article 3. t).
[17] DSA Recital 9.
[18] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (hereinafter: EC Proposal) 3.5.
[19] EC Proposal 3.5.

and in an inappropriate manner, which would result in a disproportionate damage by obliging the disclosure of know-how and business information (trade secrets).[20]

The DSA imposes transparency reporting obligations on service providers to counteract moderation decisions based on automated decision making.[21] These obligations are set out cumulatively at three levels: firstly, in the context of minimum content elements for all intermediary service providers,[22] secondly, for service providers operating an online platform,[23] and thirdly, for service providers operating a very large online platform or a very popular online search engine.[24]

According to Articles a5 and 42 of the DSA, the most popular social networking service providers that are considered to be a very large platform are required to include the following information in their transparency reports on automated decision-making:

a) "Meaningful and comprehensible information about the content moderation engaged in at the providers' own initiative, including the use of automated tools, the measures taken to provide training and assistance to persons in charge of content moderation, the number and type of measures taken that affect the availability, visibility and accessibility."[25]
b) "Any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied."[26]
c) "The indicators of accuracy and related information" referred to in point (b).[27]

According to Ad (iii), one of the main objectives of the DSA was to improve the availability of effective remedy mechanisms.[28] In order to ensure that users can effectively exercise their right to remedy, all hosting providers are required to provide adequate justification for their decisions (not only when deleting content, but also, for example, when demonetising or demotioning content) when moderating content, as justification is an instrumental element for the exercise of the right to effective remedy. The obligation of social media service providers to state the reasons of their decisions applies to any restrictions imposed on the grounds of illegality or incompatibility with the contractual terms of the content (information) shared by the user, and even if with limited content,[29] the intermediary service provider must also state the reasons for not acting on any requests or notifications. The DSA follows the structure and logic prescribed for judicial decisions, with the exception that the legal consequence applied is also set out in the statement of reasons instead of in the operative part. In addition, the statement of the facts, the legal basis or contractual terms and information on the use of automated (AI-assisted) tools in reaching the decision are mandatory elements. The statement of reasons should also include a clause on possible appeals, as in judicial decisions.

According to Ad (iv), while the obligation to state reasons under Article 17 of the DSA applies largely when the service providers are active, the electronic and free effective complaint mechanism[30] under Article 20 of the DSA, which must be guaranteed for at least six months, applies to a wider range of decisions, also providing the possibility to lodge a complaint even if the platform remains passive in case of a notification under Article 16.[31] The complaint procedure necessarily requires staff intervention, and the service provider must inform the complainant without undue delay of its reasoned decision and of the possibilities for out-of-court dispute resolution or other means of redress. However, the phrase "without undue delay" is as broadly formulated as the immediate removal obligation in the E-Commerce Directive.[32] In the absence of further specification, this expression is likely to be a source of uncertainty and fragmentation, like the assessment of the immediacy of the removal obligation.[33]

Regarding Article 17, the question rightly arises: to what extent can a notifier lodge a complaint, whose notification did not incide an action, therefore the service provider was not obliged to state reasons for the decision under Article 17. The difference in the obligation to state reasons in the case of a restriction and the lack of such obligations in case of the passivity of the intermediary service provider (Ortolani, 2023) may result in that a user whose notice had no consequences (service

[20] EC Proposal point 3.5.
[21] DSA Recital 65.
[22] DSA Article 15.
[23] DSA Article 24.
[24] DSA Article 42.
[25] DSA Article 15 (1) c).
[26] DSA Article 15 (1) e).
[27] DSA Article 42 (2) c).
[28] DSA Recital 9.
[29] DSA Article 16 (5).
[30] DSA Article 20 (1).
[31] DSA Article 17.
[32] 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), OJ L 178, 17.7.2000, 1–16.
[33] See *Delfi AS v Estonia* App no. 64569/09 (ECtHR, 16 June 2015); *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt. v Hungary* App no. 22947/13 (ECtHR, 2 February 2016); *Pihl v Sweden* App no. 74742/14 (ECtHR, 9 March 2017).

provider's inaction) does not have the same chances of a successful redress a user whose content has been removed by the service provider with a reasoned decision.[34] However, looking at the entirety of obligations, it can be seen that the service provider is obliged to provide a statement of reasons covering the circumstances and the facts of the complaint decided under the internal complaint handling mechanism, thus complementing the lack of reasoning obligation in case of passivity. As the exhaustion of the complaint mechanism is a prerequisite for access to out-of-court dispute resolution (Lendvai, 2024) this obligation enables the user to challenge a reasoned decision before the Article 21 out-of-court dispute resolution forum. (Husovec, 2023).

Ad (v) Out-of-court dispute resolution is of paramount importance because in the past, excluding Facebook (Meta) Oversight Board' almos four years of activity, the only way to resolve disputes between users and social media platforms was through national (and possibly arbitration) courts. As of 17 February 2024, following the implementation of the DSA, users have the right to out-of-court dispute resolution, which, due to the form of the regulation, must be provided by all Member States, with a significant share of the costs to be borne by social media service providers. (Ruschemeier *et al.*, 2024) The out-of-court dispute resolution mechanism outlined the Regulation is most similar to the alternative fora provided for by the German Netzwerkdurchsetzungsgesetz[35] (Act on the Improvement of Enforcement in Social Networks, hereinunder: NetzDG). While the NetzDG only applies to social media platforms with more than two million German users,[36] the obligation under the DSA applies to all providers of online platforms larger than micro and small enterprises that are located or established in the EU, irrespective of the place of establishment of these intermediary service providers.[37]

Although the decisions resulting from the procedure before the panel under Article 21 of the DSA are not binding, except in the case of a submission, they allow a panel of experts to decide in a timely and cost efficient way, to settle disputes outside the judicial organisation, that the users of the online platform could not resolve directly with the service provider.[38] These procedures do not affect the right to go to court, but they certainly have the advantage of having a panel of experts instead of a lay court.

3.2. Content Moderation, Transparency and the AI Act

The AI Act brings significant changes to the way online platforms use artificial intelligence in content moderation, and brings both opportunities and challenges. The AI Act takes a risk-based approach, classifying AI systems into four categories: unacceptable, high, systemic and minimal risk (Ebers, 2024) Content moderation tools used by online platforms generally fall into the category of high or limited risk systems, depending on their potential impact on fundamental rights such as freedom of expression and non-discrimination.

The AI Act prohibits the use of artificial intelligence systems that pose an unacceptable risk to fundamental rights. This includes AI systems that manipulate human behaviour or exploit vulnerable persons in ways that could cause them harm.[39] Regarding content moderation, this category would include AI systems that intentionally or systematically censor lawful content in order to manipulate public opinion (Veale & Zuiderveen Borgesius, 2021).

High-risk artificial intelligence systems are subject to the strictest rules.[40] Content moderation systems used by very large online platforms such as Facebook, YouTube and X/Twitter can be classified as high-risk because of their potential impact on users' fundamental rights. These platforms often use artificial intelligence to automatically detect and remove content that violates community standards, such as hate speech, misinformation and extremist content. Because ofthe sensitivity of these functions, the EU imposes specific obligations on high-risk AI systems, in particular with regard to transparency, supervision and accuracy (Wagner, 2021).

Content moderation tools that have less impact on users' rights are considered systemic and minimal risk.[41] These may include basic automatic filtering systems that, for example, block spam or detect certain copyright infringements. For limited risk systems, the requirements are less stringent but still include transparency obligations. Artificial intelligence with minimal risk, such as general-purpose algorithms used in recommender systems or language filters with limited moderation functionality, are subject to even fewer regulatory obligations (Floridi, 2021).

Several articles of the AI Act directly address the responsibility of platforms using AI to moderate content, focusing on transparency, data management, accountability and human supervision.

Proper data management is essential to ensure the fair and unbiased operation of the artificial intelligence systems used to moderate content (Boone, 2023). Article 10 of the AI Act requires that the

[34] DSA Article 16 (5).
[35] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz, NetzDG), BGBl. I S. 3352.
[36] NetzDg § 1 (2).
[37] DSA Article 21.
[38] DSA Article 21 (2).
[39] AI Act Article 5.
[40] AI Act Articles 6–27.
[41] AI Act Articles 51–55.

data used to train high-risk AI systems should be of high quality, free from bias and be representative of all relevant groups. This is particularly important in content moderation, where biased training data may result in unfair treatment of certain groups or overly strict supervision of marginalised communities (Binns, 2018). For example, if an AI system used to flag hate speech is trained on data that is predominantly from dialects or languages, it may disproportionately flag content from these groups, which may raise concerns about discrimination and freedom of expression.

In order to prevent over-reliance on automated decisions, Article 14 of the AI Act subjects high-risk AI systems to human supervision. In the context of content moderation, this means that although AI can automate the detection of harmful content, human moderators should be involved in the final decisions, especially those affecting users' fundamental rights. This human approach is essential to address the limitations of AI, such as its inability to fully grasp or appropriately understand context. In case of content moderation, this means that AI tools reliably detect harmful content while minimising false positives, where legitimate content is wrongly flagged or removed, and false negatives, where harmful content is missed (Gillespie, 2018). The provisions on human supervision reinforce the importance of balancing automated decisions and human judgement. While artificial intelligence can effectively moderate large amounts of content, human moderators remain essential to deal with complex issues requiring contextual understanding.

The MI Act also contains relatively strict transparency requirements and definitions, similar to the DSA, in order to reduce risks relating to certain content moderation issues. To clarify the meaning of transparency, for example, Recital 27 states that:

Transparency in the AI Act means that AI systems should be developed and used in a way that allows for adequate traceability and explainability, while making people aware that they are communicating or interacting with an AI system, while adequately informing users of their rights, as well as of the capabilities and limitations of the AI system.[42]

Building on this, Article 50 of the AI Act includes transparency obligations for AI systems that (a) interact with people, (b) are used to detect emotions or to determine (social) categories based on biometric data, or (c) create or manipulate content, in order to balance the rights and interests of different stakeholders and to reduce information asymmetry.[43] Persons should be informed when interact with AI systems or have their emotions or characteristics recognised with automated tools. Where an AI system is used to create or manipulate images, audio or video content that bears a perceptible resemblance to authentic content, it is mandatory to disclose that the content was created by automated means, with the exception falling under the legitimate purpose exception. This transparency allows individuals to make informed decisions or to withdraw from a given situation.[44]

4. Artificial Intelligence in the Practice of Content Moderation

Artificial intelligence has become an essential tool for the content moderation of online platforms, where managing the huge amount of user-generated content is a critical task. Platforms such as Facebook, X/Twitter, YouTube and Instagram process huge volumes of posts, images, videos and comments on a daily basis, and AI offers an effective way to automate the monitoring, flagging and removal of content that violates community guidelines. While AI can greatly improve content moderation, it is not without challenges. Its application has resulted in both successful and flawed content classification, leading to numerous debates around accuracy, fairness and ethical implications. In this chapter we examine the effectiveness, limitations and specific cases where AI has performed well or poorly in content moderation.

Artificial intelligence usually works by combining machine learning algorithms, natural language processing and computer vision to moderate content. These technologies allow AI to quickly examine and analyse large amounts of data and identify patterns or signals that may indicate violations of community guidelines.

Algorithms are trained on large data sets containing labelled examples of acceptable and unacceptable content. Artificial intelligence then uses these learned patterns to predict whether new content meets community guidelines (Gorwa *et al*., 2020). These models are constantly evolving as they encounter new data, enabling dynamic improvements in content recognition. *Natural Language Processing* (NLP) is essential for analysing text-based content and detecting inappropriate phrases. NLP models can sometimes recognise the context of a word or phrase, distinguishing between benign and harmful uses (Schmidt & Wiegand, 2017). For example, X/Twitter uses NLP to flag tweets containing offensive language or hate speech. To moderate images and videos, computer vision

[42] AI Act Recital 27.
[43] AI Act Article 50.
[44] EC Proposal 5.2.4.

algorithms analyse visual content to detect inappropriate images such as nudity, graphic violence or terrorist propaganda (Araujo *et al*., 2020).

Artificial intelligence is very effective in identifying harmful content in many cases. The ability to process huge amounts of data quickly is essential for platforms with billions of users. Facebook, for example, has reportedly invested heavily in AI to detect and remove terrorism-related content. The company reported in 2024 that more than 99% of terrorism-related posts were flagged by AI before they were reported by users (Facebook, 2024). During the COVID-19 pandemic, YouTube used AI to combat the spread of health misinformation, and AI was used to identify 94% of videos flagged for policy violations (Parabhoi *et al*., 2021). Twitter's AI proactively handled more than 99% of offending accounts in 2024 without relying on user reports (X, 2024).

Despite these successes, AI-based content moderation still faces significant limitations, which have led to high-profile failures. These often stem from AI's inability to understand the full context of content, cultural nuances and language complexities. While these systems are constantly evolving, flawed content moderation can also harm the reputation of companies. In 2016, Facebook's AI incorrectly removed the iconic Vietnam War Napalm Girl photo, claiming that its nudity violated its community guidelines (Gillespie, 2018). AI failed to recognise the historical significance of the image and treated it as inappropriate content based solely on its nudity. During the COVID-19 pandemic, YouTube relied heavily on AI to moderate content, as human moderators worked from home. This led to a significant increase in false positives when educational and news-related content about COVID-19 was removed as they were classified as misinformation. Satire also often poses a challenge to AI, it is often identified as infringing content, because AI is oblivious to its context and tone of voice, which is different from normal speech (Lymn & Bancroft, 2021).

Problems of content moderation stems from that despite the huge amount of data fed to AI, it is still a challenge to fully understand contexts, different cultures and language differences. For example, it would be important to be aware that language is constantly evolving, new slang and colloquial expressions are emerging rapidly, especially in social media. Artificial intelligence models trained on outdated datasets may not be able to capture new terms, which can lead to inaccurate moderation. In addition, language varies significantly across cultures, making it difficult for global platforms to develop moderation systems that can be used uniformly and consistently in all countries. It should be remembered that AI models are not better than the data sets they are trained on. If the data is biased or incomplete, the end result will produce biased results.

5. Summary, or the Relationship between Moderation by Artificial Intelligence and Humans

Given the limitations of artificial intelligence, many platforms use a hybrid approach to content moderation, combining the speed of artificial intelligence with human supervision. Human moderators review tagged content to ensure that contextual and linguistic nuances and cultural differences are taken into account. This collaborative approach allows platforms to balance the power of AI with the human ability to make informed decisions. As AI evolves, content moderation systems are likely to become more sophisticated. Deep learning, contextual analysis and AI's improved ability to understand cultural differences could increase the accuracy of moderation systems.

EU's Artificial Intelligence Regulation and DSA will play an important role in shaping the future of AI-driven content moderation on online platforms. These Regulations impose strict requirements on AI-powered systems and aim to ensure that content moderation tools are transparent, fair and accountable. They offer a way to ensure that the quality of content moderation is improved, and fundamental rights are protected in the digital age.

In summary, while AI plays a vital role in moderating content on online platforms, its use and efficiency is not unlimited (Oversight Board, 2024). The balance between artificial intelligence and human moderation will continue to evolve as online platforms adapt to new regulations, in particular as platforms are forced to mitigate the problems that arise. However, human oversight will remain key in the foreseeable future to manage the subtleties and complexities that AI cannot yet understand (Lendvai, 2024).

Funding

This article was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

Alqodsi, E. M., & Gura, D. (2023). High tech and legal challenges: Artificial intelligence-caused damage regulation. *Cogent Social Sciences*, *9*(2), 1–12. https://doi.org/10.1080/23311886.2023.2270751.

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *10*(81), 149–159.

Blackman, R., & Ammanath, B. (2022, June 20). Building transparency into AI projects. *Harvard Business Review*. https://hbr.org/2022/06/building-transparency-into-ai-projects.

Boone, T. S. (2023). The challenge of defining artificial intelligence in the EU AI Act. *Journal of Data Protection and Privacy*, *6*(2), 180–195. https://doi.org/10.69554/QHAY8067.

Bory, P., Natale, S., & Katzenbach, C. (2024). Strong and weak AI narratives: An analytical framework. *AI & Society*, *70*, 337. https://doi.org/10.1007/s00146-024-02087-8.

Brown, M. (2023, March 27). *EU Digital Services Act's Effects on Algorithmic Transparency and Accountability*. Lexology. https://www.lexology.com/library/detail.aspx?g=b9d6250f-7200-45e7-af96-0e8d4298b501.

Ebers, M. (2024). Truly risk-based regulation of artificial intelligence: How to implement the EU's AI Act-European Union Law Working Papers No. 101 (Report). Stanford–Vienna Transatlantic Technology Law Forum. https://law.stanford.edu/wp-content/uploads/2024/10/EU-Law-WP-101-Ebers.pdf.

Facebook. (2024). *Community Standards Enforcement Report: Q2 2024*. Meta. https://transparency.meta.com/reports/community-standards-enforcement.

Floridi, L. (2021). The European legislation on AI: A brief analysis of its philosophical approch. *Philosophy & Technology*, *34*(1), 215–222. https://doi.org/10.1007/s13347-021-00460-9.

Fuchs, T. (2024). Understanding Sophia? On human interaction with artificial agents. *Phenomenology and the Cognitive Sciences*, *23*(1), 21–42. https://doi.org/10.1007/s11097-022-09848-0.

Gartner. (2024, August 21). *2024 Hype Cycle for Emerging Technologies Highlights Developer Productivity, Total Experience, AI and Security*. Gartner. https://www.gartner.com/en/newsroom/press-releases/2024-08-21-gartner-2024-hype-cycle-for-emerging-technologies-highlights-developer-productivity-total-experience-ai-and-security.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press. https://doi.org/10.12987/9780300235029.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945.

Gosztonyi, G. (2023). Content management of censorship?. In G. Gosztonyi (Ed.), *Censorship from Plato to social media: The complexity of social media's content regulation and moderation practices* (pp. 7–19). Springer. https://doi.org/10.1007/978-3-031-46529-1_2.

Higby, L. (2021). Navigating the speech rights of autonomous robots in a sea of legal uncertainty. *Journal of Technology Law & Policy*, *26*(1), 33–53.

Hines, M. (2019). I smell a bot: California's S.B. 1001, free speech, and the future of bot regulation. *Houston Law Review*, *57*(2), 405–435.

House Energy and Commerce Subcommittee on Communications and Technology. (2021). *Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation (Testimony of Mark Zuckerberg)*. 117th Congress. https://www.congress.gov/event/117th-congress/house-event/111407.

Husovec, M. (2023). Will the DSA work?: On money and effort. In J. Van Hoboken, I. Buri, J. P. Quintais, R. Fahy, N. Appelman, M. Straub (Eds.), *Putting the DSA into practice: Enforcement, access to justice, and global implications* (pp. 19–34). Verfassungsblog.

Kaminski, M. E., & Jones, M. L. (2024). Constructing AI speech. *The Yale Law Journal Forum*, *133*(1), 1212–1266.

Kattnig, M. (2024). Assessing trustworthy AI: Technical and legal perspectives of fairness in AI. *Computer Law & Security Review*, *55*, 1–18. https://doi.org/10.1016/j.clsr.2024.106053.

Koltay, A. (2024). *Media Freedom and the Law: The Regulation of a Common European Idea*. Routledge. https://doi.org/10.4324/9781003321569.

Koops, B. -J., Hildebrandt, M., & Jaquet-Chiffelle, D. -O. (2010). Bridging the accountability gap: Rights for new entities in the information society? *Minnesota Journal of Law, Science & Technology*, *11*(2), 497–561.

Lendvai, G. F. (2024). A Facebook Ellenőrző Bizottság működése és bíráskodása a gyűlöletbeszéd kontextusában (The functioning and judging of the Facebook Oversight Board in the context of hate speech). *Medias Res*, *13*(1), 195–221. https://doi.org/10.59851/imr.13.1.11.

Lymn, T., & Bancroft, J. (2021, August 5). *The Use of Algorithms in the Content Moderation Process*. GOV.UK. https://rtau.blog.gov.uk/2021/08/05/the-use-of-algorithms-in-the-content-moderation-process.

Meta. (2021, December 8). *Meta's New AI System to Help Tackle Harmful Content*. Meta Newsroom. https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content.

Montagnani, M. L., Najjar, M. -C., & Davola, A. (2024). The EU regulatory approach(es) to AI liability and its application to the financial services market. *Computer Law & Security Review*, *53*(2), 1–19. https://doi.org/10.1016/j.clsr.2024.105984.

Nikolinakos, N. T. (2023). *EU Policy and Legal Framework for Artificial Intelligence, Robotics, and Related Technologies: The AI Act*. Springer. https://doi.org/10.1007/978-3-031-27953-9.

Ortolani, P. (2023). If you build it, they will come: The DSA procedure before substance approach. In J. Van Hoboken, I. Buri, J. P. Quintais, R. Fahy, N. Appelman, M. Straub (Eds.), *Putting the DSA into practice: Enforcement, access to justice, and global implications* (pp. 151–166). Verfassungsblog.

Oversight Board (2024, September 17). *Content Moderation in a New Era for AI and Automation*. Oversight Board Report. https://www.oversightboard.com/wp-content/uploads/2024/09/Oversight-Board-Content-Moderation-in-a-New-Era-for-AI-and-Automation-September-2024.pdf.

Parabhoi, L., Sahu, R. R., Dewey, R. S., Verma, M. K., Seth, A. K., & Parabhoi, D. (2021). YouTube as a source of information during the COVID-19 pandemic: A content analysis of YouTube videos published during January to March 2020. *BMC Medical Informatics and Decision Making*, *21*(1), 255–265. https://doi.org/10.1186/s12911-021-01613-8.

Pizzetti, F. G. (2021). Embryos, organoids, and robots: 'Legal subjects'? *BioLaw Journal*, *9*(1), 345–352. https://doi.org/10.15168/2284-4503-755.

Pázmándi, K. (2018). Digitalizáció, technológiai fejlődés, jogi paradigmák (Digitalisation, technological development, legal paradigms). *Gazdaság és Jog*, *26*(12), 10–14. https://doi.org/10.21637/GT.2018.01.02.

Ribeiro, B. A., Coelho, H., Ferreira, A. E., & Branquinho, J. (2024). Metacognition, accountability, and legal personhood of AI. In H. S. Antunes, P. M. Freitas, A. L. Oliveira, C. M. Pereira, E. V. de Sequeira, L. B. Xavier (Eds.), *Multidisciplinary perspectives on artificial intelligence and the law* (pp. 169–185). Springer. https://doi.org/10.1007/978-3-031-41264-6_9.

Robison, K. (2024, February 7). *Inside the Shifting Plan at Elon Musk's X to Build a New Team and Police a Platform 'So Toxic it's Almost Unrecognizable'*. Fortune. https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation.

Ruschemeier, H., Quintais, J. P., Nenadic, I., De Gregorio, G., & Eder, N. (2024, September 10). *Brave New World: Out-of-Court Dispute Settlement Bodies and the Struggle to Adjudicate Platforms in Europe*. Verfassungsblog. https://verfassungsblog.de/ods-dsa-user-rights-content-moderatin-out-of-court-dispute-settlement/.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L. Ku, C. Li (Eds.), *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*(1), 98–119. https://doi.org/10.1111/misp.12032.

Stahl, B. C., Brooks, L., Hatzakis, T., Santiago, N., & Wright, D. (2023). Exploring ethics and human rights in artificial intelligence: A Delphi study. *Technological Forecasting and Social Change*, *191*, 1–17. https://doi.org/10.1016/j.techfore.2023.122502.

Vadymovych, S. Y. (2017, April 12). Artificial personal autonomy and concept of robot rights. *CyberLeninka*, *17*(21), 17–21.

Vasiliki, P. (2022). Transparency in artificial intelligence: A legal perspective. *Journal of Ethics and Legal Technologies*, *4*(1), 25–40.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402.

Wagner, G. (2021). Liability for artificial intelligence: A proposal of the European Parliament. *Working Paper of the Research Institute for Law and Digital Transformation*, *9*, 1–34. https://doi.org/10.2139/ssrn.3886294.

Wu, T. (2013). Machine speech. *University of Pennsylvania Law Review*, *161*(6), 1495–1533.

Wulf, A. J., & Seizov, O. (2020). Artificial intelligence and transparency: A blueprint for improving the regulation of AI applications in the EU. *European Business Law Review*, *31*(4), 611–640. https://doi.org/10.54648/EULR2020024.

X. (2024). *Global Transparency Report H1 2024*. X Transparency Reports. https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf.

Yampolskiy, R. V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. Chapman & Hall. https://doi.org/10.1201/9781003440260.

Świerczyński, M., & Więckowski, Z. (2023). Statut Jednolity Sztucznej Inteligencji (Artificial Intelligence Uniform Statute). *Zeszyty Prawnicze*, *23*(1), 217–253. https://doi.org/10.21697/zp.2023.23.1.09.