Artificial Intelligence in Content Moderation – Legal Challenges and EU Legal Framework

Ralitza Dimitrova
Department of Law and Human Sciences, Faculty of Management
Technical University of Sofia
8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria
{rvd}@tu-sofia.bg

2022 10th International Scientific Conference on Computer Science (COMSCI) | 978-1-6654-9777-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMSCI55378.2022.9912595

Abstract – **The article is dedicated to the concerns and challenges that the use of AI in content moderation poses to law. In the first part the legal issues and challenges discussed in the literature, as well as the proposed solutions, are summarized. In the second part a brief analysis of the current legal framework of content moderation at EU level and in different Member States is offered. The proposal for new EU legislation is also presented focusing on the provisions concerning automated content moderation.**

Keywords – **artificial intelligence, algorithmic decision-making, content moderation, illegal online content, online platforms.**

I. Introduction

Nowadays online platforms and especially social media exert tremendous influence on our lives. Platforms accumulate and disseminate an increasing volume of user-generated content. Unfortunately, part of this content is illegal and has the potential to harm users and societies. Thus, the effective online content moderation becomes indispensable. Adequate content moderation gets even more important when events, which take much attention – like the presidential elections in the USA, the terrorist attacks in France, Germany and UK, the COVID-19 pandemic, war conflicts, etc., occur.

There are plenty of definitions of "moderation" and "content moderation" in the literature – from J. Grimmelman's "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" to Bloch Wehba's "a platform's internal decision-making on whether user-generated content violates its rules, and if so, what the penalty might be" [1].

European Union (EU) law doesn't provide for a harmonized legal definition of "illegal content". The meaning of this term varies from one jurisdiction to another, and from one study to another. Some studies distinguish between illegal content, content that is legal but harmful, and content that is legal but not harmful [2]. Others differentiate illegal content, harmful content and disinformation (fake news), and exclude the last two types of information from their analyses [3]. However, on the basis of the sectoral instruments in the analyzed field, it can be assumed that under EU law four types of content - child sexual abuse/pornography materials; racist and xenophobic content; content constituting a public provocation to commit a terrorist offence; and content violating intellectual property rights, are deemed illegal.

At EU level the European Commission called for the adoption of effective proactive measures to detect and remove illegal online content, and pointed out that the technologies for automated detection and filtering gain increasing role in tackling illegal online content [4]. Recently adopted pieces of national legislation impose on service providers heavy obligations to effectively tackle illegal online content at an early stage within extremely short periods of time (e.g. 24 hours). Given the increasing volume of user-generated content this task, if performed only by humans, seems quite unachievable. Both governments and technological giants are eager to deploy artificial intelligence (AI) in content moderation. Consequently, content moderation increasingly relies on the use of AI systems and, in particular, on automated (algorithmic) decision-making.

Automated content moderation has at least two significant advantages. First, it helps increase the capacity of reviewing large quantities of online content. Second, it takes the burden off the shoulders of human moderators. Without the use of AI systems they bear the hard task to consider an immense volume of disturbing material which affects their wellbeing. Consequently, major online platforms have deployed AI tools in order to tackle harmful content on their platforms both pro and retroactively [5].

Notwithstanding its advantages automated content moderation also raises serious legal and ethical concerns and poses significant challenges to law. Scholars, experts, and institutions outline a variety of legal issues deriving from the development, deployment and use of AI. Lack of transparency and accountability, algorithmic bias, discrimination and unfairness have the potential to infringe fundamental rights of individuals.

Therefore, an adequate legal framework is needed to address the legal concerns and challenges relating to online content moderation and, in particular, to content moderation based on AI systems. The analysis shows that the current legal framework both at EU level and in different jurisdictions is incomplete, fragmented and requires improvement in many aspects in order to effectively tackle illegal content and ensure safe online environment while at the same time safeguard the fundamental rights.

Therefore, studies in the field of online content moderation, especially in the context of use of AI systems, are necessary and useful in order to facilitate the development of the legislation and its application. Such studies are even more important in Bulgaria where the legal framework and its enforcement are under-developed and the online content moderation relies mainly on self-regulation

rules. Recently, the opinions expressed in the social media get very polarized, at the same time the rules on the online content moderation are not well-known by the general public. Consequently, users of social media become infuriated against moderators and argue that moderation is biased, unfair, not transparent, especially due to the suspicion that it is based on automated decision-making without adequate rules for its deployment and use.

The first part of the present study summarizes the different legal issues and challenges relating to automated content moderation as well as some of the proposed solutions.

The second part offers a brief analysis of the current legal framework of online content moderation at EU level and in some Member States, as well as of the proposed new EU legislation in the field.

In the last part conclusions are drawn on the use of AI tools in content moderation and its adequate legal regulation at EU level and in Bulgaria.

II. AUTOMATED CONTENT MODERATION – LEGAL CONCERNS, CHALLENGES AND SOLUTIONS

A. Place of automation in content moderation

Content moderation can be based on decisions which are entirely automated, entirely human-made or a combination of them. Scholars argue that today, the content moderation decisions that platforms often aim to automate, require more contextual analysis and human judgment than simple search for unlawful content [1]. Thus, pre-moderation is usually automated, while post-moderation is usually based on a combination of automated tools and human review [6]. It is believed that AI systems are able to perform successfully some tasks in content moderation, but that human consideration is still needed to differentiate between acceptable and unacceptable online content. In order to achieve accountability, best practices used by the major online platforms aim to obtain „appropriate synergy of human and automated elements" [5].

For example, the dating application Tinder uses AI for content moderation, in particular to scan private messages for inappropriate language, and human moderators to review suspicious profiles, activity and user-generated reports.

The live streaming platform Twitch has long history of using automated moderation tools. Its automated moderator (bot) AutoMod combines machine learning and natural language processing algorithms. AI is used to detect and alert inappropriate content, which is afterwards reviewed by human moderators.

Social media applications like Twitter and Instagram use AI for the content moderation of public posts, although they rely on large teams of human moderators.

B. Legal concerns, challenges and solutions

Online content moderation and, in particular, automated moderation, pose specific legal issues and challenges that should be addressed through adequate legislation measures. The question arises as to how to establish a safe online environment and efficient procedures for noticing and removal of illegal online content while at the same time guarantee the freedom of expression and prevent the "chilling effect of over-removal of content" [7]. In other words, such issues and challenges concern the rule of law and the fundamental rights.

Some authors point out that intermediary liability rules in the United States and in Europe have traditionally conferred too much respect to platforms' rules for governing user speech. Thus, platforms were enabled to elaborate rules and technologies for blocking, filtering, and monitoring user speech [1][3]. The rules of content moderation are often regarded as carrying out a law-like function, setting the limits of participation in an online community and the sanctions for infringement of such rules [1]. Thus states have become dependent on the platforms' d3ecisions [7]. Such broad power of private actors in the field together with the insufficient legislative, judicial and regulatory oversight poses serious threats for fundamental rights of citizens.

The new regulatory approach to encourage platforms to use technology to prevent the spreading of unlawful online content before it is seen or disseminated has received both criticism and praise. Supporters of these initiatives claim that ex ante screening obligations will motivate platforms to encourage more appropriate online speech and that new attempts to regulate platforms' "content moderation" practices restrict technological giants' power by obliging platforms to take the necessary degree of responsibility. The opponents of the new approach argue that such new regulation poses serious risk of infringement of civil liberties, and especially freedom of expression and privacy, and seriously benefits platforms' power [1].

Some authors underline that the use of algorithmic decision-making systems is non-transparent and creates a substantial risk of violation of users' fundamental rights [7] and worsens existing accountability shortcomings [1], and thus undermines foundations of democracy [6].

Besides these major concerns, different reports on the current legal framework and the online platforms moderation practices outline several particular problems: fragmentation of legislation; different content moderation practices in EU Member States; lack of a common definition of "illegal content"; lack of effective rules and measures fit for finding the appropriate balance with fundamental human rights; lack of access to well-established user-friendly 'notice-and-takedown' procedures; emergence of many new types of digital services which were not encompassed when the E-Commerce Directive 2000/31/EC of 8 June 2000 (ECD) [8] was adopted (for example cloud computing, content delivery networks, social media, "sharing economy" services, etc.); dispersing of regulatory competence and oversight for digital services between different sectoral regulators in the Union; inapplicability of ECD to service providers established outside the EU [2] [3].

The analysis shows that the main concerns relating to online content moderation in the light of the current legal framework worldwide encompass the fears that lawmakers and governments have granted too broad power to the online platforms to establish and enforce the rules and technologies for detecting, blocking and removal of illegal online content; the lack of adequate judicial or public supervision over the online platforms' procedures for blocking and removal; the lack of adequate procedures for the users to contest the decision for blocking or removal of content, all of which

lead to an obvious lack of transparency and accountability and risk for the fundamental rights and the rule of law.

Several ways and measures are proposed in the literature to address the legal issues relating to content moderation, and especially that based on algorithmic decision-making. The use of AI systems (and in particular automated filtering technologies) in this field should be governed by adequate rules that ensure transparency and accountability and appropriate legislative and judicial supervision [3][7]. Obligations of service providers to produce transparency reports and disclose information on their engagement in content moderation, as well as provisions for audits by competent authorities should be laid down [1]. Human review over the decisions on the user-generated content and contextual expertise should be guaranteed [2]. Safeguards of users' fundamental rights should be provided [3][7]. Harmonized and transparent 'notice-and-action' procedures, obligation to inform the users about blocking or removal of their content [2][3], as well as procedures that enable users to contest the platforms' decisions on blocking or removal of their content (including the possibility of judicial review) should be introduced [1][6][7]. The scope of the ECD should be updated with a view to the services included, as well as to encompass the service providers established outside EU. The liability regime should be updated too [3].

The analysis of the possible solutions proposed in the literature shows that a detailed legal framework should be introduced at EU level to guarantee transparency and accountability in content moderation. Clear rules should be laid down on the obligation of platforms to produce reports on the content moderation performed by them, to provide users with information about the content moderation procedure, to inform the users about the reasons for the decision to remove or block their content, and to establish possibility for the users to contest the decision and ask for its review by the platform (redress). Besides, the new framework should establish adequate regulatory mechanisms in the field of content moderation and tackling illegal content.

III. Legal framework of online content moderation – current status and future developments

A. Current EU legislation

The current EU legal framework on online content moderation is very complex and contains hard-law, soft-law and self-regulation instruments, vertical rules and horizontal rules.

The hard-law EU instruments in the field of online content moderation include a set of horizontal and vertical rules. The basic legal framework - the E-Commerce Directive, is applicable to all kinds of online platforms and to all kinds of illegal content. These are horizontal rules, among which the most important are the exemption from liability of service providers under certain conditions and the prohibition of general monitoring measures in order to safeguard fundamental rights. Other instruments, the so-called sector- and problem-specific legislation, either apply to a certain kind of platform (the revision of the Audio-Visual Media Services Directive in 2018 extended some audiovisual provisions to video sharing platforms and social media services) or to a certain kind of illegal content (for example Directive (EU) 2017/541 on combating terrorism, Directive 2011/93/EU on combating the sexual abuse and sexual exploitation of children and child pornography, 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, and Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market). These are also vertical rules, which impose additional obligations to certain kind of platforms or tougher rules concerning specific type of content (child pornography, provocation to terrorism, copyright, etc.).

The main ECD provisions relating to online content moderation regulate the exemption of liability of service providers for third party content and the prohibition to establish a general obligation for service providers to monitor the information or actively look for facts or circumstances showing an illegal activity.

The introduction of the exemption of liability regime under the ECD aimed to guarantee the normal functioning of the internal market, the development of cross-border services and the undistorted competition which could be prevented by the disparities in Member States' legislation and case law in the field (recit. 40). The exemption of liability regime lies on the principle of knowledge-based liability [9]. The Member States' national law regulates the prerequisites for liability of information society service providers. However, as far as the so-called liability for third party content is concerned, the ECD provisions exempt three types of information service providers ("intermediaries"), of course, on certain conditions [3] [9].

The conditions under which a service provider can be exempted from liability are differentiated due to the specific nature of the respective information society service. First, under the "Mere conduit" (Art.12) rule, the ECD establishes an exemption from liability in favor of the intermediary service providers for transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network, if the provider refrains from an active role. Second, under art.13 the provider of "caching" services is exempted from liability if it is not involved with the information transmitted and acts expeditiously to remove or to disable access to the information it has stored upon obtaining actual knowledge of the fact that the information has been removed from the network, or access to it has been disabled, or that a court or an administrative authority has ordered such removal or disablement. Third, under art. 14 the provider of "hosting" services is not liable for the information stored at the request of a recipient of the service, on two alternative conditions: (a) the provider is not aware of illegal activity or information and, of facts or circumstances which show an apparent illegal activity or information; or (b) the provider, upon obtaining awareness, acts quicky to remove or to disable access to the information.

According to recit. 42 the exemptions from liability laid down in the ECD cover only activity which is of a mere technical, automatic and passive nature. This means that the service provider must not have knowledge and control over the information which is transmitted or stored.

The liability exemption regime established in the ECD comes to guarantee that the activity of intermediaries and

their role shall not be hindered by the realization of their liability under national laws to an extent which affects the normal functioning of the internal market.

The other core ECD provision relating to online content moderation is art. 15 that prohibits Member States to provide a general obligation for providers, in their activity under Articles 12-14, to monitor the information which they transmit or store, nor a general obligation actively to look for facts or circumstances indicating illegal activity. However, the provision doesn't mean that some monitoring obligations can't be imposed, for example by national authorities under the relevant national law.

The soft-law instruments, which are legally non-binding, include at least two important acts – the Code of Conduct (it is a self-regulatory initiative) and the Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. The Code of Conduct was drawn up in 2016 on the initiative of the Commission. Under the Code some IT companies (Facebook, Microsoft, Twitter and YouTube) committed themselves to prevent and counter illegal hate speech on the Internet. They were joined by Instagram, Snapchat and Dailymotion (2018), Jeuxvideo.com (2019), TikTok (2020), and LinkedIn (2021). The implementation of the Code of Conduct is subject to an evaluation on a regular basis. A set of organizations situated in the different EU countries checks how the IT companies are implementing the commitments laid down in the Code. By the initiative to elaborate and adopt such soft-law instruments the EU has demonstrated its ability and willingness to set the standards in the field. Under the other soft-law instrument - the Commission Recommendation, Member States and hosting service providers, are encouraged to take effective, appropriate and proportionate measures to fight illegal content, in compliance with the principles laid down in the Recommendation and in conformity with the Charter of Fundamental Rights of the EU, in particular the right to freedom of speech and information, and other applicable rules of Union law, in particular those concerning the data protection, competition and e-commerce.

B. National legislation

It should be mentioned that some Member States have recently adopted their own rules on online content moderation, especially as far as hate speech and online disinformation are concerned. Such a step is probably prompted by the loopholes and insufficiencies of the existing relevant EU legal framework and thus may seem justified. On the other hand, often diverging national laws may prove incompatible with the EU rules, bring additional fragmentation in the field and hinder the normal functioning of internal market.

Germany passed the Network Enforcement Act (NetzDG) in 2017 to tackle hate speech and fake news in social networks. On June 28, 2021, an amendment to NetzDG entered into force with the aim to improve the user-friendliness of the reporting mechanisms for complaints about illegal content, to introduce an appeals procedure for measures taken by the social network provider, etc. On May 13, 2020 the French parliament (after a long legislative procedure) passed into law the "Loi Avia" that aimed to fight various forms of online hate speech, terrorist speech and child pornography. On June 18, 2020, the Constitutional court declared key provisions of the bill unconstitutional (among them those on the removal of certain types of illegal content within 24 hours or within one hour, depending on the type of the content). The bill, without the provisions struck down by the Constitutional court, was signed into law by President Macron on June 24, 2020. Besides, on 20 November 2018, after being rejected twice by the Senate, draft ordinary and organic laws on the fight against the manipulation of information have been adopted the French National Assembly. In 2019 in the UK was adopted the Online Harms White Paper which lays down the government's plans for a best worldwide package of online safety measures that also fosters innovation and a robust digital economy. The analysis of the above-mentioned national acts leads to several conclusions. First, the adoption of national laws was obviously a long and difficult process (the UK instrument is not adopted yet). Second, these acts got a lot of criticism by lawmakers and experts and consequently got partially struck down as unconstitutional or underwent serious amendments. Third, such national measures are similar in some respects and differ in others. For example, common features are the establishment of transparency obligations, appeal mechanisms, obligation for removal of illegal content within short periods, provisions on heavy sanctions for non-compliance. On the other hand, differences are seen in their scope - the NetzDG is dedicated to online content which is considered illegal under its provisions if it meets the requirements of certain offences established in the German Criminal Code. The two French laws on manipulation of information apply to information that is objectively false, misleading and threatens the fairness of upcoming elections, together with the requirement that the dissemination of such information should be artificial or computerized, deliberate and large-scale). The UK initiative has broader scope - it encompasses not only illegal, but also harmful speech. As a whole, the analysis shows that national laws on content moderation seek to guarantee effective tackling of illegal content and address the various legal concerns mentioned above. The approaches and measures vary from one Member State to another. The national initiatives seem justified in the absence of harmonized EU-wide rules, but they may exacerbate the existing fragmentation and hinder the enforcement at EU level. However, national laws on content moderation and tackling illegal content can have also some positive impact on the EU legal framework in this field. The DSA can step on the experience already gained at national level. First, the drafts, the adopted acts and their implementation have provoked a wide and useful discussions among lawmakers, experts, digital services providers and users, on the positive and negative features of such pieces of national legislation. Second, the legislative procedures on the adoption of the original texts and their successive amendments, as well as the Constitutional Court review in France, demonstrate what provisions are inacceptable and disproportionate in the light of the safeguard of fundamental rights and particularly the freedom of expression.

Bulgarian legislation on content moderation consists mainly of the relevant provisions in the Electronic

Commerce Act of 2006 [10]. Besides, Bulgaria has adopted measures to transpose the sector- and problem- specific instruments with exception of the Copyright in Digital Single Market Directive. In the Bulgarian legislation there is no legal definition of "illegal content". However, due to the transposition of the above-mentioned four EU instruments, online content relating to terrorism, child pornography and racist speech should be considered illegal. Violation of intellectual property rights is prohibited under civil and criminal law, therefore it can be assumed that online content which infringes intellectual property rights is illegal too. The Electronic Commerce Act transposes the ECD in the Bulgarian law. In compliance with the Directive the exemption of liability regime (art. 13 (access and transmission), art. 15 (caching) and art. 16 (hosting and linking)) as well as the prohibition to impose a general obligation for monitoring of information (art. 17) are introduced. The Bulgarian lawmaker has laid down a separate provision on the exemption of liability of service providers of automated search of information (art. 14). Besides, online platforms have elaborated and published their internal self-regulation rules. Unfortunately, there is not enough relevant case law to be mentioned.

The analysis shows that Bulgaria lags in the field of online content moderation regulation and enforcement. Public and professional discussions on the development of such legislation are not conducted actively. Relatively small group is interested in and has knowledge on the problems of the use of AI systems in content moderation.

C. Proposed EU legislation

In order to address the challenges and concerns relating to content moderation and in particular automated content moderation a proposal was elaborated for a Regulation on a Single Market for Digital Services (Digital Services Act) [11].

The proposal for DSA notes that the requirements for the provision of intermediary services should be harmonized because diverging national legislations have negative impact on the internal market (recit. 2). The new instrument shall encompass particularly "mere conduit", 'caching' and 'hosting' services because of the proliferation of those services (recit.5). The scope of application of the DSA is broadened in comparison to the scope of the ECD. The DSA provisions should be applicable to all intermediary services providers which provide their services in the EU, wherever they are established (recit. 7). The new instrument is created not to replace but to amend and complement the ECD and it should not affect the application of the above-mentioned sector- or content-specific instruments in the field of content moderation. The obligations of service providers are differentiated according to their size, role and influence on the digital market. Special rules are provided for a certain subcategory of providers of hosting services – the online platforms, such as social networks and online marketplaces, as well as for the so-called very large online platforms.

The detailed analysis of the proposal for DSA exceeds the goal and the volume of the present article, but the most relevant provisions should be discussed.

First, definitions of content moderation, illegal content and online platform, are provided which will benefit the enforcement.

Second, the exemption of liability regime is retained – the proposed provisions are very similar to the current ones. The provision stating that on service providers should not be imposed an obligation to monitor the information or actively seek facts indicating an illegal activity is preserved too (art.7 of the proposal).

Third, detailed rules are laid down to guarantee the necessary degree of transparency and accountability and to better protect the rights and interests of users. Intermediary services providers shall be obliged to include in their terms and conditions information on any policies, procedures, and tools used for content moderation, including automated decision-making and human review. Such information shall be formulated clearly and unambiguously and the users should have easy access to it (art. 12, para 1). Providers of intermediary services should produce annual reports, in compliance with the harmonized requirements, on the content moderation they engage in (art. 13). It should be underlined that online platforms shall be obliged to provide in their reports also information on the use of automatic means in content moderation. In particular, they should specify the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and the safeguards which they apply (art.23).

All hosting services providers shall be obliged to lay down and implement user-friendly notice and action mechanisms to enable users to inform the service provider about pieces of information that the user regards as illegal content. Such mechanisms should be based on harmonized rules at EU level (art. 14). Where a hosting services provider decides to remove or disable access to user-generated content, it shall inform users of the decision and clearly explain its reasons. The minimum content of the statement of reasons is listed in detail in para 2. If automated means are used in taking the decision (also where the decision relates to a content which was detected using such means), this fact should be indicated in the statement too (art. 15).

Recipients should be enabled to appeal without difficulty particular online platforms' decisions because the information provided by the recipients is either illegal content or unsuited with its terms and conditions, including decisions for removal, or disabling the access to that user-generated content, through internal complaint-handling system (article 17, para. 1); out-of-court dispute (art. 18, para. 1); judicial redress (recit. 42); lodging a complaint against the provider of intermediary services claiming a violation of the regulation with the Digital Services Coordinator (art. 43).

Fourth, in order to ensure more efficient public oversight and better enforcement, several provisions are included. For example the Digital service coordinators shall have the power to carry on-site inspections and impose fines for infringement of the Regulation (art. 41). Special provisions on supervision, investigation, enforcement and monitoring are laid down regarding the very large online platforms (section 3). Obligation for Member States to lay down effective, proportionate and dissuasive sanctions for violation of the regulation are also laid down (art. 42)

There are already various opinions expressed on the proposal for DSA. De Gregorio recapitulates that the EU initiative can be regarded as a first decisive step towards a new model of content moderation under which online platforms are required to operate as responsible players in the light of their gatekeeping role in the digital environment [6]. Other experts point out that the proposal has some positive elements: it was elaborated taking account of human rights issues, provides clear transparency obligations, and was drafted with the participation of all interested parties. However, for UN Human Rights, some problems remain, for example the risk to provide too broad liability for platforms for user-generated content, and the insufficient judicial oversight [12].

It can be assumed that the DSA addresses in a relatively adequate manner the outlined legal concerns. Its seems that its provisions in the above mentioned four aspects establish improved legal rules for effective content moderation and tackling illegal online content and guarantee better protection of users' rights and interests. More important, the proposed horizontal framework contains specific provisions on the use of AI in content moderation in order to ensure transparency, accountability and protection of fundamental rights. However, only future will show whether the proposed regulation will achieve its goals and solve the issues deriving from the automated content moderation.

II. Conclusion

The automated content moderation has significant advantages in terms of moderation efficiency and wellbeing of human moderators. It is obvious that in future content moderation will increasingly rely on AI systems despite of the legal concerns arising thereof. Although serious and well-grounded, such concerns don't mean that we should give up the use of AI. As it was discussed above, a plenty of solutions have already been proposed. It can be concluded that the deployment of AI in content moderation should be performed on two conditions: (1) the right balance between automated and human content moderation should be found; and (2) specific rules that ensure accountability and transparency and safeguard fundamental rights should be carefully elaborated and adopted.

The shortcomings and the loopholes in the existing EU legal framework, the diverging national law of Member States and the delayed elaboration and adoption of a new EU instrument hinder the effective tackling of illegal online content. Moreover, currently the EU doesn't have the right legal framework to welcome the deployment of AI in content moderation and address the legal concerns arising thereof.

It seems that the DSA is the first step in the right direction. It has the potential to address the legal concerns and challenges posed by the deployment of AI in content moderation. The type of the instrument – a regulation with direct effect in all Member States, its scope – applicability to all service providers irrespective of their place of establishment, as well as its provisions on the use of AI, gives the DSA the chance to become, together with the ECD and the sectoral instruments, an adequate legal framework of automated content moderation and effective tool in tackling illegal online content.

In this context Bulgaria should consider different options – to adopt its national legislation following the example of Germany, France and the UK or wait passively the adoption and the entry into force of the DSA. In the first case, the lawmaker takes the risk to lay down rules, some of which may prove incompatible with the EU law. In the second case, Bulgaria risks to meet the problems deriving from the dissemination of illegal online content without the proper legislation and means to tackle it, relying mainly on the online platforms' self-regulation rules. Which path shall it take?

Acknowledgment

The author would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

References

[1] H. Bloch-Wehba, Automation in Moderation, Cornell International Law Journal, 41 (2020) Vol. 53, pp. 42-96.

[2] A. De Streel et al., Online Platforms' Moderation of Illegal Content Online, Study for the committee on Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2020.

[3] A. Hoffmann, A. Gasparotti, Liability for illegal content online. Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a "Digital Services Act", March 2020, cep Study – Center for European policy.

[4] Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions, Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms, Brussels, 28.9.2017 COM(2017) 555 final.

[5] M. Killeen, Leading platforms keep humans in the content moderation loop, Euractive.com, https://www.euractiv.com/section/politics/news/content-moderation-policies-continue-to-face-core-dilemmas/.

[6] G. De Gregorio, Democratising Online Content Moderation: A Constitutional Framework, 36 Computer Law and Security Review 105374, 2020.

[7] C. Castets-Renard, Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement, JOURNAL OF LAW, TECHNOLOGY & POLICY , Vol. 2020 No. 2, pp.284-322.

[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.

[9] F. Wilman, The EU's system of knowledge-based liability for hosting service providers in respect of illegal user content – between the e-Commerce Directive and the Digital Services Act, 12 (2021) JIPITEC 317 para 1.

[10] Bulgarian E-Commerce Act, State Gazette No. 51 of 23 June 2006.

[11] Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, Brussels, 15.12.2020, COM(2020) 825 final.

[12] OHCHR, Moderating online content: fighting harm or silencing dissent?23 July 2021 https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent