LEGAL ASPECTS OF CONTENT MODERATION ON SOCIAL MEDIA PLATFORMS: A COMPARATIVE PERSPECTIVE

Assoc. Prof. Denitza Toptchiyska, PhD,
New Bulgarian University

Abstract

The paper analyses in a comparative perspective the legal aspects of content moderation on social media platforms, including its specificities in the context of the rule of law standards. The study compares the legal frameworks governing content moderation in the EU and the US, focusing on the role of government regulation and self-regulation by the platforms as well as the implications of content moderation concerning free speech, censorship, and privacy.

Additionally, the paper focuses on the obligations and responsibilities of social media platforms when moderating content, as well as the rights and freedoms of users and the procedures for appealing moderation decisions. The practical application of the two regulatory models is explored through the practice of the Meta Oversight Board and their adopted approach to resolving controversial cases and the new EU Digital Services Act.

Key words: social media, online platforms, regulation, soft law, content moderation

At the beginning of the 21st century, social media platforms represent a widely accessible technological communication environment that enables billions of users to share, create, and modify content in various forms and in real time. Besides being an economic driver and a successful business model, they are also recognized as a primary forum where modern individuals exercise their right to freedom of expression and access to information. Although there is a fast-growing international consensus that fundamental human rights, as established and protected in the territorial space, should be projected into the digital environment, the specificities of the online space, and particularly social media platforms, pose challenges to legal regulation.

The present publication aims to analyse, from a comparative legal perspective, the regulatory models in the United States (USA) and the European Union (EU) concerning the moderation of content on social media platforms. In this regard, the fundamental specifics of the digital environment and the combination of regulatory factors interacting within it will first be outlined.

1. Specifics and normative layers of the digital environment

The digital environment, often referred to as cyberspace, is characterized by specific features concerning normative layers and the effectiveness of regulatory mechanisms within it compared to the physical or territorial space. In the digital space, law, social norms, and markets as regulatory layers are complemented by technological architecture. While some theories suggest that technological innovations solely act as a driving force that alters future expectations, thus challenging the social consensus embedded in regulation,[1] other authors argue that technology represents a distinct normative level that dynamically interacts with other regulatory factors.[2] In their various combinations, these normative layers shape the regulatory models within the digital realm.

Law, as a form of public regulation, and technological architecture, as a form of private regulation, have their own advantages and disadvantages when it comes to regulating the digital space. The legitimacy of law as a social regulator is linked to the concept of the rule of law, which encompasses requirements for the laws such as generality and formal justice, clarity, precision, comprehensibility, absence of contradictions, limitations on retroactive norms, stability, consistency, publicity, and accessibility. The law should be adopted through a transparent law-making process, based on the principle of separation of powers, publicly promulgated, equally enforced, and independently adjudicated, ensuring equality before the law. However legal regulation is associated with certain drawbacks given the specifics of the digital space. The reach of legal rules is limited to the jurisdiction of the state that has enacted them, while digital networks have a global scope. The duration of procedures for adopting, amending, or repealing regulatory legal acts restricts the ability to timely respond to risks and hazards arising from rapidly evolving technologies. This is why legal regulation is often associated with a rigidity that restricts or diminishes the effectiveness of regulatory responses. Additionally, the abstraction inherent in legal norms can lead to ambiguities in the process of their application.

The technological architecture of the digital environment is constructed by the software and hardware that define its capabilities and functionalities available to its users, thereby directing their behaviour. Unlike legal rules, whose adoption is associated with a lengthy public process, technological standards embedded in networks can be easily and rapidly changed as they are determined by technology developers. The application of technological rules is not functionally limited to the territorial borders of a specific state jurisdiction but is determined by the scope of the network upon which they are applied, which is often transnational. Contrary to legal rules, where enforcement relies on subsequent coercive action through the state institutional mechanism in case of violation, the technological architecture can ensure ex ante rule enforcement by defining technical parameters for network usage and restricting

---

[1] Santos, B. de S. (2020). *Toward a new legal common sense law, globalization, and emancipation* (Third edition.). Cambridge University Press, p. 3

[2] Lessig, L. (1999). The Law of the Horse: What Cyberlaw Might Teach. *Harvard Law Review*, Vol.113(2), 501-549.

users' ability to modify them.[3] Aligning user behaviour with the rules of the technological architecture can be effectively achieved through integrated technological mechanisms such as filtering or tracking.

However, the regulations embedded in the technological design of Internet architecture predominantly reflect the interests of their developers, who are primarily commercial companies or corporations. Consequently, technological regulation tends to prioritize private interests that may not always align with the values agreed upon in society. The regulatory framework implemented through technological architecture often lacks transparency, leading to information asymmetry and unfairly positioning users at a disadvantage vis-a-vis private entities engaged in technological development. Furthermore, the enforcement of technologically embedded rules, relying on tracking and monitoring mechanisms, can pose substantial threats to fundamental social values, such as the freedom of expression and the right to privacy.

The effectiveness of regulatory mechanisms in the digital space largely depends on the interaction between legal norms and technological rules with the aim of leveraging their advantages and mitigating their shortcomings. The complexity faced by contemporary society is to achieve regulation of transnational technological networks that effectively reflects the balance of values around which different societies have reached a consensus. Although human rights are perceived as fundamental standards in international instruments, their specification and the balance between them largely depend on national political, economic, and legal traditions. In this context, the regulatory approaches of the USA and the EU regarding content moderation on social media platforms can be examined and analysed from a comparative legal perspective.

2. Regulatory approach to content moderation on social media platforms in the US

The evolution of the Internet and Internet services in the United States during the 1990s highlighted the necessity of introducing legal regulations concerning the liability of Internet service providers. Considering the specific characteristics of the digital landscape and the emerging business model of Internet platforms and other intermediary services, section 230 of the Communication Decency Act (CDA) was enacted in 1996, granting them immunity with regard to content created and shared by third parties. The adoption of Section 230, DCA is justified by the significant importance of the Internet and other interactive computer services, as they provide a platform that enables diverse political discussions, facilitates cultural development, and fosters intellectual engagement. The main purpose of the law is to promote the development of the Internet and other interactive computer services and media, as well as to preserve a dynamic free market unrestricted by legal regulations. Additionally, the aim is to promote the development of technologies that enable user control over content, including blocking and filtering mechanisms, to safeguard children from inappropriate online content.

---

[3] Reidenberg, J. (1998). Lex Informatica: The Formulation of Information Policy Rules through Technology. *Texas Law Review*, Vol.76(3), 553-594.

Section 230 of the CDA grants immunity to providers of interactive computer services, shielding them from civil liability, provided that they act in good faith when removing or moderating third-party content that they or a user find to be obscene or offensive, even if it involves constitutionally protected speech. The case law surrounding Section 230 takes a broad interpretation, acknowledging its explicit prohibition on courts considering lawsuits that would cast computer service providers as publishers. It is understood that the primary objective of this legislation is to incentivize providers of online interactive services to actively monitor the Internet for harmful content and remove obstacles to implementing self-regulatory measures. Consequently, the court maintains that Section 230 bars the imposition of publisher liability on online service providers for their exercise of editorial and self-regulatory functions, as the potential legal risks would discourage them from blocking and screening offensive materials.[4] As a result, legal actions seeking to hold service providers accountable for engaging in traditional editorial functions of a publisher, such as making decisions regarding publication, withdrawal, postponement, or modification of content, are prohibited.

In accordance with Section 230 of the CDA, providers of interactive computer services are exempt from liability even when they are notified of the presence of potentially harmful content. The court acknowledges that imposing such liability, based on information or notification, would hinder providers from regulating the dissemination of content within their own services, as it would place them in a constant dilemma between suppressing controversial speech or assuming responsibility. In this context, it would contradict the purpose of the legal provision, which is precisely to shield providers from such obligations.[5] For users whose rights are violated by other users on social media platforms, there remains the possibility of direct legal action against the infringing party. In cases of anonymous speech, the court must assess, based on the evidence presented by the complainant, whether there is sufficient grounds to believe that a legal violation (defamation or invasion of privacy) has occurred in order to request the intermediary to disclose the identity of the user who posted the content.[6]

It should be noted that the immunity granted under Section 230 is limited regarding the obligation to remove content that infringes copyright, sexually explicit material, and other content that violates federal criminal laws. In 1998, the Digital Millennium Copyright Act (DMCA) was enacted, providing safe harbours for Internet service providers, including social media platforms, against copyright liability for infringing material uploaded by their users under specific conditions. Unlike the regime under Section 230 of the CDA, where regardless of whether the service provider has been notified of the harmful content, they are exempt from liability, to apply the immunity under the DMCA exemption regime, it is required that upon receiving a notice, the provider removes the material claimed to infringe copyright. Furthermore, although federal law provides a range of protections that can grant social media platforms immunity against claims based on the conduct of their users, this immunity is not

---

[4] Zeran v. Am. Online, Inc. - 129 F.3d 327 (4th Cir. 1997); Blumenthal v. Drudge - 992 F. Supp. 44 (D.D.C. 1998)

[5] Zeran v. Am. Online, Inc. - 129 F.3d 327 (4th Cir. 1997).

[6] John Doe No. 1 v. Cahill - 884 A.2d 451 (Del. 2005)

universal and can be lost under certain circumstances. Additionally, the immunity does not shield social media platforms from liability arising from their own conduct.

The legal immunity granted to social media platforms along with other internet intermediary service providers aims to incentivize them to develop self-regulatory policies. Furthermore, various agencies are responsible for overseeing the activities of social media platforms to ensure compliance with legislation, issuing advisory opinions and assessments regarding the operations of social media platforms.[7] The U.S. Federal Trade Commission (FTC) oversees compliance with numerous laws related to personal privacy and online advertising and also possesses the general authority to investigate unfair or deceptive trade practices. The FTC develops and publishes guidelines on how it will exercise its authority and encourages organizations to adopt their own policies in this area within the context of self-regulation. Although the guidelines issued by the FTC constitute soft law and have a quasi-legal nature, they help the addressees by providing clarity regarding the Commission's requirements.

In 2018, leveraging its regulatory freedom, Facebook, currently Meta company, undertook measures to create an Oversight Board as a means of legitimizing its content moderation practices. The Board, comprised of representatives from the international academic community and civil society, has the primary task of assisting the social platform in content moderation. The Board's decisions are binding on the company unless they violate established laws in any of the jurisdictions where the company operates. The objective of the Board is to promote freedom of speech and safeguard it by making principled and independent judgments on Facebook and Instagram content, while also providing recommendations on the applicable Facebook Company Content Policy.

The Oversight Board makes decisions regarding the removal or preservation of content based on Facebook's Community Standards, Facebook's values, and the Relevant Human Rights Standards. These standards include the UN Guiding Principles on Business and Human Rights (UNGPs), which were endorsed by the UN Human Rights Council in 2011 and establish a voluntary framework for businesses' human rights responsibilities. Additionally, the Board considers the International Covenant on Civil and Political Rights (ICCPR), the Human Rights Committee General Comment No. 34 on freedom of opinion and expression (2011), and the reports of the UN Special Rapporteur on freedom of opinion and expression.

According to Article 19, paragraph 3 of the International Covenant on Civil and Political Rights (ICCPR), restrictions on freedom of expression are permitted when the following three conditions are met: legality, legitimacy, and necessity. When resolving disputes related to content moderation, the Oversight Board consistently assesses whether each of these standards has been upheld. With regard to the legality standard, the Oversight Board conducts an evaluation of the Meta policy rule to assess its compliance in terms of clarity and specificity. In its Case Decision 2020-006-FB-FBR, dated January 28, 2021, the Oversight Board highlighted certain concerns regarding Facebook's patchwork of rules and policies displayed on various sections of its website. The lack of

[7] For ex. the National Labor Relations Board, the Securities and Exchange Commission

clear definitions for key terms, such as „misinformation," and the varying standards regarding whether a post „could contribute" or actually contributes to imminent harm, create difficulties for users in understanding the prohibited content.

When assessing the legitimacy standard, the Oversight Board evaluates whether Facebook's decision to moderate content is driven by a legitimate aim. When applying the necessity and proportionality standard, the Oversight Board assesses whether Facebook has chosen the least intrusive measures to achieve its legitimate public interest objective. In order to meet this requirement, Facebook must demonstrate that the public interest objective cannot be achieved without limiting speech, that it has chosen the least intrusive measure among those that restrict speech, and that the selected measure is effective and not counterproductive in achieving the intended goal.

The approach adopted by the Oversight Board in resolving content moderation disputes reflects the methodology commonly employed by international courts in addressing cases related to freedom of expression. The issue lies in the fact that only a limited number of disputes are examined by the Oversight Board, leaving the decisions of content moderators practically final and without the possibility to be effectively challenged in the remaining cases. In conclusion, we can highlight that the regulatory approach in the United States towards content moderation by social media platforms provides a substantial degree of freedom. It primarily relies on self-regulation by online service providers and soft legal instruments.

3. Regulatory approach to content moderation on social media platforms in the EU

In 2000, the EU adopted Directive 2000/31/EC on electronic commerce,[8] in which social media platforms, along with internet search engines, blogs, discussion forums, wiki applications, and photo and video-sharing social networks, are recognized as services of the information society. The directive aims to harmonize the national legislation of the Member States in the field of the Single Digital Market. According to Article 14 of the Directive, online service providers functioning as mere conduits, caching, or hosting service providers are not held accountable for the information they transmit or host. To qualify for liability exemption, two conditions must be satisfied: 1) they must lack actual knowledge of illegal activity or information, and 2) if they become aware of such content, they must promptly remove or disable access to it. Additionally, the Directive prohibits national governments from imposing a general monitoring obligation on these intermediaries. The primary goal is to establish efficient procedures for swiftly and reliably removing and disabling access to illegal content. Online intermediaries are encouraged to proactively implement measures while retaining their liability exemption under the e-Commerce Directive.

Regarding content moderation on social media that is not illegal but falls within the scope of harmful content, the European institutions initially undertake initiatives

[8] 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, p. 1-16

for self-regulation, aiming to engage social media platforms in these efforts. In September 2018, the Code of Practice on Disinformation was published, joined by Facebook, Google, Twitter, Mozilla, representatives of the advertising industry, and later by Microsoft and TikTok. In December 2018, an Action Plan against Disinformation was adopted, which provided for the European Commission (EC) to continuously monitor and analyse the implementation of the Code in collaboration with the Group of European Regulators for Audiovisual Media Services (ERGA) and the European Audiovisual Observatory. The EC report for the first year reached positive conclusions that the Code provides a valuable framework for a structured dialogue between online platforms and the EC, enhancing transparency and accountability in their disinformation policies.[9] However, further analyses revealed significant shortcomings, such as the inconsistent and incomplete application of the Code by platforms and Member States, lack of uniform definitions, and other limitations pertaining to insufficient public protection.[10] In accordance with the recommendations, the measures in the Code should be reviewed and supported by new guarantees, while simultaneously inviting a broader range of organizations to join.

As an indication of the inefficiency of the existing self-regulation mechanisms, the dissatisfaction expressed by Meta - Facebook users in Bulgaria in 2022 regarding the established content moderation mechanisms on the platform can be highlighted. The war in Ukraine has elicited a strong reaction in Bulgarian society, often expressed on social media platforms, especially on Facebook, which currently is the most used social media with the greatest influence in Bulgaria. The dispute arose over numerous cases of blocked civil voices and journalists on the social media platform, whose profiles were blocked mainly for posting against Vladimir Putin and the war that Russia is waging in Ukraine. In January 2023, representatives from Meta and TELUS Bulgaria (responsible for content moderation in Bulgarian, Russian, and Turkish languages) were invited to the National Assembly for a hearing.[11] During the hearing, Meta explained that their platform utilizes a blend of artificial intelligence and over 15,000 content reviewers, who are native speakers of the respective languages. These reviewers operate in over twenty global locations and to ensure linguistic comprehension, local residents serve as moderators for specific languages, including Bulgarian for local content oversight. After the hearing, in February 2023, Bulgarian media reported that Meta will terminate its contract with TELUS Bulgaria, with content moderation being relocated to Germany starting in July 2023. While some media have presented the outcome as a victory for civil society, there are still concerns regarding the evaluation in Germany of the content generated in Bulgaria to combat effectively the dissemination of disinformation and hate speech.

Parallel to the self-regulatory regime concerning the moderation of harmful content on social media platforms, the EU adopts a distinct approach regarding the moderation

---

[9] European Commission. SWD (2020)180 final. Assessment of the Code of Practice on Disinformation.

[10] European Commission. COM (2021) 262 final. *Guidance on strengthening the Code of Practice on Disinformation*

[11] Комисия в НС претупа скандала с „Фейсбук". Сега [онлайн], 26.01.2023 [прегледан на 20.05.2023]. Достъпен на: https://www.segabg.com/hot/category-bulgaria/komisiya-ns-pretupa-skandala-feysbuk

of content associated with the right to be forgotten, specifically when media platforms or intermediary service providers act as data controllers. In its judgment on the Google Spain Case (Case C-131/12),[12] the Court of Justice of the European Union (CJEU) establishes that the activity of search engines can be considered as the processing of personal data on its own grounds, even though they act as intermediaries and that this activity relates to information that has already been published and remains unchanged. As the processing of data, including searching for individual names and obtaining structured results of published information, can have a significant impact on individual's rights to privacy and data protection, the Court held that a fair balance must be struck between the legitimate interests of search engines and the rights protected under Article 7 and 8 of the European Convention on Human Rights (ECHR). The individuals have the right to ask search engines like Google to delist certain results for queries related to a person's name. The responsibility for decision-making is assigned to the data controller, in this case, the search engine, and their decision is subject to appeal before national supervisory authorities for personal data protection or through judicial proceedings.

Despite some concerns that this approach would have a chilling effect on freedom of speech, as data controllers would prefer to delete information rather than be responsible for their decision, the EU reaffirms its approach in the adopted General Data Protection Regulation in 2016 (article 17, Right to be forgotten). In its judgment of 8 December 2022 in Case C-460/20,[13] the CJEU provides guidance on the burden of proof on the individuals requesting de-referencing, as well as the obligations and responsibilities of the search engine operator. According to the guidance provided by the CJEU the individual requesting de-referencing based on the grounds of inaccurate content is required to demonstrate the clear and evident inaccuracy of the information contained in that content, or at least a substantial portion thereof that is not insignificant in relation to the content as a whole. The individual is only expected to provide evidence that, considering the specific circumstances of the case, can reasonably be expected of them to seek to establish the manifest inaccuracy. Regarding the data controllers, the court's guidance is that they should consider all the rights and interests involved, as well as the particular circumstances of the case. However, the data controller cannot be obligated to actively search for facts that are not substantiated by the de-referencing request in order to assess its validity. Additionally, the data controller is not required to investigate or engage in an adversarial debate with the content provider to obtain further information regarding the accuracy of the referenced content. In case of rejecting the request, the data subject must be able to bring the matter before the supervisory authority or the judicial authority competent to conduct appropriate investigations and order the data controller to take the necessary actions.

In cases concerning the protection of personal data, the objective of the regulatory approach in the EU is to combine self-regulation with opportunities for administrative, civil, and criminal legal protection in case of violations. A specific feature of the European model for personal data protection is the requirement for states to establish

[12] Judgment of the Court (Grand Chamber), 13 May 2014, Case C 131/12.

[13] Judgment of the Court (Grand Chamber) of 8 December 2022; Case C-460/20.

an independent authority responsible for monitoring compliance with the legal framework. In practice, this helps to address the shortcomings associated with self-regulation, such as impersonal decision-making or indefinite timelines for resolving cases.

With the adoption of the Digital Services Act[14] in October 2022, the EU embraces a comprehensive approach that combines private and public legal regulations regarding the moderation of content by online platforms and other providers of intermediary digital services. The EU Regulation provides for measures against illegal as well as harmful content, including disinformation. The aim is to create the so-called „co-regulatory backstop" that supplements self-regulatory mechanisms with legal safeguards, ensuring transparency and accountability on behalf of the digital platforms to regulators and users, without censoring the content.[15] The Digital Services Act introduces requirements regarding the handling of user notifications regarding illegal content, which must be addressed promptly, diligently, impartially, and objectively. Providers are obligated to inform both the user who submitted the moderation request and the user who uploaded the moderated content about their decision, including information about legal remedies. Specific requirements are outlined for online platforms regarding the establishment of an Internal Complaints Handling System, which allows users to challenge decisions made by the online platform. Additionally, provisions are made for out-of-court dispute resolution by certified bodies, as well as general oversight of platform activities by designated Digital Services Coordinators in each EU member state. The practical implementation of the provisions of the Digital Services Act is yet to be discussed among the Member States of the Union.

Conclusions

The Internet's worldwide reach and its vast user base, consisting of both service providers and end-users, diminish the efficacy of conventional legal frameworks, necessitating more adaptable and responsive regulation that accounts for participants' behaviour and the rapid pace of technological advancements. Concerning content moderation on social media platforms, two distinct regulatory approaches are currently emerging in the US and the EU. While in the US, the focus remains on self-regulation and soft law as primary regulatory instruments, in the EU, attention is directed towards the combined use of a common legislative framework aimed at addressing the limitations of self-regulation in the context of the digital environment. Despite the differences in regulatory approaches, which can largely be explained by variations in legal traditions, both jurisdictions share the view that the digital space should be regulated in accordance with internationally recognized human rights. This shared idea is likely to contribute to finding effective solutions in moderating the content on global social media platforms to support freedom of speech while ensuring that users are protected from illegal and harmful content.

---

[14] 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)

[15] Огнянова, Н. Няма да има Министерство на истината в ЕС. Публикувана в: „Дезинформацията: новите предизвикателства", Изд. СУ „Св. Климент Охридски", 2021 (Международна научна конференция - октомври 2021)

Bibliography

- European Commission. COM (2021) 262 final. Guidance on strengthening the Code of Practice on Disinformation
- European Commission. SWD (2020)180 final. Assessment of the Code of Practice on Disinformation.
- Lessig, L. (1999). The Law of the Horse: What Cyberlaw Might Teach. Harvard Law Review, Vol.113(2), 501-549.
- 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)
- Reidenberg, J. (1998). Lex Informatica: The Formulation of Information Policy Rules through Technology. Texas Law Review, Vol.76(3), 553-594.
- Santos, B. de S. (2020). Toward a new legal common sense law, globalization, and emancipation (Third edition.). Cambridge University Press, p. 3
- Комисия в НС претупа скандала с „Фейсбук". Сега [онлайн], 26.01.2023 [прегледан на 20.05.2023]. Достъпен на: https://www.segabg.com/hot/category-bulgaria/komisiya-ns-pretupa-skandala-feysbuk
- Огнянова, Н. Няма да има Министерство на истината в ЕС. Публикувана в: „Дезинформацията: новите предизвикателства", Изд. СУ „Св. Климент Охридски", 2021 (Международна научна конференция - октомври 2021)