Revija za kriminalistiko in kriminologijo / Ljubljana 75 / 2024 / 4, 290–303

Governing Digital Spaces: Addressing Illegal and Harmful User-Generated Content on Online Platforms

Manja Skočir[1]

The article analyses the European Union's legal framework regulating user-generated content on social media platforms. It begins by presenting the development of content moderation practices on social media platforms and describes the shift from platform self-regulation of user-generated content to more structured legal frameworks. The key focus is on whether the existing legislative framework adequately addresses the social risks arising from user-generated content. The article demonstrates that while the European Union's Digital Services Act (2022/2065) imposes obligations on platforms regarding the removal of illegal user-generated content, the regulation of harmful but legal content remains insufficient, as the moderation of such content is largely left to the platforms themselves. This highlights a gap in the European Union's approach to online safety, particularly considering the unique characteristics of the online environment, where the potential for harm is often amplified in ways that significantly differ from the offline world. In conclusion, the article emphasises the need for a more robust regulatory framework that goes beyond merely aligning online regulations with offline norms. It questions whether the principle that 'what is illegal offline must also be illegal online' adequately addresses the complexity of the digital environment. The article suggests that future regulations should adopt a harm assessment methodology that allows for the proper evaluation of the consequences of harmful but legal content. It stresses the particular importance of focusing on reducing the risks posed by the algorithmic amplification of specific content (which reveals that intermediaries play more than just a neutral role) and highlights the need to acknowledge the broader societal impacts of harmful user-generated content, including harm to third parties.

Keywords: user-generated content, Digital Services Act, platform regulation, intermediary liability, content moderation.

UDC: 34:077

1 Introduction

In November 2022, an unknown attacker brutally stabbed four University of Idaho students to death. The small college town of Moscow, Idaho, with a population of 25,000 – mostly students – had not witnessed a homicide in seven years. With no suspects, no motive and no weapon found, the investigation appeared stalled (Jackson, 2022). In the days following the crime, millions of people worldwide engaged in obsessive efforts, scouring social media platforms for clues. On TikTok alone, videos generated by true crime enthusiasts analysing the case and speculating about who the perpetrator might be garnered nearly two billion views. Some amateur sleuths travelled to Idaho, filming content at the crime scenes, interviewing locals, and speculating about potential suspects. However, the viral sleuthing soon spiralled out of control, as innocent people were falsely accused and had their private information shared online, resulting in harassment that forced some into hiding (Yang, 2022). As local authorities struggled to manage the offline investigation, they also faced the challenge of battling misinformation and rumours spreading across social media platforms. Eventually, the police had to publicly address the online speculations in an effort to curb the harassment of individuals wrongly targeted by internet sleuths (Spring, 2023).[2]

Another incident that highlights growing societal concerns about the inadequate oversight of user-generated content on social media platforms, and the role these platforms

[1] Manja Skočir, LL.M., M.A., Young Researcher, Institute of Criminology at the Faculty of Law Ljubljana, Doctoral Student, Faculty of Law, University of Ljubljana, Slovenia. E-mail: manja.skocir@inst-krim.si

[2] The documentary "The Idaho Murders: Trial by TikTok" explores how social media platforms, particularly TikTok, transformed the investigation of the tragic murders of four University of Idaho students into a viral spectacle. Directed by Zara McDermott and produced by BBC Three, the film investigates the role of amateur sleuths and the broader impact of unchecked user-generated content, including the harassment of innocent people and interference in the police investigation.

play in amplifying harmful behaviour, involves the deaths of several minors who participated in the so-called Blackout Challenge on TikTok. This challenge encourages participants to intentionally restrict their oxygen supply to induce unconsciousness. Many young users, exposed to the challenge through TikTok's algorithm, attempted it, with some succumbing to accidental asphyxiation (French, 2024).

The families of some victims have filed lawsuits against TikTok, arguing that the platform's algorithms bear responsibility for their children's deaths (Paul, 2022). The key legal issue at the heart of these disputes is: Who should bear responsibility for the tragic consequences stemming from user-generated content? Holding the minors accountable is clearly untenable. While the individual who created and uploaded the videos may bear ethical culpability, there has been no legal action against them. Does TikTok share responsibility? The platform not only hosted the content but, as alleged in the legal complaints, actively promoted the dangerous challenges through its algorithm, repeatedly placing them on children's "For You" page, increasing their exposure to harmful – yet not strictly illegal – content.

User-generated content has become a central element of the modern internet, transforming online platforms into spaces of both opportunity and risk. Various service providers, referred to as intermediaries, play a pivotal role in facilitating the exchange of such content.[3] These intermediaries, which include social media platforms, web hosting providers, search engines, and website operators, serve as conduits for user-generated content that spans a wide spectrum—from harmless interactions to clearly illegal (child sexual exploitation imagery, human trafficking material, terrorist propaganda, hate speech) or harmful (but legal or not strictly illegal) materials (misinformation, fake news, manipulative material, promotion of self-harm, extremist views, etc.) (Arora et. al, 2023). Harmful content occupies a grey zone—while it may not breach any specific legal norm, it still causes harm and could, in some cases, be interpreted as illegal. However, as will be discussed later, the European Union's (hereinafter EU) legal framework is limited to addressing illegal content (i.e., content not in compliance with national or EU legislation), leaving harmful but not strictly illegal content largely unregulated.[4] Social media platforms, in particular, have assumed a position of substantial significance due to their extensive reach and unprecedented capacity to disseminate user-generated content on a global scale.

This article is divided into four sections. Following the introduction, the governance of user-generated content by platforms is discussed, followed by an exploration of the European Union's legal framework regulating such content on social media platforms. The article traces the evolution of platform regulation, from self-regulation to increasing legal oversight through frameworks such as the »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) (Text with EEA relevance)« (2022) (hereinafter DSA) and the »Online Safety Act 2023« (2023). It examines how current regulations aim to protect users from illegal and harmful material while balancing free expression and platform accountability. The article addresses two key legal questions: who should bear responsibility for illegal user-generated content, and should regulatory frameworks also cover harmful but legal content?

1.1 The Regulation of User Generated Content

The regulation of user-generated content on platforms began with self-regulation, where platforms exercised their own discretion in managing and moderating content. This approach was solidified by early legal frameworks such as the »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)« (2000) (hereinafter ECD) in the European Union and the »Communications Decency Act of 1996« (1996) (hereinafter CDA) in the United States, which granted intermediaries immunity from liability for third-party content (Floridi, 2021; Klonick, 2017). Legislators sought to foster innovation by shielding platforms from excessive liability, which could have stifled the development of new technologies. As a result, platforms were given the option—but not the obligation—to moderate content provided by their users.

In recent years, however, it has become clear that self-regulation alone is insufficient to protect users and society from harmful content. The ad hoc systems developed by platforms often lack transparency and accountability,

[3] As stated in the Organisation for Economic Co-operation and Development (OECD, 2010) report The Economic and Social Role of Internet Intermediaries (April 2010), intermediaries are described as the "pillars of the Internet" because they bring together or facilitate transactions between third parties online.

[4] This raises an interesting question within the philosophy of law: is it accurate to assume that everything not explicitly illegal is, by default, legal? This assumption, often simplified into a binary distinction, overlooks the nuanced grey areas where certain actions or content, while not strictly illegal, may still be harmful or ethically questionable. The complexity of this legal-philosophical question suggests that the equation "not illegal equals legal" may be overly simplistic.

leading to inconsistent moderation practices. Germany's Network Enforcement Act (»Netzwerkdurchsetzungsgesetz (NetzDG)«, 2017) marked the first significant legal intervention, imposing specific obligations on platforms to moderate harmful user-generated content. Subsequently, newer regulatory frameworks, such as the DSA (2022) in the European Union and the UK »Online Safety Act 2023« (2023), have imposed more stringent controls on intermediaries.

While new regulatory frameworks mark significant progress in platform governance, they still fall short of comprehensively addressing the broader societal harms associated with user-generated content. The focus remains largely on the removal of illegal content, leaving gaps in how harmful but non-illegal behaviours – such as disinformation, fake news, self-harm material, and cyberbullying – are managed. These gaps are particularly concerning for third parties, who may suffer indirect consequences from platform activity that is inadequately regulated or inconsistently enforced. A criminological perspective offers valuable insights for evaluating and understanding these limitations in platform governance. With its focus on harm, criminology (Hillyard & Tombs, 2004) provides a useful framework for assessing existing regulation. Routine activity theory (Cohen & Felson, 1979) could also be applied to digital environments, suggesting that, in the absence of adequate oversight or capable guardianship, online platforms may become breeding grounds for new forms of deviant behaviour. This underscores the need for more robust external regulation to address the societal harms that neither platforms' self-regulation nor existing legal frameworks – which focus solely on prohibiting illegal content while leaving harmful but legal content unregulated – can sufficiently mitigate.

2 Governance by Platforms

The internet was initially envisioned as a space free from speech constraints, where users could engage without external interference. However, it soon became evident that activities conducted online could violate fundamental human rights, facilitate criminal activities and give rise to new forms of deviance. As the internet became flooded with illegal, harmful and misleading content, the idealistic vision of a completely free and open internet began to erode (Wu, 2011). Initially, these challenges were primarily discussed in academic circles, focusing on issues such as privacy, bias, content moderation, intellectual property, fake news, disinformation and the digital divide. However, with the rise of the so-called commercial web, driven largely by user-generated content, these concerns took on new urgency (Floridi, 2021). As online platforms became central to everyday life and commerce, ethical and legal issues transformed into practical regulatory challenges.

The transformation of the internet through user-generated content fundamentally reshaped the digital landscape, allowing individuals to share their voices and build communities like never before. Platforms democratised communication, empowering users to participate actively in the digital public sphere. However, this openness also introduced significant risks, challenging the vision of a completely "open" platform (Kelty, 2014). While platforms, like the broader internet, were initially envisioned as spaces where everyone could freely express themselves (Barlow, 2019),[5] it soon became evident that, as these platforms grew and their influence expanded, they needed to implement systems to manage user-generated content, detect violations and enforce compliance. Effective content governance became essential not only to maintain user trust but also to protect the integrity and growth of these services. Some researchers argue that the growth of social media into a multibillion-dollar industry has largely depended on platforms' ability to regulate user-generated content effectively (Chen, 2014), particularly when compared to the more chaotic and less regulated environments that preceded these governance systems (Edwards, 2009; Lehdonvirta, 2022). Most users simply prefer not to encounter their family photos alongside the most vile content imaginable.

Platform governance, which is a form of self-regulation, emerged as a practical solution to the ethical concerns surrounding user-generated content, which initially consisted primarily of user comments. It encompasses the collection of systems, rules and practices that online platforms implement to manage user interactions and behaviours.

This section will explore the evolution of self-regulatory governance by platforms, focusing on its role, various forms, and the factors driving the increasing need for legal regulation. It will also address the challenges and trade-offs inherent in this type of governance. Although the primary focus of this article is on the legal regulation of platform governance (i.e., governance of platforms) – which concerns how platforms should moderate content rather than directly regulating the content itself – understanding these self-regulatory dynamics is essential for analysing how harmful and illegal content is managed through legal mechanisms.

---

[5] In his famous manifesto "A Declaration of the Independence of Cyberspace", John Perry Barlow (1999) calls on governments to withdraw from regulating cyberspace, which he describes as a free and independent zone, separate from the physical world. The manifesto was written in 1996 in response to the CDA (1996), which sought to regulate content on the internet. In his declaration, Barlow advocates for the internet as a space that should not be governed by traditional mechanisms of state authority, arguing that cyberspace was created as an expression of free thought and should remain independent from old political and social structures.

2.1 Hard and Soft Platform Governance Mechanism

Platform governance began to develop during a period when legal frameworks were either absent or offered minimal guidance on content moderation (providing platforms with immunity from liability and no strict rules on how to manage content, a topic that will be discussed in the next section). In this regulatory vacuum, platforms were left to create their own governance structures in order to address ethical and practical concerns, while also ensuring the growth of their services. The absence of strict legal obligations effectively positioned platforms as both the judges and legislators of content, a role that led to opaque decision-making processes and limited external oversight.

Platforms govern user-generated content and interactions through a combination of soft and hard governance mechanisms (Klonick, 2017). Soft governance mechanisms include design decisions, user interfaces, algorithmic features and other elements that shape user experience without direct intervention, while hard governance mechanisms encompass content moderation practices such as content removal, account suspension and the enforcement of community standards.

While hard governance mechanisms tend to be more visible in their enforcement of explicit guidelines, it is the subtle yet pervasive influence of soft mechanisms that fundamentally shapes user experience and behaviour. These soft mechanisms are not merely a necessity but an intrinsic component of every platform's design. As Gillespie (2018) notes, no platform can operate without some form of structured guidelines. Through elements such as design choices, user interfaces and platform architecture (Gorwa, 2024; Klonick, 2017), platforms inherently shape user behaviour, dictate content boundaries and influence interactions. They act as norm-setters, interpreters of laws, arbiters of taste and adjudicators of disputes, profoundly impacting both individual interactions and broader societal outcomes. Furthermore, they play critical political and gatekeeping roles by determining which topics are open for discussion, defining the boundaries of acceptable behaviour, and establishing criteria for what constitutes spam, hate speech or harassment.[6] These governance methods often reflect broader cultural, political or commercial priorities and can significantly influence not only individual user behaviour but also societal outcomes.

2.1.1 Content Moderation: A Hard Governance Mechanism Within a Soft Law Framework and Its Challenges

Content moderation on platforms functions as a hard governance mechanism, characterised by the strict enforcement of rules and standards on user-generated content. At the same time, as a self-regulatory mechanism, it operates under soft law due to its reliance on self-regulation, allowing platforms to flexibly adapt policies to evolving societal norms and user expectations (Edwards, 2009).

Despite the benefits of self-regulation – such as the ability to manage risks while fostering a user base that feels empowered to participate within predefined boundaries – significant challenges persist. These include the lack of transparency in decision-making, limited accountability to users and the potential for platforms to bypass democratic oversight (Caplan, 2018; Klonick, 2017). The opacity of platform content moderation policies[7] contributes to the perception of objectivity while masking the inherently subjective nature of these practices. This concentration of power allows platforms to act simultaneously as adjudicators – deciding the legality of content – and as legislators, defining what content is acceptable, often without public input. Given that these platforms serve user bases larger than the populations of many nations, their behind-closed-doors governance processes raise serious concerns about the lack of democratic oversight.

Widely publicised content moderation decisions have highlighted significant challenges, particularly inconsistencies and biases in platform enforcement. Examples include Facebook's banning of images depicting breastfeeding mothers (Sweney, 2008), the removal of a photo featuring two fully clothed men kissing (Hudson, 2011), and the deletion of the iconic Vietnam War photograph, the Napalm Girl (Scott & Isaac, 2016), while allowing live streams of shootings on the platform (Graham-McLay, 2019; Ingram, 2016).

Over the past decade, scholars and journalists have highlighted another significant but under-examined issue: the reli-

[6] The motivations behind these governance practices are shaped by commercial interests—emphasising the profit-driven nature of online platforms—as well as political pressures from policymakers, civil society and concerned public groups aimed at mitigating illegal, unsafe or otherwise harmful impacts (Gorwa, 2024).

[7] It is worth mentioning that platform governance is not a monolithic concept. Caplan (2018) distinguishes three major categories of platform companies based on their size, organization and content moderation practices: 1) the artisanal approach, where governance is performed case-by-case by 5 to 200 workers; 2) the community-reliant approach, which typically combines formal policy-making at the company level with volunteer moderators; and 3) the industrial approach, characterised by tens of thousands of workers whose efforts are increasingly supported by automated tools to flag offensive content. The main concerns regarding content moderation are linked to the industrial approach, which is defined by its outsourced and profit-driven nature (Roberts, 2018).

ance on low-paid, outsourced workers from the Global South for content moderation. Major platforms such as Facebook and X employ these workers to manage content, often under severe time constraints and with minimal psychological support (Roberts, 2016). These workers, often based in countries with weaker labour protections and far removed from the platform's headquarters, face poor labour conditions and inadequate mental health support, despite the emotionally taxing nature of moderating vile and shocking material – working to hide it from public view (Chen, 2014; Perrigo, 2023; Roberts, 2016).

In Silicon Values (2021), Jillian York connects platform governance, particularly content moderation, to the broader phenomenon of surveillance capitalism, as defined by Shoshana Zuboff in The Age of Surveillance Capitalism. Zuboff (2019) describes surveillance capitalism as an economic model in which platforms collect, analyse and commodify personal data for profit, often without users' full awareness. Content moderation serves this model by shaping user behaviour to create environments attractive to advertisers, all while platforms monitor and influence public discourse. By creating environments free of objectionable content, platforms ensure user engagement, which is crucial for retaining advertisers and driving revenue within the surveillance capitalism framework (York, 2021).

2.2 The Shift From Governance by Platforms to Governance of Platform

Over the past decade, legislators at both national and supranational levels have increasingly recognised the complex challenges posed by user-generated content in the digital age. As online platforms expanded in scale and influence, concerns over harmful content, privacy violations, misinformation and the protection of fundamental user rights intensified. Legislators responded by seeking to regulate the governance practices that technology companies had previously managed through self-regulation.

This shift was prompted by several high-profile incidents, including the rise of disinformation on social media and its impact on public discourse, the proliferation of hate speech following the 2015 migrant crisis in Europe (Gorwa, 2024), and the (predictable and preventable) Facebook-Cambridge Analytica scandal in 2018 (Floridi, 2021). Despite platforms' attempts to develop hundreds of codes, guidelines, manifestos and public commitments, these self-regulatory efforts proved inadequate in addressing the growing societal risks associated with their operations. It became evident that self-regulation was inadequate, and existing legal frameworks – established around the turn of the millennium – were insufficient to address the scale, complexity and societal impact of modern online platforms. As a result, a regulatory shift emerged, often referred to as the "Brussels Effect," where soft governance mechanisms were increasingly replaced by legally binding compliance measures and penalties (Floridi, 2021). This marked a turning point, as legal acts began to replace platform governance, signalling the end of the era where self-regulation alone could manage the challenges posed by digital platforms.

The following section will first review the legislation enacted in the U.S. and the EU around the turn of the millennium, which, in its largely unchanged form, remained in effect until recently. This legislation not only granted immunity to certain intermediary services but, more importantly, allowed intermediaries to determine – often without transparency – what content was permissible on their platforms. Consequently, online platforms were positioned as both adjudicators and legislators regarding the moderation of user-generated content.

In the subsequent analysis, I will focus on the legal frameworks developed in the EU in response to the growing realisation that platform self-regulation has proven ineffective and that the existing legal structures are inadequate for addressing current challenges.

3 Legal Regulations: The Question of Liability and Definition of Restricted Content

As highlighted before, the regulation of user-generated content raises two pressing legal questions: 1) who should be held accountable for the harmful consequences of user-generated content, and 2) what content should be restricted – only illegal content, or also harmful but legal content? This section explores how both U.S. and EU legal frameworks have evolved to address these issues, focusing on platform liability, the immunity regimes that protect platforms, the challenges related to the definition of restricted content, and the need for modern regulations.

3.1 U.S. Legal Framework

Given the rapid early development of the internet in the United States, it is unsurprising that U.S. legal frameworks were the first to address the regulation of online platforms. U.S. regulations, particularly the CDA (1996) and the Digital Millennium Copyright Act (1998) (hereinafter DMCA), have had an outsized influence on global internet governance, providing a blueprint for many jurisdictions, including the European Union.[8]

[8] For historical and political reasons, the U.S. has established separate liability regimes for internet service providers concerning intellectual property infringements and other types of civil and

A pivotal aspect of both the CDA (1996) and the DMCA (1998) is the inclusion of "safe harbor" provisions. Section 230 of the CDA (1996) and Section 512 of the DMCA (1998) state that no provider or user of an interactive computer service shall be treated as the publisher or speaker of information provided by another content provider. These provisions offer broad immunity to intermediaries, shielding them from liability for user-generated content. Under Section 230 of the CDA (1996), platforms are not held legally responsible for third-party content, yet they are encouraged to moderate harmful material in 'good faith' without fear of legal reprisals. This immunity enabled platforms to host public comments and user interactions without facing catastrophic legal liability, allowing them to innovate and expand rapidly without the constant threat of litigation (Kosseff, 2019).

However, Section 230 has sparked controversy, particularly as platforms have gained unprecedented influence over public discourse. While Section 230 has been instrumental in allowing platforms to flourish, it has become a focal point of debate in the U.S., especially with the rise of misinformation, hate speech and extremist content. Critics argue that the broad immunity granted under Section 230 enables platforms to avoid responsibility for harmful content while wielding excessive power over public discourse without transparency or accountability (Kosseff, 2019). This issue will be revisited in the discussion of EU regulations, where increasing legal oversight seeks to fill the gaps left by self-regulation.

Section 230 has had a global impact, influencing how platforms in other jurisdictions approach content moderation and liability. For example, as Chander (2022) notes, the "International Law of Facebook" demonstrates how global platforms apply U.S.-style content policies to their operations worldwide, often imposing U.S. standards on speech and moderation globally.

3.2 EU Legal Framework

The European Union's legal framework for regulating intermediary platforms began with the adoption of the ECD (2000), which introduced liability exemptions for platforms similar to those found in U.S. legislation, such as Section 230 of the CDA (1996). However, unlike U.S. law, the E-Commerce Directive imposes conditional immunity based on the type of service and the platform's role in managing content. This section explores the E-Commerce Directive and the shift towards a more modernised regulatory approach with the development of the DSA (2022).

3.2.1 E-Commerce Directive

The ECD (2000) was introduced to regulate intermediary services across the European Union and establish a consistent legal framework to promote e-commerce and technological innovation. Much like U.S. regulations, the ECD (2000) shields platforms from liability for third-party content, but with notable differences in the conditions for immunity (Genc-Gelgec, 2022). The Directive grants liability exemptions to intermediaries – mere conduits, caching services and hosting services – provided they meet specific conditions. The liability exemptions are determined based on the type of service provided and the degree of control exercised over the content. Mere conduit and caching services are exempt from liability as long as they remain passive, meaning they do not alter or interfere with the transmission of illegal content. Hosting services, however, are only immune if they act swiftly to remove or disable access to illegal content once they gain actual knowledge of its existence (ECD, 2000).

To maintain immunity, hosting services must meet two key conditions. First, they must act passively,[9] functioning as neutral conduits[10] without altering or interfering with the content they transmit. Second, while they are not required to actively monitor content, they must act promptly to remove or disable access to unlawful content (under national or EU law) once notified of its existence. This process, known as the Notice and Takedown (NTD) mechanism, allows users to notify platforms of illegal content[11] and requires platforms to

criminal liabilities. The Digital Millennium Copyright Act addresses the former, while the Communications Decency Act covers the latter (Damjan, 2010).

[9] The distinction between passive and active intermediaries is often unclear, and the European Court of Justice (2010, 2011, 2022) has, in several rulings, sought to clarify when an intermediary can be considered passive. The Joined Cases C-682/18 and C-683/18 concerned the liability of online platforms for copyright-infringing content uploaded by users. In Frank Peterson v. Google LLC (regarding YouTube) and Elsevier Inc. v. Cyando AG (regarding the file-hosting platform Uploaded), the Court of Justice of the European Union examined whether platforms could benefit from safe harbour protections under Article 14 of the E-Commerce Directive (Court of Justice of the European Union, 2021). The Court of Justice of the European Union held that platforms are not liable if they act as neutral intermediaries without knowledge or control over the uploaded content, but may lose this protection if they play an active role in organizing or promoting infringing content.

[10] Under the general principles of tort and criminal law, providers are liable for illegal content that they themselves produce, as they are not acting in the capacity of intermediaries in such instances.

[11] The ECD (2000) does not impose obligations for the removal of harmful but legal content. The management of such content often falls under platforms' voluntary self-regulation practices.

take action once informed (Angelopoulos, 2016; de Streel & Husovec, 2020).

However, the ECD (2000) does not provide detailed guidance on how this NTD process should be implemented, leaving Member States to regulate the specifics within their domestic laws. This lack of harmonisation has resulted in fragmented rules and procedures for NTD mechanisms across the EU (Genc-Gelgec, 2022). Consequently, platforms also establish their own content moderation and NTD policies, contributing to this fragmentation.

This fragmentation created two major risks. First, platforms often engage in the over-removal (over-censorship) of lawful content. In an effort to avoid potential legal repercussions, many platforms adopt a conservative approach to content removal, erring on the side of caution by taking down content that might be considered problematic (Gillespie, 2018). This conservative stance risks stifling freedom of expression and contributes to a chilling effect on user participation. Second, the absence of harmonised enforcement across Member States has allowed illegal and harmful content to persist on platforms. The lack of consistent legal requirements for swift and uniform content removal enables platforms to delay or avoid taking action, particularly in jurisdictions with weaker enforcement mechanisms (de Streel & Husovec, 2020).

3.2.2 The Need for a New Regulation

The rapid evolution of the internet and digital business models has exposed the limitations of the ECD (2000), which was designed to regulate a vastly different digital landscape. As the digital environment changed, the Directive struggled to keep pace with these developments. Initially, the Directive sought to facilitate the development of a single digital market by harmonising rules for intermediary liability and establishing legal certainty across Member States. However, with the dramatic growth of user-generated content, the proliferation of illegal and harmful content, and the rise of very large online platforms, it became clear that the existing framework was no longer fit for purpose.

One of the core challenges of the ECD (2000) is its failure to achieve harmonisation across Member States, particularly concerning the notice and takedown mechanism. This led to fragmented national interpretations and applications of the law, resulting in inconsistent enforcement of content removal processes across the EU. Germany, for example, implemented stricter rules for content removal under its national »NetzDG« (2017) law, while other Member States applied more lenient approaches, creating disparities in how platforms responded to illegal content notifications. The disparity in national NTD procedures not only created legal uncertainty for platforms but also undermined the Directive's goal of establishing a well-functioning digital single market.[12]

The lack of clear procedural rules in the ECD (2000) left a vacuum that has been filled by platforms themselves, which have become de facto lawmakers in determining what content is permissible online. In the absence of clear guidance from the ECD (2000), platforms were allowed to develop their own content moderation and takedown policies, often with little transparency or accountability (Genc-Gelgec, 2022). This self-regulation resulted in uneven enforcement practices, with platforms exercising vast power over user-generated content, frequently behind closed doors and without external oversight. The imbalance of power has led to growing concerns over the concentration of influence in a few dominant platforms, which effectively regulate public discourse according to their internal policies rather than uniform legal standards (de Streel & Husovec, 2020).

As platforms adopted more advanced technologies for content moderation, such as artificial intelligence, new challenges emerged that the ECD (2000) did not foresee. AI-driven moderation systems, while efficient, introduced risks of errors and biased decisions due to their reliance on algorithms. The ECD's (2000) inability to impose clear due diligence obligations on platforms further compounded these issues, as there were no standard requirements for platforms to ensure the reliability or fairness of their moderation practices.

In response to the growing inadequacies of the ECD (2000) the European Union introduced the DSA (2022) and the »Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act)« (2022) (hereinafter DMA). The DSA (2022) represents a significant regulatory shift, aiming to impose stricter obligations on platforms – particularly very large online platforms (VLOPs)[13] – to safeguard users' fundamental rights and ensure greater accountability in addressing illegal and harmful content. By standardising enforcement practices and establishing due diligence requirements for platforms, the DSA (2022) seeks to rectify the legal fragmentation caused by the ECD (2000) and adapt to the complexities of the modern digital ecosystem (Turillazzi et al., 2023).

---

[12] The German NetzDG (2017) law will be briefly analysed later in this article to illustrate its impact on content regulation within the broader EU context.

[13] Accordning to the DSA (2022) a platform is designated as a VLOP if it has at least 45 million monthly active users in the European Union, which corresponds to around 10% of the EU population.

Even before the DSA (2022), the EU recognized the need for sector-specific regulations, such as the »Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC« (2019), which introduced direct liability for intermediaries hosting copyright-infringing content under Article 17. This Directive serves as a lex specialis relative to the ECD (2000), imposing specific obligations on platforms that host copyrighted works.[14] These earlier legislative efforts highlight the EU's incremental approach to addressing gaps in the ECD (2000), focusing on specific sectors before introducing broader reforms through the DSA (2022) and DMA (2022).

3.2.3 Digital Service Act

The Digital Services Act, proposed by the European Commission in December 2020, passed by the European Parliament in July 2022, and in force since 2024, marks a significant regulatory shift in regulating digital platforms within the European Union. Building upon the ECD (2000), the DSA (2022) addresses its shortcomings, with a primary focus on enhancing platform accountability and transparency. The DSA (2022) forms a core component of the EU's digital strategy, along with the DMA (2022), aimed at fostering fairness, trust and safety in the online ecosystem (Husovec & Roche Laguna, 2023).

The DSA (2022) applies to all digital services targeting the EU market or having a substantial number of EU users, ensuring that large non-EU technology companies, particularly Big Tech, are also bound by its provisions. This reflects the EU's intent to regulate powerful online platforms and mitigate their societal impact.

The DSA (2022) introduces several key innovations designed to address the limitations of the ECD (2000). While retaining the immunity exemptions for intermediaries (such as hosting, caching and mere conduit services) the DSA (2022) established a more structured, tiered system of obligations to increase accountability and transparency for digital service providers, especially VLOPs and very large online search engines (VLOSEs). A central innovation is the introduction of four levels of due diligence obligations (Husovec & Roche Laguna, 2023), aimed at safeguarding the fundamental rights of stakeholders. These obligations directly relate to user-generated content, requiring platforms to take proactive measures to prevent the spread of illegal content and ensure the protection of users' rights. According to Article 3(h) of the DSA (2022), illegal content is defined as any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law.

DSA (2022) categorises its due diligence obligations based on the type, size and impact of the digital service (Chapter III of the directive). These obligations are structured into four levels. First, universal obligations apply to all intermediary services that are eligible for liability exemptions. Second, basic obligations require hosting intermediaries to implement measures that ensure responsible content management. Third, advanced obligations target hosting platforms that exceed the threshold of small firms, recognising their significant role in public information distribution. Finally, special obligations are imposed on VLOPs and search engines, acknowledging the heightened responsibility that accompanies their scale and influence.

The following part will explore the obligations imposed by the DSA (2022), with a specific focus on how they pertain to the management of user-generated content.

Universal due diligence obligations. While there is no overarching obligation to moderate content (as per Article 8), all intermediary services must clearly outline in their terms and conditions whether, and if applicable, how they moderate third-party content, including the use of automated tools. Additionally, intermediaries (except for micro and small enterprises) are subject to transparency obligations (Article 15). They must publish an annual report detailing the orders they receive from authorities and explaining their content moderation practices, particularly in relation to the use of automation (DSA, 2022).

Basic obligations for all hosting services. Hosting intermediaries, responsible for storing third-party information, must implement easily accessible and user-friendly systems for reporting illegal content. Reports must be specific and justified, enabling the identification of illegal material (Article 16). Platforms must process reports promptly, confirm receipt and inform users of their decisions. If content is restricted (re-

[14] In Poland v. European Parliament (C-401/19) the European Court of Justice (2022) confirmed that the Copyright Directive assigns online services providers specific responsibilities regarding copyright protection. Article 17(4) requires platforms to implement effective measures to prevent unauthorized access to copyright content, balancing stakeholder interests and ensuring robust enforcement of intellectual property rights in the digital environment. The Court emphasised that this provision is not only appropriate but essential for the protection of intellectual property rights in the digital environment. It is necessary to balance the interests of various stakeholders and ensure that copyright is effectively enforced online.

moved or suspended), a clear explanation, including the legal basis, must be provided, along with information on appeal mechanisms (Article 17). Furthermore, hosting services must notify authorities when illegal content constitutes a criminal offence (Article 18) (DSA, 2022).

Advanced obligations for online platforms.[15] Online platforms, defined as hosting services that publicly disseminate third-party content, face stricter content moderation and transparency requirements. These platforms must have an internal complaint-handling system for disputes regarding content removal and ensure that the Notice and Takedown mechanisms are accessible to users (Articles 20-21). They must also design their interfaces to prevent manipulative practices that impair users' ability to make informed decisions (Article 25). Additional obligations include ensuring privacy, safety and security, particularly for minors (Article 28), and disclosing how online advertisements target users (Article 26). Online platforms must report biannually on their average number of monthly users and their dispute resolution practices (Article 24) (DSA, 2022).

Special obligations for very large online platforms and search engines. VLOPs, those with more than 45 million active monthly users, face the most stringent content moderation and transparency obligations. VLOPs must create a repository of online advertisements displayed on their interface and provide more frequent transparency reports – every six months – detailing their content moderation processes, including the human resources dedicated to it. These platforms must also give regulatory authorities and vetted researchers access to their data to ensure oversight (Articles 34–40) (DSA, 2022).

VLOPs are obligated to identify, analyse and mitigate systemic risks, including the spread of illegal content, fundamental rights violations and other significant societal impacts (Article 34). To mitigate these risks, VLOPs must implement reasonable, proportionate and effective measures. These include enhancing content moderation practices, refining algorithms and advertising systems, fostering collaboration with trusted flaggers and introducing specific safeguards for vulnerable users, such as minors (Article 35). Additionally, VLOPs are required to regularly submit detailed risk reports (Article 35(2)) and maintain a crisis response mechanism to address extraordinary situations (Article 36). Although the DSA (2022) primarily addresses illegal content, these provisions are also potentially relevant for harmful (but legal) content. However, it remains to be seen how effectively these provisions will be applied to harmful content in practice.

[15] All due diligence obligations applicable to online platforms carve out small and micro firms.

The DSA (2022) also mandates independent audits for VLOPs to ensure compliance with these obligations (Article 37), with a particular focus on content moderation practices, advertising and algorithmic transparency (Articles 39–42). Similar special obligations apply to VLOSEs, mirroring those of VLOPs. However, as the focus of this article is on online platforms, the specific obligations for VLOSEs will not be discussed in detail.

3.3 German Network Enforcement Act (NetzDG)

However, the DSA (2022) was not the first legislative act in the EU to directly impose obligations on platforms to moderate illegal online content. The »NetzDG« (2017), which came into force in 2018, was an earlier and significant piece of legislation aimed at addressing illegal content online. The policy process that led to the »NetzDG« (2017) began in 2015, a period marked by a major influx of refugees, during which Germany experienced a sharp rise in anti-immigration and racist posts (Gorwa, 2024). Politicians and public figures were also frequently harassed online. Initially, accusations were directed at platforms such as Facebook and Twitter, whose content moderation practices were viewed as insufficient. As dissatisfaction with the status quo grew, key German decision-makers pushed for stronger regulatory measures (Genc-Gelgec, 2022).

4 Discussion

More than ever, posting content freely on digital platforms, especially the largest ones, which are now subject to clear obligations regarding the handling of illegal content, is significantly regulated. To maintain immunity under the DSA (2022), platforms must comply with its obligations, including removing or disabling access to content that violates EU or Member State law as soon as they become aware of it. Beyond these legal requirements, platforms also develop their own terms of service, which set additional rules governing the publishing of content and user behaviour on their platforms. The shift towards a more structured regulatory environment marks a gain in terms of legal clarity, platform accountability and the protection of fundamental rights (Genc-Gelgec, 2022).

At the beginning of this article, two examples were introduced where (lawful) user-generated content led to harmful consequences for both platform users (viewing users) and third parties (non-users). Content that caused individuals to flock to the scene of a quadruple homicide and videos hypothesising about the perpetrator remain available. Although TikTok has taken steps to block searches for the fatal Blackout Challenge and issues warnings when specific terms are used,

the challenge continues to circulate on the platform under different names or through alternative methods. Even more troubling is that other harmful challenges still evade detection by reappearing under different aliases, posing a continuous risk to users, particularly minors.

This section critically evaluates the EU's new legislative framework for online intermediaries, particularly how it governs harmful user-generated content. Two central issues emerge: 1) an evaluation of the gains and losses resulting from the DSA (2022) decision to directly regulate the moderation of illegal content while leaving harmful but legal content mostly unregulated, and whether the analogy that "what is illegal offline should also be illegal online" fully captures the complexity of the digital environment; and 2) the issue of focusing harm regulation only on users of online services, without sufficient recognition of how harm can extend to third parties. Challenging the principle that "what is illegal offline should be illegal online",[16] this article proposes adopting the harm assessment methodology developed by Greenfield and Paoli (2022) to improve existing legal frameworks. This approach offers a more nuanced understanding of harm, including the recognition of the diverse parties affected. Such a comprehensive perspective is essential for the DSA (2022) to effectively fulfil its goal of creating a safer online environment for EU users.

4.1 Gains and Losses of Addressing Only Illegal Content: Balancing Harmful Content and Freedom of Speech

While the DSA (2022) provides a standardised approach to the removal of illegal content, harmful but legal content does not fall under the same strict regulatory framework and is generally protected by freedom of expression. It is only addressed indirectly.

Harmful user-generated content encompasses a wide spectrum of materials that, although not necessarily illegal, can inflict significant damage on both individuals – particularly vulnerable groups – and society as a whole. Exposure to idealised (though entirely legal) images on social media platforms has been associated with negative mental health outcomes, particularly among adolescent girls. Research has demonstrated links between this exposure and increased rates of anxiety, depression, body dissatisfaction and, in some cases, a higher risk of suicidal ideation (Fardouly et al., 2015; Perloff, 2014). Repeated exposure to violent or graphic content can lead to desensitisation and emotional distress, potentially fostering aggressive behaviours (Desmurget, 2022). Some content, while not overtly violent, can become highly dangerous when amplified by platform algorithms, as demonstrated by the TikTok challenges described in the introduction (Clark, 2022).

Another significant category of harmful but legal content is disinformation. The spread of COVID-19 conspiracy theories serves as a stark illustration of how unchecked misinformation can have real-world consequences, undermining public trust in democratic institutions and aggravating public health crises. Algorithmic amplification, driven by engagement metrics, ensures that such content reaches large audiences, reinforcing harmful ideas and fostering echo chambers that polarise public discourse. The viral nature of harmful content, combined with recommendation algorithms – studies have shown that misinformation spreads faster than mainstream news – amplifies these adverse effects, posing a serious challenge to both public health and democratic integrity (Vosoughi et al., 2018).

Although the DSA (2022) mandates risk assessments (for content that may affect fundamental rights, democratic processes and public health, particularly regarding the safety of minors) and encourages platforms to manage this type of content (through voluntary codes of conduct, such as The 2022 Code of Practice on Disinformation (European Commission, 2022) it stops short of requiring platforms to remove harmful content unless it is explicitly illegal, and it does not impose any mandatory obligations on how platforms should handle harmful but legal material. As a result, harmful content – protected through freedom of speech – can still pose significant societal risks, especially in an online environment where algorithms amplify its visibility and reach.

Regulating harmful but legal content is complex, as it touches upon the protection of free speech and fundamental rights. By focusing primarily on illegal material, the DSA (2022) seeks to protect these fundamental rights, ensuring that platforms do not over-censor content and thereby infringe on users' freedom of expression. But on the other hand, this approach represents a missed opportunity to address the growing and complex issue of harmful but legal content. The reliance on self-regulation for these types of content risks perpetuating the inconsistencies and lack of accountability that have historically plagued content moderation efforts (as shown in this article). Moreover, by not providing binding legal requirements, the EU leaves the door open for platforms to prioritise commercial interests over the safety and well-being of their users, particularly when it comes to non-illegal content that can still have harmful effects.

[16] This position has been reiterated multiple times during the adoption process of the Digital Services Act, including in the following statement by the Council of the European Union (2021): What is illegal offline should be illegal online: Council agrees on position on the Digital Services Act.

The UK's »Online Safety Act 2023« (2023) takes a direct approach to addressing harmful content, building on the Online Harms White Paper and culminating in the Online Safety Bill. While the final definition of harm in the act is narrower than in earlier drafts (Colegate, 2023), it still defines harm as both physical and psychological (Section 234). The act primarily mandates that platforms, particularly those classified as Category 1 services (similar to VLOPs under the DSA, 2022), implement measures to mitigate the spread of harmful content, such as disinformation, self-harm material and cyberbullying, particularly when such content affects children or vulnerable users.

Some critics of the »Online Safety Act 2023« (2023) warn that the broad definitions of harmful content could lead to excessive censorship, posing a threat to fundamental rights, particularly freedom of speech (Bliss, 2022), while others argue that these provisions are too vague to deliver the act's promise of making the UK "the safest place in the world to go online" (Colegate, 2023).

4.2 The Problematic "Offline/Online" Analogy

While the principle that "what is illegal offline should also be illegal online" offers a sound regulatory foundation by aligning digital spaces with established legal norms, it also oversimplifies the complex and unique nature of the digital environment. This environment has some special characteristics that makes this analogy problematic.

In contrast to the offline world, where illegal or harmful activities are often limited by physical and social barriers, the online world operates under different dynamics. One key difference lies in the viral potential of online content. Algorithmic amplification enables harmful material to spread rapidly, reaching millions within a matter of hours (a scale and speed unimaginable in offline settings). This exponential spread heightens the risk of social harm, particularly when harmful content such as disinformation, cyberbullying, fake news or dangerous challenges reaches vast audiences. For instance, algorithmically driven disinformation campaigns have been shown to significantly influence public discourse, undermine democratic processes and foster social instability, and to reach more readers than mainstream news (Vosoughi et al., 2018).

Moreover, the anonymity afforded by many online platforms exacerbates this problem by shielding perpetrators from accountability. Unlike in the physical world, where individuals can often be more easily identified and held responsible for their actions, the digital environment allows harmful actors to operate with relative impunity. This complicates enforcement efforts, as identifying and addressing harmful content becomes more challenging when users remain anonymous.

Additionally, the business model of many platforms incentivises engagement, often prioritising content that provokes strong emotional responses, regardless of its harmful nature. This economic structure starkly contrasts with traditional media environments, where regulatory oversight and ethical standards generally act as barriers to the dissemination of harmful content. The continuous and pervasive exposure to harmful ideas online, whether through disinformation, radicalisation or harmful challenges, poses long-term risks to public health, societal cohesion and democratic processes. The far-reaching impacts of these risks – such as influencing elections, inciting violence or contributing to mental health crises – underscore the need for more stringent regulatory frameworks in the digital realm, frameworks that account for the distinct ways in which online harm can proliferate and persist beyond control. Social media platforms, driven by user engagement metrics, create echo chambers that reinforce harmful ideas, making it increasingly difficult for users to escape or counteract their influence. This also challenges the notion of platforms as neutral intermediaries. Can platforms that prioritise content based on engagement metrics, often amplifying harmful but legal material, truly be considered neutral when their algorithms expose users to harmful content that shapes public discourse and social behaviour? Can we, in fact, define the algorithm as the platform's language – one that actively 'speaks' to users through content selection – and hold the platform liable for this language and its consequences?

In light of these differences, we can assert that the digital environment requires a more tailored and rigorous regulatory approach, one that not only addresses illegal content but also the unique risks associated with harmful, yet legal, material. The DSA (2022), in its current form, does not adequately respond to these challenges, leaving significant gaps in the protection of users and society from the far-reaching consequences of harmful, though legal, digital content. A more comprehensive regulatory strategy must take into account not just the content but also the technological infrastructure that amplifies and perpetuates such harm in the digital environment.

4.3 Harm Beyond Users: The Impact on Third Parties

Another critical aspect that demands attention in understanding online harm is that it is not confined to direct users or those engaging with harmful content, even though both legislation and literature predominantly focus on individual viewing users. The spread of harmful material can have far-reaching impacts on third parties—individuals who may never engage with the platform but still suffer the consequences

of viral content. In this context, harm extends beyond the screen, disrupting social and public life offline.

A striking example of harm beyond users is the phenomenon of "TikTok Frenzies", one of which was already presented in the introduction. These frenzies illustrate situations where, due to algorithmic amplification and recommendation systems, a large number of users engage with certain content, leading to real-world consequences. For instance, these have included involvement in criminal investigations, obstructing the course of investigations, violent school protests, and mass offline gatherings that result in property damage, injury and other forms of public disruption (Spring, 2023).

These examples underscore the need to rethink the DSA's (2022) ability to fully address the societal risks posed by harmful, albeit legal, content and how we assess and mitigate harm in digital spaces. It seems that current frameworks, such as the DSA (2022), fall short in addressing this extended harm. While the DSA (2022) mandates risk assessments for illegal content, it does not sufficiently account for how harmful content is amplified by algorithms, nor does it fully tackle the societal risks that such amplification can cause.

5 Conclusion

The optimistic narratives promoted by tech companies often mask the profound risks embedded within their platforms. As new digital spaces emerge, they introduce unprecedented dangers that challenge existing regulatory frameworks. The necessity for a robust legal response is clear, requiring not only regulatory measures from platforms and large tech companies but also cohesive national and supranational policies. While initial steps have been taken through legislative instruments such as the UK's »Online Safety Act 2023« (2023) and the DSA (2022) and DMA (2022) – grounded in the principle that what is illegal offline should also be illegal online – these efforts may still prove insufficient. The unique nature of digital spaces introduces risks that may not have direct parallels in the physical world, raising the question of whether actions not considered illegal offline should, nonetheless, be prohibited in virtual environments.

A particularly promising approach from criminology is the harm assessment methodology proposed by Greenfield and Paoli (2022). This framework offers a structured way to systematically identify and evaluate harms by considering both the direct and indirect effects of online content. The methodology, which has been successfully applied in other regulatory contexts such as drug regulation, helps assess harm by focusing on the severity, scale and scope of its impact – whether it be physical, psychological or societal. By using this approach, future legal frameworks can more effectively address the complexities of digital environments, where harmful content can be algorithmically amplified, leading to widespread harm that is difficult to contain.

The harm assessment methodology could be adapted to digital platforms to bridge the gap between illegal and harmful but legal content. By focusing on harm reduction, this approach offers a way to design regulations that are both proportionate and effective, ensuring that the unique risks posed by digital platforms are managed responsibly. Regulation must evolve not only to prevent illegal behaviour but also to mitigate harm – whether direct or indirect – ensuring that the digital realm is a safer space for all, including those beyond direct platform users.

References

1. Angelopoulos, C. (2016). *European intermediary liability in copyright. A Tort-Based Analysis.* Wolters Kluwer.
2. Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2023). Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Computing Surveys, 53*(3), 1–17.
3. Barlow, J. P. (2019). A declaration of the independence of cyberspace. *Duke Law & Technology Review, 18*, 5–7.
4. Bliss, L. (22. 3. 2022). Online Safety Act: Ambiguous definitions of harm could threaten freedom of speech – Instead of protecting it. *The Conversation.* https://theconversation.com/online-safety-bill-ambiguous-definitions-of-harm-could-threaten-freedom-of-speech-instead-of-protecting-it-179514
5. Caplan, R. (2018). *Content or context moderation? Artisanal, community-reliant, and industrial approaches.* Data & Society Research Institute.
6. Chander, A. (2022). Section 230 and the international law of Facebook. *Yale Journal of Law & Technology, 24*, 209–256.
7. Chen, A. (23. 10. 2014). The laborers who keep dick pics and beheadings out of your Facebook feed. *Wired Magazine.* https://www.wired.com/2014/10/content-moderation/
8. Clark, M. (8. 7. 2022). The TikTok 'blackout challenge' has now allegedly killed seven kids. *The Verge.* https://www.theverge.com/2022/7/7/23199058/tiktok-lawsuits-blackout-challenge-children-death
9. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608.
10. Colegate, E. (2023). The lost clause – Exploring the potential impact of amendments to the definition of harm to children in the UK's Online Safety Bill. *Communications Law, 28*(4). https://nottingham-repository.worktribe.com/output/30146986/the-lost-clause-exploring-the-potential-impact-of-amendments-to-the-definition-of-harm-to-children-in-the-uks-online-safety-bill
11. Communications Decency Act of 1996. (1996). *U.S.C. §*, (230/1996).

12. Council of the European Union. (25. 11. 2021). *What is illegal offline should be illegal online: Council agrees on position on the Digital Services Act.* https://www.consilium.europa.eu/en/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act/
13. Court of Justice of the European Union. (2021). CJEU Joined Cases C-682/18 and C-683/18 / Judgment, Frank Peterson v Google LLC and Others and Elsevier Inc.v Cyando AG dated June 22, 2021. https://fra.europa.eu/sl/caselaw-reference/cjeu-joined-cases-c-68218-and-c-68318-judgment
14. Damjan, M. (2010). Odškodninska odgovornost internetnih posrednikov [Liability of internet intermediaries]. *Pravni letopis, 1,* 139–155.
15. De Streel, A., & Husovec, M. (2020). *The e-commerce Directive as the cornerstone of the internal market: Assessment and options for reform.* Policy Department for Economic, Scientific and Quality of Life Policies, Directorate-General for Internal Policies, European Parliament.
16. Desmurget, M. (2022). *Screen damage: The dangers of digital media for children*. Polity Press.
17. Digital Millennium Copyright Act. (1998). *U.S.C. §,* (512/1998).
18. 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). (2000). *Official Journal of the European Communities,* (178/1).
19. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Copyright Directive). (2019). *Official Journal of the European Communities*, (130/92).
20. Edwards, L. (2009). The fall and rise of intermediary liability online. In L. Edwards, & C. Waelde (Eds.), *Law and the internet* (pp. 47–88). Hart Publishing.
21. European Commission. (2022). *The 2022 Code of practice on disinformation.* https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
22. European Court of Justice. (2010). Google France SARL v. Louis Vuitton Malletier, C-236/08 and C-238/08, 2010 E.C.R. I-02417 dated March 23, 2010. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62008CJ0236
23. European Court of Justice. (2011). L'Oréal SA v. eBay International AG, C-324/09, 2011 E.C.R. I-06011 dated July 12, 2011. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62009CJ0324
24. European Court of Justice. (2022). Poland v. European Parliament and Council of the European Union, C-401/19, EU:C:2021:596 dated apRIL 26, 2022. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62019CJ0401
25. Fardouly, J., Diedrichs, P. C., Vartanian, L. R., & Halliwell, E. (2015). Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. *Body Image, 13*, 38–45.
26. Floridi, L. (2021). The end of an era: From self-regulation to hard law for digital industry. *Philosophy & Technology, 34*, 619–622.
27. French, D. (5. 9. 2024). Opinion: The viral blackout challenge is killing young people. Courts are finally taking it seriously. *The New York Times.* https://www.nytimes.com/2024/09/05/opinion/tiktok-blackout-challenge-anderson.html
28. Genc-Gelgec, B. (2022). Regulating digital platforms: Will the DSA correct its predecessor's deficiencies? *Croatian Yearbook of European Law and Policy, 18*, 25–60.
29. Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
30. Gorwa, R. (2024). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Oxford University Press.
31. Graham-McLay, C. (2. 5. 2019). Death toll in New Zealand mosque shootings rises to 51. *The New York Times.* https://www.nytimes.com/2019/05/02/world/asia/new-zealand-attack-death-toll.html
32. Greenfield, V. A., & Paoli, L. (2022). *Assessing the harms of crime: A new framework for criminal policy.* Oxford University Press.
33. Hillyard, P., & Tombs, S. (2004). Beyond criminology? In P. Hillyard, C. Pantazis, S. Tombs, & D. Gordon (Eds.), *Beyond criminology: Taking harm seriously* (pp. 10–29). Pluto Press.
34. Hudson, J. (22. 4. 2011). The controversy over Facebook's gay kissing ban isn't over. *The Atlantic.* https://www.theatlantic.com/technology/archive/2011/04/controversy-over-facebooks-gay-kissing-ban-isnt-over/349921/
35. Ingram, M. (7. 7. 2016). Facebook live streams the death of a black man shot by police. *Fortune.* https://fortune.com/2016/07/07/facebook-live-death/
36. Jackson, P. (19. 11. 2022). University of Idaho students stabbed to death in their beds. *BBC.* https://www.bbc.com/news/world-us-canada-63689176
37. Kelty, C. M. (2014). The fog of freedom. In P. J. Boczkowski, T. Gillespie, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 195–220). The MIT Press.
38. Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985
39. Kosseff, J. (2019). *The twenty-six words that created the internet.* Cornell University Press.
40. Lehdonvirta, V. (2022). *Cloud empires: How digital platforms are overtaking the state and how we can regain control.* MIT Press.
41. Netzwerkdurchsetzungsgesetz (NetzDG) [Network Enforcement Act]. (2017). *BGBI,* (149/2017).
42. Online Safety Act 2023. (2023). *UK Public General Acts*, (c. 50).
43. Organisation for Economic Co-operation and Development (OECD). (2010). *The economic and social role of internet intermediaries.* OECD Publishing. https://www.oecd.org/internet/ieconomy/44949023.pdf
44. Paul, K. (6. 6. 2022). Families sue TikTok after girls died while trying 'blackout challenge. *The Guardian.* https://www.theguardian.com/technology/2022/jul/05/tiktok-girls-dead-blackout-challenge
45. Perloff, R. M. (2014). Social media effects on young women's body image concerns: Theoretical perspectives and an agenda for research. *Sex Roles, 71*(11-12), 363–377.
46. Perrigo, B. (18. 1. 2023). OpenAI used Kenyan workers on less than $2 per hour: Exclusive. *Time.* https://time.com/6247678/openai-chatgpt-kenya-workers/
47. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act). (2022). *Official Journal of the European Union*, (265/1).
48. 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) (Text with EEA relevance). (2022). *Official Journal of the European Union*, (277/1).

49. Roberts, S. (2016). Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications, 12*. https://ir.lib.uwo.ca/commpub/12/
50. Scott, M., & Isaac, M. (9. 9. 2016). Facebook restores iconic Vietnam war photo it censored for nudity. *The New York Times*. https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html
51. Spring, M. (20. 9. 2023). Inside Tiktok's real-life frenzies – From riots to false murder accusations. *BBC*. https://www.bbc.com/news/technology-66719572
52. Sweney, M. (30. 12. 2008). Mums furious as Facebook removes breastfeeding photos. *The Guardian*. https://www.theguardian.com/media/2008/dec/30/facebook-breastfeeding-ban
53. Turillazzi, A., Taddeo, M., Floridi, L., & Casolari, F. (2023). The Digital Services Act: An analysis of its ethical, legal, and social implications. *Law, Innovation and Technology, 15*(1), 83–106.
54. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
55. Wu, T. (2011). *The master switch: The rise and fall of information empires*. Vintage.
56. Yang, M. (25. 12. 2022). Professor sues TikTok accuser for linking her to Idaho students' murders. *The Guardian*. https://www.theguardian.com/us-news/2022/dec/25/tiktok-sleuth-ashley-guillard-rebecca-scofield-defamation-idaho-murders
57. York, J. (2021). *Silicon values: The future of free speech under surveillance capitalism*. Verso.
58. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Upravljanje digitalnih okolij: Obravnava nezakonite in škodljive vsebine, ki jo ustvarijo uporabniki na spletnih platformah

Manja Skočir, mag. prav., mlada raziskovalka, Inštitut za kriminologijo pri Pravni fakulteti v Ljubljani, doktorska študentka, Pravna Fakulteta, Univerza v Ljubljani, Slovenija. E-pošta: manja.skocir@inst-krim.si

V prispevku je analiziran pravni okvir Evropske unije za urejanje vsebin, ki jih na platformah družbenih medijev ustvarjajo uporabniki. Začne s prikazom razvoja praks moderiranja vsebin na platformah družbenih medijev in opiše prehod od samoregulacije uporabniških vsebin s strani platform k bolj strukturiranim pravnim okvirom. Ključni poudarek je na tem, ali obstoječi zakonodajni okvir ustrezno obravnava družbena tveganja, izhajajoča iz vsebin, ki jih ustvarijo uporabniki. Članek pokaže, da zakonodaja Evropske unije - Akt o digitalnih storitvah (DSA) - platformam sicer nalaga obveznosti v zvezi z odstranjevanjem nezakonite vsebine, ki jo ustvarijo uporabniki, pri urejanju škodljive, vendar zakonite vsebine, pa je zakonodaja pomanjkljiva, saj je moderiranje takšne vsebine v veliki meri prepuščeno platformam samim. To kaže na vrzel v pristopu Evropske unije k spletni varnosti, zlasti glede na edinstvene značilnosti digitalnega okolja, v katerem je možnost škode pogosto povečana na načine, ki se bistveno razlikujejo od analognega sveta. Članek v zaključku poudarja potrebo po robustnejšem regulativnem okviru, ki presega zgolj uskladitev spletnih predpisov z normativi zunaj spleta. Preizprašuje, ali načelo, da, mora biti "vse, kar je nezakonito v analognem, nezakonito tudi v digitalnem okolju", v zadostni meri obravnava zapletenost digitalnega okolja. Članek predlaga, da se v prihodnjih predpisih sprejme metodologija ocenjevanja škode, ki omogoča ustrezno presojanje posledic škodljivih, vendar zakonitih vsebin. Poudarja, da se je še posebej pomembno osredotočiti na zmanjševanje tveganj, ki jih povzroča algoritemsko ojačevanje določenih vsebin (ki razkriva, da posredniki igrajo več kot zgolj nevtralno vlogo), ter izpostavlja potrebo po obravnavi širših družbenih učinkov škodljivih vsebin, ki jih ustvarjajo uporabniki, vključno s škodo tretjim osebam.

Ključne besede: vsebine, ki jih ustvarjajo uporabniki, Akt o digitalnih storitvah, regulacija platform, odgovornost posrednikov, moderiranje vsebine

UDK: 34:077