Content Moderation: How the EU and the U.S. Approach Striking a Balance between Protecting Free Speech and Protecting Public Interest

Rrita Rexhepi*

Abstract: The topic of content moderation is becoming increasingly relevant, as we are in an era of acute politicization and social media are now used to achieve political goals . This means that regulation is necessary to preserve democratic standards and simultaneously encourage a healthy online environment. This article aims at analyzing and comparing how content sharing is regulated respectively in the EU and U.S. and at identifying the benefits and shortcomings of both methods. It does so by using information from government agencies, social media companies, and specific cases which reflect the policies in both regions. It is evident that while both the U.S. and the EU have taken steps to regulate online content, there are significant differences. The EU chooses a more centralized approach and values the protection of users and public interest, whilst the U.S. adopts a more decentralized approach and tends to opt for the protection of free speech. Lack of transparency, over-removal, under-removal, and vague social media standards are the difficulties that both the EU and the U.S. face in regulating online content. This article recommends potential answers to these problems, including regulating platform transparency, increasing accountability, and establishing oversight bodies. Moreover, platforms are encouraged to invest in their content moderation policies by using higher-level means of finding and removing harmful content.

Keywords: content moderation; internet governance; censorship; Section 230; Digital Services Act.

1. *Introduction*

The tension between the fundamental right of free speech and the responsibility to protect public interest has become increasingly relevant in the field of digital communication, prompting both the EU and the U.S. to grapple with finding a balance between the two principles in their respective frameworks. The need for content moderation has become significant due to the rapid growth of the internet and the increasing amount of information shared online. Content moderation refers to the process of reviewing and regulating user-generated content on social media platforms followed by the removal of posts that are viewed as harmful or go against Community Standards[1]. This may include graphic, sexual, or violent content, as well as disinformation released or circulated by political figures. Considering the sheer volume of content generated by billions of users daily, and the ease with which it can be disseminated, the need for effective moderation has become essential to ensure a safe and trustworthy online environment. In recent years, tech companies have begun to take the issue more seriously. For instance, Facebook has launched an oversight board, dubbed often as Facebook's "supreme court", which is entrusted with reviewing specific content decisions made by moderators[2].

---

* Rrita Rexhepi is currently in her second year of law school at the University of Trento, where she is studying Comparative, European, and International Legal Studies. Prior to law school, Rrita actively participated in several local initiatives and projects organized by international and local organizations, as well as local governments in Kosovo. Her fields of interest include international and economic law, where she aspires to specialize upon graduation.

1. See Jennifer Grygiel and Nina Brown, *Are social media companies motivated to be good corporate citizens? Examination of the connection between corporate social responsibility and social media safety*, 43(5) Telecommunications Policy, 445–460 (2019).

2. *Independent judgment. transparency. legitimacy. Oversight Board*, available at https://www.oversightboard.com/.

Issues relating to content moderation have proven to be problematic for the European Union, which recognizes freedom of expression as a right protected by the Charter of Fundamental Rights (hereinafter: CFREU) under Article 11[3], while also claiming a responsibility to protect the public interest against hate speech[4], and disinformation[5]. The EU has not adopted any strict limit to the use of the term "public interest" but has established it as a potential ground for the restriction of one of the fundamental freedoms guaranteed by EU law[6]. In *Omega* (C-36/02), the ECJ held that public interest can be used by Member States to justify restrictions to the free movement of goods under the public policy exception provided in Article 36 of the Treaty on the Functioning of the European Union (TFEU)[7], which reads that the provisions of the previous Articles 34 and 35 shall not preclude prohibitions or restrictions on imports, exports or goods in transit justified on grounds of public morality, public policy or public security. Although the European Union does not employ a precise definition of public interest, it may be frequently found in its legislative acts. For instance, the General Data Protection Regulation (GDPR), which aims to protect users from the unlawful processing of data, affirms that controllers may process data if it is necessary for the performance

3. Art. 11, Charter of Fundamental Rights of the European Union, 7 June 2016, C 202/405, available at https://www.europarl.europa.eu/charter/pdf/text_en.pdf.

4. Council of the European Union, *Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*, November 28th, 2018, 913/JHA, available at https://eur-lex.europa.eu/eli/dec_framw/2008/913/oj.

5. See European commission, *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Online Disinformation: A European Approach*, COM/2018/236 (April 26, 2018), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236. See also European Commission, *European Democracy Action Plan* (2020), available at https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en.

6. Alexander J. Belohlavek, *Public Policy and Public Interest in International Law and EU Law,* 3 Czech Yearbook of International Law: Public Policy and Ordre Public, 117-147 (2012).

7. C-36/02, *Omega Spielhallen- und Automatenaufstellungs-GmbH contro Oberbürgermeisterin der Bundesstadt Bonn*, ECR 2004 I-09609

of a task carried out in the public interest[8]. This exemption is, however, subject to safeguards to ensure that processing is indeed necessary and proportionate to the public interest it relates to. After the COVID-19 outbreak, the European Data Protection Board adopted a set of guidelines that permitted controllers to process health data for scientific research based on public interest, stating that the EDPB considers that the fight against COVID-19 has been recognized by the EU and most of its Member States as an important public interest, which may require urgent action in the field of scientific research[9]. By using the terms "important public interest" and "urgent action", the EDPB highlights the use of assessing necessity and proportionality to balance personal interest and public interest.

While legislators found it less difficult to reach a consensus on relaxing certain protections (such as those on data processing) for public health, regulating free speech presents a more challenging task. This, inasmuch as what may be considered harmful speech to some, may be viewed as protected speech by others. The subjective nature of deciding where the limits of free speech lie have also proven to be a difficulty for content moderators. This is subsequently compounded by the fact that online platforms have global reach and must navigate the differences in cultural, legal, and political norms present in several countries.

An additional problem for the EU is regulating tech companies often based outside the region. These companies are subject to their home country's laws, which may not align with EU regulations and standards. The EU has recognized the need to effectively regulate these companies to ensure that they are taking the necessary measures to protect the public interest and promote responsible content moderation practices, and has attempted to address these challenges through regulations such as the e-Commerce Directive, the Digital Services Act (DSA), and the Audiovisual Media Services Directive (AVMSD), which aim to provide a framework for content moderation while

8. Art. 6, *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/European Commission, (General Data Protection Regulation)* April 27 2016, no. 679.

9. Art. 63 par. 7, *Guidelines for COVID 19 health data processing*, April 21 2020, no. 3.

balancing the protection of freedom of expression. These acts will be discussed ahead.

Content moderation has also become a relevant issue in the United States, especially post-Covid-19. Government efforts to regulate content moderation have mostly been conducted at the State level, although there have been talks about reforming Section 230 of The Communications Decency Act passed by Congress in 1996, which holds that companies are not liable for the content published on their platforms[10]. Similar to the CFREU, the U.S. Constitution also protects freedom of speech in its First Amendment. This protection is deeply ingrained in American constitutional culture and is seen as a cornerstone of democratic values, which results in any act attempting to reduce the threshold being met with some degree of scrutiny[11]. It is important to note, however, that private corporations are not bound by this and can remove any content, which has led to debates about the role of private companies in regulating speech online[12]. However, most mainstream sites (Facebook, Twitter, YouTube) have developed their own policies regarding content moderation, usually employing fact-checker programs to combat misinformation[13]. The Cambridge Analytica scandal[14], and foreign intervention in elections online, including the alleged use of Russian bots in campaigning and spreading disinformation[15], have further highlighted the need for effective content moderation for platforms.

---

10. Communications Decency Act, S.314(1995), available at https://www.congress.gov/bill/104th-congress/senate-bill/314.

11. Robert Allen Sedler, *An essay on freedom of speech: The United States versus the rest of the world*, 2 Mich. St. L. Rev. 377 (2006).

12. The First Amendment only applies to government action and Independence of platforms in regulating the content they allow is guaranteed by Section 230 of The Communications Decency Act.

13. See Facebook, *About Facebook Ads: Ad targeting options* available at https://www.facebook.com/business/help/259586717571940?id=673052479947730 and Google, *Choose where your ads appear on YouTube* available at https://support.google.com/youtube/answer/9229632?hl=en.

14. Antonio Peruzzi, Fabiana Zollo, Walter Quattrocchi and Antonio Scala, *How news may affect markets' complex structure: The case of Cambridge Analytica*, 20(10) Entropy 765 (2018).

15. Darin E. W. Johnson, *Russian election interference and race-baiting*, 9(2) Columbia Journal of Race and Law 191-213 (2019).

Despite the fact that the U.S. places significant importance on personal freedoms, it has also enacted laws aimed at protecting the public interest, even when such measures have entailed a degree of personal cost. One of the most important (and arguably most controversial) of such legislation is the PATRIOT Act of 2001, which was adopted after the 9/11 attacks to increase counterterrorism efforts and defend public security. Some of the provisions of the PATRIOT Act, such as the authorization of "roving wiretaps"[16], were believed to be infringing upon privacy, but national security concerns were so high that they trumped certain privacy protections[17]. Regarding free speech, in particular, the Supreme Court, in the landmark decision of Brandenburg v. Ohio, held that speech that is directed to inciting or producing imminent lawless action and is likely to incite or produce such action is unlawful and cannot be protected by the First Amendment[18]. In other words, speech that incites or brings about violence does not fall under the First Amendment and is not considered free speech. More recently, in 2021, the COVID-19 Consumer Protection Act was passed and made any disinformation regarding the virus unlawful[19]. It is evident, then, that there are situations in which the U.S. government is willing to restrict freedoms to protect the public and national interest.

2. *The European Union: A Toolbox for Content Moderation*

The EU has been involved in attempting to regulate different aspects of online content, initially through the e-Commerce Directive which was adopted in 2000. The e-Commerce Directive established a legal framework for online service providers and their responsibilities for the content they host but, due to its status as a directive, gave space for Member States to expand on the rules as they pleased,

16. Roving wiretaps are wiretaps that follow specific surveillance targets across private communications, instead of specific devices.

17. John T. Soma, M. M. Nichols, Stephen D. Rynerson, Lance A. Maish, Jon David Rogers, *Balance of Privacy vs. Security: A Historical Perspective of the USA PATRIOT Act*, 31 U.B.C. Law Review, 285 (2005).

18. *Brandenburg v. Ohio*, 395 U.S. 444 (1969).

19. *COVID-19 Consumer Protection Act of the 2021 Consolidated Appropriations Act*, Pub. L. No. 116-260, 134 Stat. 1182, Division FF, Title XIV, § 1401.

thereby affecting the internal market[20]. The e-Commerce Directive did not explicitly refer to online platform regulation, although it did stipulate that platforms can be held liable for hosting illegal content under Article 14, provided that the platform had knowledge of the illegal activity and did not act to disable or remove it. Nevertheless, the e-Commerce Directive (ECD) did not establish any monitoring or control obligations for platforms to root out unlawful content. In fact, Article 15(1) of the ECD explicitly provides that Member States shall not impose a general obligation on providers, when providing the services covered (by Articles 12, 13, and 14), to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity[21].

ECJ case law demonstrates that Article 14 of the ECD was indeed used to hold platforms liable for the hosting of illegal content. A prominent example is Glawischnig-Piesczek v Facebook Ireland Ltd, in which the CJEU insisted that Facebook can be ordered to remove illegal/defamatory content posted by users, even if the users reside outside of the EU[22]. The case concerned Austrian politician Eva Glawischnig-Piesczek, who had solicited Facebook to remove a defamatory user comment about her, a request Facebook dismissed. The ECJ first held that Facebook's hosting services fell under Article 14 of the ECD. The court also held that article 15 of the ECD, which asserted no obligation for providers to monitor the content they host, does not preclude national courts from ordering them to take down content if it is unlawful or "equivalent"[23]. The case raised questions about platform liability. when it comes to user-generated content and

20. European Parliament and Council of the European Union, *Directive of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce)*, OJ L, 178, 1-16 (2000), available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32000L0031 (last visited April 6, 2023).

21. See *ibid*.

22. C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited*, ECLI:EU:C:2019:821..

23. The court described "equivalent content" as "information conveying a message the content of which remains essentially unchanged and therefore diverges very little from the content which gave rise to the finding of illegality".

the legitimacy of a national order triggering the removal of content globally[24].

In 2010, along with the ECD, the EU also adopted the Audiovisual Media Services Directive. The AVMSD, while regulating broadcasting, television, and radio, also provides rules for video-sharing platforms, such as YouTube, to protect users from harmful content[25]. Specifically, Article 28b requires platforms to protect the public from content whose dissemination is criminal in EU law, such as terrorism, child pornography, or offenses concerning racism or xenophobia[26]. Regulation 2021/784 on online terrorist content requires hosting services to remove any terrorist content within one hour of getting a "removal order" from a designated national authority[27]. This indicates that the platforms are not themselves required to search for terrorist content but must rapidly remove any such material when detected by competent authorities.

However, the most comprehensive act adopted by the EU regarding content moderation is the 2022 Digital Services Act, a regulation that modernized the rules governing online platforms under the e-Commerce Directive[28]. The DSA, which will be applied to all regulated entities later in 2024[29], intends to regulate the sharing of "illegal

24. Luc von Danwitz Danwitz, *The Contribution of EU Law to the Regulation of Online Speech*, 27 Michigan Technology Law Review, 167 (2020), available at https://www.congress.gov/bill/116th-congress/house-bill/133/text#toc-H6A24A7F9B-1B04FF2AEF09C41F028FC12 (last visited April 04, 2023).

25. European Parliament and Council of the European Union, *Directive on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive)*, OJ L 95/1 (March 10, 2010).

26. European Commission, *Communication from the Commission Guidelines on the practical application of the essential functionality criterion of the definition of a 'video-sharing platform service' under the Audiovisual Media Services Directive*, C/2020/4322 OJ C 223/3 (July 7, 2020).

27. European Parliament and Council of the European Union, *Regulation on addressing the dissemination of terrorist content online*, Regulation (EU) 2021/784 OJ L 172/79 (April 29, 2021).

28. European Parliament and Council of the European Union, *Regulation on a Single Market For Digital Services and amending Directive 2000/31/EC,* Regulation (EU) 2022/2065 L 277/1 (October 27, 2022).

29. Due to its status as a regulation, the DSA is self-executing and directly applicable to all EU member states. It was entered into force in November 2022 but its

content, online disinformation or other societal risks"[30]. Under the DSA, platforms will be required to implement stronger measures to prevent the dissemination of illegal content, such as hate speech, terrorist propaganda, and child abuse material, which were previously dealt with by specific instruments[31]. As per the DSA, online platforms will not be held liable for the content hosted if the platform does not have actual knowledge of the illegal activity or illegal content and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or illegal content is apparent; or, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the illegal content[32].

While the DSA has yet to fully apply, the aforementioned Article 6 of the DSA is identical to Article 14 of the previous e-Commerce Directive. Attempting to create a healthier and safer online environment for users, the DSA is a significant development in the regulation of the digital economy and while the impact it will have on online platforms, which operate within the EU, is yet to be seen, it is sure to be notable. To begin with, the DSA provides that intermediary services will not lose their liability exemption if they carry out voluntary initiatives aimed at investigating, detecting, or removing unlawful content in good faith and a diligent manner[33]. This is a guarantee to the platforms that, for as long as they comply with said standards and have their own practices for detecting unlawful content, they will not be subject to legal action or fines, as well as an incentive for them to demonstrate they are acting with due diligence and good faith to address these issues and maintain their liability exemption.

Under the DSA, platforms operating in the EU have to designate a point of contact for direct communication with authorities in the

---

full application will start in February 2024 (European Commission, Digital Services Act Package).

30. European Parliament and Council of the European Union, *Regulation on a Single Market For Digital Services* (cited in note 28).

31. See Caroline Cauffman and Catalina Goanta, *A new order: The Digital Services Act and consumer protection*, 12(4) European Journal of Risk Regulation, 758-774 (2021).

32. European Parliament and Council of the European Union, *Regulation on a Single Market For Digital Services* (cited in note 28).

33. Art. 7, *Regulation on a Single Market For Digital Services* (cited in note 28).

Member States to increase cooperation and transparency[34]. Discussing increasing transparency, Article 14 of the DSA provides that intermediary services must make their content moderation policies and procedures in their terms and conditions of use. This article is followed by Article 15, which obliges providers to release annual public reports on any content moderation they engaged in, including the number of national orders and complaints received, the number of notices submitted and processed, any content moderation conducted at their own initiative and any use of automated means of moderation. In addition, very large online platforms[35] must include the human resources dedicated to content moderation and the qualifications of the persons involved and indicators of the accuracy of automated means of moderation[36]. The DSA also aims to harmonize notice and action procedures, which the previous ECD did not do[37], by obliging hosting providers to put in place user-friendly mechanisms to allow "any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content."[38]. The providers must respond to these reports without delay and provide the reporting user with a statement explaining the grounds for their decision[39]. This is intended to create a more streamlined and transparent process for addressing unlawful content. The DSA also requires certain platforms to establish out-of-court dispute settlement bodies, which would help resolve disputes arising out of content moderation practices and enforce the terms and conditions[40]. Similarly, even though the ECD encouraged creating out-of-court mechanisms to solve disputes, it did not explicitly require platforms to establish such bodies, unlike the DSA. On that account, the DSA

34. Art. 11, *Regulation on a Single Market For Digital Services* (cited in note 28)..

35. Under Article 33, "very large online platform" applies to any platform that has a number of average monthly active recipients of the service in the Union equal to or higher than 45 million.

36. Art. 42, *Regulation on a Single Market For Digital Services* (cited in note 28)..

37. See Sebastian Felix Schwemer, *Digital Services Act: A reform of the e-Commerce Directive and much more*, prepared for A Savin, Research Handbook on EU Internet Law (2022), available at https://ssrn.com/abstract=4213014 or http://dx.doi.org/10.2139/ssrn.4213014 (last revised October 13, 2022).

38. Art. 16, *Regulation on a Single Market For Digital Services* (cited in note 28)..

39. *Id.* art. 17.

40. *Id.* art. 2 §1.

establishes formal requirements for content moderation, notice and action procedures, dispute settlements, and complaint procedures, as well as aims to enhance platform transparency when it comes to the restrictive measures employed.

The DSA recognizes the need to take into consideration fundamental freedoms stating in its preamble that the restrictions should not be arbitrary or discriminatory and that providers of very large online platforms should "pay due regard to freedom of expression and of information, including media freedom and pluralism.". It emphasizes that very large online platforms should be proportionate in their measures and avoid unnecessary restrictions on the use of their service, considering the potential negative effects on those fundamental rights. While the DSA does not specifically refer to balancing free speech with the public interest, its emphasis on fundamental freedoms and proportionality indicates a recognition of the need to balance these competing interests.

In addition to these regulations and directives, the EU has also taken measures to deal with disinformation and fake news, mainly through soft law instruments. The Code of Practice on Disinformation, adopted in 2018 and strengthened in 2022, is a voluntary framework for firms to fight disinformation[41]. This was adopted after the Facebook-Cambridge Analytica scandal, in which consulting firm Cambridge Analytica harvested unauthorized personal data from Facebook users in order to influence political outcomes. The scandal resulted in mass scrutiny regarding Facebook's data policy and the EU proposal for the Code specifically referred to it: "The recent Facebook/Cambridge Analytica revelations demonstrated exactly how personal data can be exploited in electoral contexts, and are a timely reminder that more is needed to secure resilient democratic processes."[42] The Code of Practice asserts that social media companies should enhance transparency regarding political advertisements, as well as calls for platforms to work with fact-checkers and to proactively remove fake accounts

41. European Commission, *The 2022 Code of Practice on Disinformation* (June 16, 2022), available at https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

42. European Commission, *Tackling online disinformation: Commission proposes an EU-wide Code of Practice*, Press release (Brussels April 26, 2018), available at http://europa.eu/rapid/press-release_IP-18-3370_en.htm.

used to spread disinformation[43]. While it is not a legally binding act, it has been signed by Google, Facebook, and Twitter among others. To fight disinformation, the EU has also launched the European Digital Observatory, a group consisting of fact-checkers and media literacy experts meant to analyze and understand disinformation trends on online platforms, identify practices to counter the spread of disinformation and work with policymakers[44]. The European Digital Observatory was proposed by the European Commission in its 2020 Democracy Action Plan, which set out to address the broader challenges facing democracy in the digital age[45].

Another soft law instrument regarding content moderation is the Code of Conduct on countering illegal hate speech online, drawn up in 2016. Signed by several companies like Facebook, TikTok, Twitter, and YouTube, the Code of Conduct is a commitment by IT companies to review any report of hate speech on their platform and remove or disable such content[46]. In its preambulatory clauses, the Code of Conduct also stresses the importance of protecting free expression, stating that the IT Companies and the European Commission also emphasize the need to defend the right to freedom of expression as well as that the spread of illegal hate speech online not only negatively affects the groups or individuals that it targets, but also those who speak out for freedom, tolerance, and non-discrimination in our open societies. This implies that, while the EU recognizes the importance of freedom of expression, hate speech comes at the expense of open and democratic discourse and therefore cannot be protected under the guise of free speech[47]. These instruments have played a crucial role in shaping content moderation practices within the EU. The Union

43. See European Commission, *The 2022 Code of Practice on Disinformation* (cited in note 41), Chapter III on political advertisements and Chapter VII on fact-checkers.

44. European Commission, *Communication* (cited in note 6).

45. See *id.*

46. European Commission, *Code of Conduct on Countering Illegal Hate Speech Online* (2016), available at https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

47. Similarly, the Code of Practice on Disinformation specifically refers to the need of finding a balance between free speech and freedom from harm, its preamble reading that the parties are mindful of the fundamental right to freedom of expression, freedom of information, and privacy, and of the delicate balance that must be

pushes platforms to take a proactive approach in removing harmful content to protect the public interest, which has led to most providers developing their own moderation policies to detect and remove all such content. While the EU attempts to balance free speech and protecting users and public interest, its comprehensive guidelines suggest that they prioritize defending users from harmful content in order to create a healthy online environment, as well as promote a culture of accountability and transparency in content moderation. However, this approach may fall short when it comes to stimulating innovation, as newer companies may be discouraged by the over-regulation, and social networks may begin to over-moderate, which means removing content that is not harmful, in order to avoid potential retribution[48].

3. *United States: a Liberal Approach to Content Moderation*

While the EU aims to actively regulate content moderation, the U.S. approach is more hands-free and noninterventionist, largely based on the First Amendment of the Constitution, which protects free speech from any Congress legislation[49]. It is important to note that the First Amendment only refers to acts that restrict free speech made by the State. This means that social media platforms, which are private actors, are allowed to restrict speech as they please because they are not bound by the First Amendment. That said, this is becoming more controversial, especially in relation to potential social media political bias. Content moderation policies are also further affected by Section 230 of the Communications Decency Act, passed in 1996, which protects online platforms as intermediaries that cannot be held liable for posts made by users[50]. In other words, Section 230 grants immunity to sites that host harmful content, even if the site has moderation policies of its own.

---

struck between protecting fundamental rights and taking effective action to limit the spread and impact of otherwise lawful content.

48. See Michal Lavi, *Do Platforms Kill?*, 43(2) Harvard Journal of Law & Public Policy, 477 (2020).

49. 1st Amendment, Constitution of the United States (1791).

50. Section 230, CDA. 47 U.S.C. § 230 (1996).

The federally enacted CDA allows for free expression online by protecting companies from unforeseeable legal problems, but this has been challenged through some state-level laws, which seek to hold platforms accountable for the content posted by their users. In 2021, Texas introduced a bill, which would allow some social media users to sue social media platforms if their posts get taken down, or if their accounts get deleted based on their political views[51]. This "censorship law" was quite controversial and was blocked by a federal judge in Texas through an injunction, as it was seen as violating the platform's First Amendment[52]. The case was later contested by the Court of Appeals for the Fifth Circuit, where the preliminary injunction was lifted, although it was subsequently reinstated by the Supreme Court until a further ruling by the Fifth Circuit, in which the judge denied the injunction arguing that platforms are not newspapers, and their censorship is not speech"[53]. Even though the impact of the law on social media platforms is uncertain, as there have been no actual cases on its application so far, its legality may still be questioned on the basis of it contradicting Section 230, which allows social media platforms to moderate content as they see fit. Similar legislation has come into effect in Florida, which passed a law that prohibits platforms from suspending or banning accounts of political candidates during an election[54]. It also allows Florida citizens to sue Big Tech if they are treated unfairly, although it does not provide a definition for what exactly constitutes unfair treatment[55]. A challenge to State level legislation arises in the balancing test established in Pike v. Bruce Church in 1970. The case involved an Arizona statute challenged as it placed an undue burden on interstate commerce, which is protected under the Commerce Clause of the Constitution. The Supreme Court held that State laws, which excessively burden out-of-state businesses or individuals, may be struck down as unconstitutional, thus establishing the "Pike balancing test"[56]. In the context of content moderation,

51. Texas House Bill 20, Tex. H.B. 20, 87th Leg., Reg. Sess. (. 2021).

52. Leslie Y. Garfield Tenzer and Hayley Margulis, *A 180 on Section 230: State Efforts to Erode Social Media Immunity*, 49 Pepp. L. Rev. (2022).

53. *Ibid*.

54. Florida Senate Bill 7072, Fla. Stat. § 106.115(2) (2021).

55. See *id.*

56. *Pike v. Bruce Church*, 397 U.S. 137 (1970).

this means that, if multiple states have their own specific laws on how platforms should moderate content, it could become an unreasonable burden for platforms to stay up-to-date and consistently apply multiple different standards of moderation.

The U.S. Supreme Court has often given precedence to the protection of speech when faced with cases related to content moderation. One of the first landmark cases which affected this area was Reno v. American Civil Liberties Union in 1997, in which the Supreme Court ruled that certain portions of the Communications Decency Act (CDA) are unconstitutional restrictions of free speech[57]. The CDA criminalized online speech that is classified as "indecent" and could be viewed by minors in an effort to protect children[58], but the court ruled that freedom of expression outweighs the benefits of such censorship on social media. The Court found the CDA's overly broad nature put an unconstitutional burden on adults and that protecting children from harmful materials does not justify an unnecessarily broad suppression of speech addressed to adults[59].

In 2015, the Supreme Court in Elonis v. United States held that threats made on social media need to be judged upon whether there was proof of *intent* to threaten rather than if the comment was reasonably perceived as a threat[60]. The case concerned threatening messages made by U.S. citizen Elonis on Facebook. When initially on trial, Elonis had argued that the State was required to prove an intent to communicate a "true threat" which was rejected by the district court that held the threshold at any communication that could reasonably be perceived as a threat[61]. When the case reached the Supreme Court, the debate surrounded whether the term "threat" included an intent to convey harm. The Court ruled that it does, and any lack thereof is a restriction on freedom of speech, ergo unconstitutional. The ruling upheld the importance of protected speech and clarified a higher standard for convicting individuals making threatening messages. This case was decided in the rapidly changing landscape of online communication and became a landmark case regarding online speech. In

57. *Reno v. American Civil Liberties Union*, 521 U.S. 844 (1997).
58. Communications Decency Act (CDA), 47 USC § 230 (1996).
59. *Reno*, 521 U.S. 844 (1997) (cited in note 57).
60. *Elonis v. United States*, 575 U.S. 723 (2015).
61. See *ibid*.

2017, the Court reinforced its commitment to protecting free speech in Packingham v. North Carolina, where it was ruled that a North Carolina statute barring sex offenders from using social media is unconstitutional and consists of a violation of free speech[62]. Specifically, the Court held that the First Amendment also includes online communication given its significance as a platform for public discourse and a source of information. The fact that courts have repeatedly ruled in favor of free speech, even when that speech is controversial or offensive, is evidence of how crucial this value is in American society and how embedded it is in its legal system.

Nonetheless, there has been another direction taken by the United States with regard to the protection of free speech, which focuses on the platforms themselves. While the Supreme Court in cases like Elonis and Peckingham has stressed the importance of safeguarding free speech in the digital era, lower courts have been defending the free speech rights of private platforms. For instance, in Prager Univ. v. Google, a California federal court ruled that YouTube did not violate Prager University's free speech right by restricting its prominently right-wing content, as YouTube is a private company[63]. Soon after, the Court of Appeals affirmed the decision and held that claims alleging censorship and denial of equal protection were meritless because the providers were not state actors[64]. Likewise, in Freedom Watch v. Google, et al., in which conservative activists claimed that multiple online platforms were violating their First Amendment rights by censoring their accounts, a Washington D.C. appeals court dismissed the case on the basis that private entities have no responsibility to respect free speech[65]. It is clear that Section 230 and the Constitution grant platforms broad discretion in regulating content, in addition to protecting them from liability for the content they host, but there is still pressure from users and lawmakers that prompt them to uphold certain policies.

Even though Congress is constitutionally prohibited from passing legislation that violates the First Amendment, and therefore cannot

62. *Packingham v. North Carolina*, 582 U.S. (2017).
63. *Prager University v. Google LLC*, U.S. Dist. (2018)
64. *Prager University v. Google LLC*, 951 F3d 991 (9th Cir 2020)
65. *Freedom Watch v Google LLC et al.*, 2018 WL 4738803 (D.D.C. Sept. 28, 2018).

act when it comes to restricting harmful speech, there have been some congressional hearings in which they investigated content moderation policies on social media platforms. In 2018, Mark Zuckerberg was called to testify in a joint Congressional hearing after the allegations that the company had allowed political consulting firm Cambridge Analytica to access millions of users' data without their consent, to target them with political ads[66]. During the hearing, he was also questioned about Facebook's content moderation policies, especially concerning the spread of fake news and hate speech on the platform, and acknowledged the importance of investing in moderation technology[67]. Additionally, two years later, in a Senate Committee of the Judiciary, Zuckerberg called for Congress to reform Section 230, in order to involve the government in privacy policies and to regulate the role of social media in elections[68]. In July 2019, a House Judiciary Committee hearing examined the influence of companies like Facebook, Google, and Twitter, focusing on how these companies moderate political speech[69].

Talks of censorship have been rapidly escalating as a result of political unrest within the country. In 2021, social media platforms including Twitter and Facebook banned President Trump in the wake of the January 6th Capitol riot[70]. These actions triggered more debate

66. See Edward Lee, *Moderating Content Moderation: Framework for Nonpartisanship in Online Governance*, 70 American University Law Review, 913 (2021).

67. U.S. Senate Committee on the Judiciary, *Facebook Social Media Privacy, and the Use and Abuse of Data*. 115th Cong., 2nd sess. Senate Hearing 115-683 (April 10, 2018), available at https://www.congress.gov/event/115th-congress/senate-event/LC64510/text?s=1&r=59 (last revised April 9, 2023).

68. U.S. Senate Committee on the Judiciary, *Breaking the news: Censorship, suppression, and the 2020 election* (November 17, 2020),available at https://www.judiciary.senate.gov/committee-activity/hearings/breaking-the-news-censorship-suppression-and-the-2020-election (last revised April 9, 2023).

69. U.S. House Committee on the Judiciary. Subcommittee on Antitrust, Commercial, and Administrative Law, *Online platforms and market power, part 1: the free and diverse press*, 116th Cong., 2nd sess. (June 11, 2019), available at https://www.congress.gov/event/116th-congress/house-event/109616.

70. See Facebook Newsroom, *In Response to Oversight Board, Trump Suspended for Two Years; Will Only Be Reinstated if Conditions Permit* (June 4, 2021), available at https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/ and Twitter, *Permanent suspension of @realDonaldTrump* (January 8, 2021), available at https://blog.twitter.com/en_us/topics/company/2020/

about the power of social networks in regulating political speech and where its limits should lie. In 2022, Elon Musk purchased Twitter and reinstated many right-wing accounts, including ones that had been banned for hate speech, after having criticized the site for suppressing and censoring conservative viewpoints, and for being politically biased[71]. These events were then discussed in the 2023 hearing on Government Interference and social media Bias, in which former Twitter personnel testified on Twitter's content moderation policies[72].

As outlined in the last paragraphs, the EU tends to focus more on terrorism, hate speech, and other 'harmful' material, while the U.S., especially post-2016 election, tends to center the conversation around political speech and platform transparency. Placing fundamental importance on free speech allows for social media to position themselves as public forums of important discussions adopting different perspectives and ideas. It also allows for more innovation, with newer platforms adopting their own moderation policies depending on the content they host. Overall, the US approach to content moderation prioritizes free speech over other concerns. This approach largely provides platforms with freedom and flexibility, but it also places a burden on content moderation, especially as the platforms grow, and it becomes more difficult to monitor every post.

4. *Key Issues in Content Moderation*

Regardless of which approach is more effective, there are still significant disadvantages to both methods of overseeing content moderation. This section will attempt to outline the key problems that platforms face in regulating online content in a globalized and

---

suspension#:~:text=After%20close%20review%20of%20recent,of%20further%20incitement%20of%20violence.

71. See The Guardian, *Elon Musk offers 'general amnesty' to suspended Twitter accounts* (November 24, 2022), available at https://www.theguardian.com/technology/2022/nov/24/elon-musk-offers-general-amnesty-to-suspended-twitter-accounts.

72. U.S. House House Committee on Oversight and Accountability, *Protecting Speech from Government Interference and Social Media Bias, Part 1: Twitter's Role in Suppressing the Biden Laptop Story*, 118th Cong., 1st sess. (February 8, 2023), available at https://www.congress.gov/event/118th-congress/house-event/115286?s=1&r=9.

interconnected world, as well as the challenges that arise from the lack of clear guidelines and the subjective nature of moderation policies across different platforms.

The EU itself has yet to narrow down exactly what constitutes "illegal content", with the only definition being "information which is not in compliance with EU or Member States Law"[73]. However, different member states have different practices and fragmented legislation becomes problematic for companies that already have to comply with an array of legal and regulatory standards. For instance, Germany gives social media platforms twenty-four hours to remove "obviously illegal" hate speech after being notified and seven days if its legal status is more problematic to determine through its Network Enforcement Act (NetzDG)[74]. This was echoed in the French 'Avia Law', which, however, was struck down by the French constitutional court holding that the deadline was too short, and the decision could pose an unnecessary or disproportionate risk to free expression[75]. Similar to the NetzDG, Austrian law provides platforms twenty-four hours upon notification to remove 'clearly' illegal content, but it also requires higher attention given to user rights and more sophisticated complaint management procedures[76]. This creates a situation where platforms may struggle to comply with different requirements across different countries. Furthermore, member states may have different standards concerning the substantive content of what is allowed to be shared. The German Network Enforcement Act imposes strict regulations on hate speech and specifically targets social media[77]. In contrast, although all EU member states have some level of hate speech regulation, countries like Poland and Hungary do not have specific laws regarding hate speech; instead, they include it in their respective criminal codes and may be more permissive on what kind of content

73. See Article 3(h), *Regulation on a Single Market For Digital Services* (cited in note 28).

74. Network Enforcement Act, NetzDG, Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 58 (2017).

75. Judit Bayer, *Procedural rights as safeguard for human rights in platform regulation*, Policy & Internet, 14 755-771 (2022)

76. See *id.*

77. See Rebecca Zipursky, *Nuts about Netz: The Network Enforcement Act and Freedom of Expression*, 42 Fordham International Law Journal, 1325-1368 (2019).

is allowed[78]. This might lead to inconsistent moderation, with some content being allowed to remain, while other similar content has to be removed, based on the country in which it is posted from.

Platforms may also decide to err on the side of caution to avoid being sanctioned, and begin to remove content that, in truth, is not harmful or "illegal". This may lead to over-removal of content, that does not violate policy but is controversial, leading to a chilling effect on free speech and freedom of expression[79]. In fact, the line between content moderation and censorship is becoming increasingly blurred and platforms are becoming no strangers to accusations of suppression or arbitrary content removal. In 2016, Facebook suspended editors and executives of two major Palestinian news publications, that covered daily news in the West Bank[80]. The editors claimed they had not violated community guidelines and were given no explanation for the suspensions. Facebook later reversed the decision claiming that it had been a mistake, although the journalists suspected it was a result of an agreement made by Facebook and the Israeli government to regulate content inciting violence[81]. The reference is to an informal agreement made by the two parties to crack down on incitements, preceded by dissatisfaction from the government and even a "Facebook bill" proposed by the Knesset, which would have granted broad authority to officials seeking court orders to compel Facebook

78. See Uladzislau Belavusau, *Hate Speech and Constitutional Democracy in Eastern Europe: Transitional and Militant? (Czech Republic, Hungary and Poland)*, 47 Israel Law Review 27 (2014).

79. See Amélie Heldt, *Borderline speech: caught in a free speech limbo?* (Leibniz Institute for Media Research, Hans-Bredow-Institut, Hamburg, Germany).

80. See Sophia Hyatt *Facebook 'blocks accounts' of Palestinian journalists*, (Al Jazeera, 2016), available at https://www.aljazeera.com/news/2016/9/25/facebook-blocks-accounts-of-palestinian-journalists.

81. This wasn't the only time Facebook was accused of political censorship. In 2016, a user uploaded a video following the aftermath of a police shooting in the U.S, which did not violate community standards, but was taken down regardless and later blamed on a glitch (see The Washington Post, *Why the Philando Castile police-shooting video disappeared from Facebook then came back*, 2016) Similarly, in 2017, Twitter suspended the account of Egyptian journalist Wael Abess who used his account to document situations of human rights abuse, without providing a reason for the suspension (see The Guardian, *Twitter under fire after suspending Egyptian journalist Wael Abbas*, 2018).

to remove content[82]. This agreement could have had unintended consequences, specifically resulting in censorship or over-restriction on pro-Palestinian speech. Although it cannot be said that every situation of censorship results from concern about regulatory penalties, it is clear that social media platforms have often had to deal with situations where the line between harmful and necessary content is unclear, and have penalized users that, though sharing controversial material, did not violate any guidelines.

The Digital Services Act applies to all platforms that offer services to EU citizens, even if the platform itself is based outside of the Union (which is the case with most major platforms including Facebook, Twitter, and YouTube)[83]. However, platforms also have to deal with contradictory legislation of other countries which mean to regulate content differently. For instance, Chinese law is highly strict on regulatory requirements for censorship, requirements, which may directly conflict with the DSA and their speech protection standards[84]. If platforms decide to comply with the DSA free speech laws by not censoring certain content, they may face penalties from China, which operates under a cyber sovereignty policy seeking to restrict foreign content[85]. A further potential problem is platforms that operate in the EU but are based in regions lacking effective cooperation mechanisms with the EU, suggesting that, while the DSA applies to them as well, it is more difficult to enforce it. Examples of this are social network sites operating from China or Russia, such as WeChat and VKontakte that are monitored by their governments[86]. This is to underline that legislative regulations can be very problematic for social media platforms, which in turn might have an easier time regulating content on

82. Sarah Koslov, *Incitement and the Geopolitical Influence of Facebook Content Moderation*, 4 Georgetown Law Technology Review, 183 (2019).

83. European Parliament and Council of the European Union, *Regulation on a Single Market For Digital Services* (cited in note 28).

84. National People's Congress of the People's Republic of China, *Cybersecurity Law of the People's Republic of China (2016)*, available at https://digichina.stanford.edu/work/translation-cybersecurity-law-of-the-peoples-republic-of-china-effective-june-1-2017/.

85. See *id.*

86. See Callum J. Harvey and Christopher L. Moore, *The client net state: Trajectories of state control over cyberspace*, *15 Policy & Internet 133* (2022), available at https://doi.org/10.1002/poi3.334.

their own guidelines, potentially even achieving more effective results. A 2018 research analysis concluded that the automated means of moderation used by platforms were more effective in identifying and removing hate speech than a group of human coders[87]. However, the scope of this article was limited to hate speech and more research is needed to fully examine the effectiveness of self-regulation.

Because most legal systems give significant discretion to platforms to decide their moderation policies[88], users and platforms often do not have clear guidelines regarding what is considered inappropriate or unacceptable behavior on the legal level. This can and does lead to discrepancies and confusion in moderation practices. Different platforms have different standards, and many of them have recently suffered accusations of bias and censorship. For instance, the U.S. takes a strong emphasis on protecting free speech, which may lead to platforms hesitating to remove controversial or harmful content for fear of being accused of censorship. YouTube has been criticized for not removing videos spreading conspiracy theories and proliferating misinformation through their algorithm[89]. On the other hand, there is the risk of over-censorship, where platforms may remove content that is not essentially harmful to avoid controversy. In 2021, YouTube was also accused of being too aggressive and of removing content that did not violate its policy, while trying to crack down on COVID and political misinformation[90]. This is where the idea of balancing the opposing interests comes into play. Platforms often have to make decisions on a case-by-case basis, to ensure that freedom of speech is being protected while removing harmful content. Whichever they choose can

87. Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber, *Automated Hate Speech Detection and the Problem of Offensive Language*, 1703 Cornell University (2017), available at https://doi.org/10.48550/arXiv.1703.04009.

88. The U.S. protects platforms through the First Amendment and Section 230, while the EU's Article 7 of the DSA allows platforms to take voluntary measures to strike down unlawful content.

89. See Mark Ledwich and Anna Zaitsev, *Algorithmic extremism: Examining YouTube's rabbit hole of radicalization*, 25 First Monday (2020), available at https://doi.org/10.5210/fm.v25i3.10419.

90. See Caroline Anders, *YouTube yanked public meeting videos over covid misinformation. Now it's backtracking* (The Washington Post, August 7, 2021), available at https://www.washingtonpost.com/technology/2021/08/07/youtube-covid-misinformation-city-council/.

lead to criticism because there is no 'perfect' solution. They are left to choose between human-based or automated methods of moderation or some degree of combination between the two. Automated moderation refers to algorithms and machine technologies being trained to filter harmful material and remove it upon detection. However, while this might be more efficient, algorithmic machines are designed to reflect society and can often exhibit bias by promoting existing societal stereotypes[91]. An example of algorithmic bias is when, in 2018, Amazon came under fire for using recruiting machine technology that penalized job applications including words like "women" and "female", which led to fewer women qualifying for the later stages of the application process[92]. Another concern is the issue of over-removal. AI cannot make contextual decisions when it is unclear if a post is violating a rule[93]. For instance, in situations of satirical content, it is difficult for AI to recognize that the post is not violating community standards. On the other hand, using automated means of moderation can be a faster and more efficient way of removing the most harmful content, as well as loosening the burden on human moderators, who are exposed to disturbing content and can face long-term emotional and psychological effects[94]. Additionally, firms with fewer resources cannot afford to pay human moderators and AI becomes the more suitable path for this job. Ultimately, while automated content moderation has its drawbacks, platforms can benefit from it for as long as they have some level of human oversight to ensure impartiality (similar to Facebook's Oversight Board).

---

91. See Céline Castets-Renard, *Algorithmic content moderation on social media in EU law: Illusion of perfect enforcement*, University of Illinois Journal of Law, Technology & Policy 283 (2020).

92. See Colin Clemente JJones, *Systematizing Discrimination: AI Vendors & Title VII Enforcement*, 171 University of Pennsylvania Law Review, 235 (2022).

93. See Robert Gorwa, Reuben Binns, and Christian Katzenbach, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance,* Big Data and Society (2020).

94. See Miriah Steiger, Timis Bharucha, Sukrit Venkatagiri, Martin J. Riedl and Matthew Lease, *The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support,* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (2021).

Most major platforms demonstrate similar rules or community guidelines regarding how they moderate their content. Facebook's Community Standards cover six categories of unacceptable content along with rationales for each policy, with Twitter and YouTube using approximately the same principles[95]. However, platforms also have internal and more exhaustive rules that moderators use to make decisions, often not accessible to the public[96]. Social media companies have faced criticism for not being transparent in their decision-making processes and their moderation policies and users have called to increase trust by making this information public[97]. Moreover, platforms have begun to use third-party fact-checkers to look for disinformation, a practice that, while useful for identifying misinformation, has been criticized because these organizations can be partisan and exhibit bias in the content they choose to flag as inaccurate[98]. Increased transparency about moderation policies and employed means (algorithms, fact-checkers, etc.) should then be used by platforms if only to build trust with their user base.

A wider platform discretion model raises another important issue to be considered. As Kyle Langvardt points out in his article, "Regulating Online Content Moderation", the largest social platforms are owned by few corporations, leaving the moderation of online speech to become the responsibility of a small number of oligarchs[99]. This means that where there are no regulatory limitations, moderation becomes influenced by market, public, and government pressures[100]. Therefore, there is a risk that moderation practices may not align with the interests of the public and may even go against the users' rights to free expression. It could also lead to dominant platforms having the ability to shape all public discourse by suppressing oppositional viewpoints. Once again, it is apparent that mechanisms enforcing platform

95. See Karanjot Gill, *Regulating Platforms' Invisible Hand: Content Moderation Policies and Progress*, 21(2) Wake Forest J. Bus. & Intell. Prop. L. 171 (2022).

96. See *id.*

97. See Evelyn Douek, *Governing online speech: From "posts-as-trumps" to proportionality and probability*, 121(3) Columbia Law Review, 759 (2021).

98. See Petter Bae Brandtzaeg and Asbjørn Følstad, *Trust and distrust in online fact-checking services*, 60(9) Communications of the ACM, 65 (2017).

99. See Kyle Langvardt, *Regulating Online Content Moderation*, 106(5) Georgetown Law Journal, 1353(2018).

100. See *id.*

transparency are of key importance to increasing public accountability. Social media may also be used by governments themselves to incite violent movements, as was the case with the Rohingya genocide in Myanmar. In fact, over one hundred Facebook accounts were used to spread hate speech against the Rohingya Muslims, some of which enjoyed over a million followers and massive engagement[101]. These posts were written entirely in Burmese, but, in 2017, when the genocide was at its peak, Facebook only had five Burmese-speaking content moderators[102]. This added to the fact that Myanmar is composed of different languages and dialects, resulted in a large amount of content being left up even if it visibly violated Facebook's guidelines. The platform belatedly began to act against these accounts in 2018, after facing negative media reactions[103]. By 2018, over 10,000 Rohingya Muslims were killed in the genocide and over 700,000 had been displaced[104].

The lack of clear standards and guidelines for content moderation, both by the state and by the platform itself, can also contribute to political polarization and extremism, as users may feel that their speech is being treated unfairly and, because of flawed algorithms, it will not be exposed to opposing viewpoints[105]. In the U.S., some critics have pushed for legislation that mandates accountability and transparency of the social networks in their moderation policies, instead of relying on Section 230 as a shield[106]. Due to the lack of regulatory pressure for

---

101. Richard Ashby Wilson and Molly K. Land, *Hate speech on social media: Content moderation in context*, 52 Conn. L. Rev, 1029-1076 (2021).

102. See Rebecca J. Hamilton, *Platform-Enabled Crimes: Pluralizing Accountability When Social Media Companies Enable Perpetrators to Commit Atrocities*, 47 Yale Journal of International Law, 121 (2022).

103. See *id.*

104. United Nations Human Rights Council, *Report of the Independent International Fact-Finding Mission on Myanmar* (Aug 27, 2018), available at https://digitallibrary.un.org/record/1643079?ln=en.

105. See Pablo Barberá, *Social media, echo chambers, and political polarization*, (Cambridge University Press 2020).

106. For instance, Senator Blumenthal has called for §230 reform because it is used "to defend keeping the bad stuff there" (Press Release at https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-on-big-techs-legal-immunities-reform-is-coming). Additionally, Senator Josh Hawley has stated that §230 has been used to "shield the Nation's largest and most powerful technology corporations from any legal consequences" (Press Release at https://www.hawley.senate.gov/hawley-files-gonzalez-v-google-amicus-brief-supreme-court-challenging-big-techs-section-230).

transparency of platforms' policies, it is difficult for users to know the point at which something is considered unacceptable, and incomprehensible to punish them by citing policies they weren't told of. As previously mentioned, platforms are not obligated to protect free speech and, therefore, are able to make arbitrary decisions, even if it results in negative feedback.

5. *Looking Ahead: the Future of Content Moderation*

While the EU places a greater emphasis on regulating harmful content and the U.S. supports the protection of speech, both attitudes seem to have their shortcomings, which has made content moderation a challenging issue for social media platforms. Over-removal, under-removal, and biases are all issues that might become more prevalent if the current approach remains, especially as social media platforms continue to grow. The consequences of this can be grave, especially in cases of terrorist content or hate speech, as was seen in the aforementioned Rohingya incident in 2017. Human-based moderation has its difficulties too. Humans are also prone to bias and error, albeit they are also able to apply contextual knowledge to their evaluation. Moreover, the amount of content that is generated is too great for such moderation to be scalable. Accommodating space for harmful content to subsist poses a threat to users and society and there should be some level of safeguarding to make sure this does not occur.

Many recommendations have been made to address these challenges. Increasing platform transparency is of utmost importance to ensure a safer online environment. Platforms should disclose what their exact moderation policies are, as well as the decision-making process, when necessary. Castets-Renard suggests that the EU set more stringent rules, requiring moderators to inform users why their content was removed on a case-by-case basis upon request[107]. Additionally, platforms need to be explicit when defining "harmful" content, since vagueness increases confusion and leads to distrust

107. See Céline Castets-Renard, *Algorithmic content moderation on social media in EU law: illusion of perfect enforcement*, 2 University of Illinois Journal of Law, Technology & Policy 283 (2020).

from users and the rest of the public. Governments may also enact legislation mandating due process, including appeal or counterclaim procedures where users can contest a decision made by the platform and have it revisited by the moderation team[108]. Additionally, they may provide training and support for platforms, to ensure that their moderators have the necessary knowledge regarding how to identify harmful content. Governments can also share information and data with platforms to aid them in identifying such content, particularly in areas like terrorism and disinformation.

While Section 230 of the Communications Decency Act currently protects platforms as intermediaries, instead of as publishers of the content they host, there have been many discussions on potential reforms. Most of these suggestions consist in limiting the scope of §230 to address challenges like cyber stalking or nonconsensual sexual content[109]. Platforms should also not be able to use §230 to invoke immunity for harmful content that they knowingly solicited or actively disregarded. Another suggestion is for the U.S. government itself to enact legislation requiring large social media sites based in the U.S. to establish independent oversight bodies (similar to Facebook's Oversight Board), to supervise and be responsible for upholding or reversing decisions that have been appealed[110]. This approach would maintain the country's commitment to free speech while also ensuring that social media platforms are accountable for their moderation practices and are more transparent in their decision-making processes.

Platforms can likewise act to improve the state of content moderation, for instance, by investing in improved AI and other algorithmic means which can detect harmful content proactively and remove it. AI is not errorless, however, it can be trained to identify harmful content and then supervised through audits or reviews to make sure that there

108. See Karanjot Gill, *Regulation platforms' invisible hand: content moderation policies and process*, 21(2) Wake Forest Journal of Business and Intellectual Property Law 171 (2022).

109. See Andrew P. Bolson, *Flawed but fixable: Section 230 of the Communications Decency Act at 20*, 50 Washington University Journal of Law & Policy, 97(2016).

110. See Trent Scheurman, *Comparing social media content regulation in the US and the EU: How the US can move forward with Section 230 to bolster social media users' freedom of expression*, 23 San Diego International Law Journal 413 (2022).

haven't been any missteps[111]. This is also when an oversight committee would be of use, as they could assess decisions made by algorithmic means, without having to be subject to so much disturbing content that is made public. Analogously, they could invest in training courses for their human moderators that ensure partiality and partisanship. Specifically, platforms should invest in moderators, who know less commonly spoken languages as these posts may go unnoticed due to the lack of moderators that can understand them. AI also falls short when referring to moderating content that requires context-specific information, like political or social situations. Ideally, there would be a 'mixed' system of both human moderators and automation to ensure accuracy[112].

Platforms can also encourage user participation by allowing them to flag or report harmful posts that are then reviewed, increasing the speed and efficacy of content moderation. They can work with specific organizations or individuals who are knowledgeable about moderation practices and have a strong understanding of the context behind posts being made, giving them the role of "trusted partner" and enabling them to monitor and flag problematic content[113]. A further recommendation, made by Evelyn Douek, is an approach focusing on proportionality and probability[114]. The suggestion is that moderating content should be done by weighing the harms and benefits of speech on a broader scale and a systemic basis, rather than looking solely at the individual post as an isolated event[115]. This would entail considering the context and potential implications of each post and using that to decide whether the potential harms outweigh the benefits of protecting speech. In other words, the decision should be made based on

111. See Yifat Nahmias and Maayan Perel, *The oversight of content moderation by AI: Impact assessments and their limitations* 58(1) Harv. J. on Legis. 145 (2021).

112. See Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, *Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts*, 31(1) Information & Communications Technology Law 123 (2022).

113. See Richard A. Wilson and Molly K. Land, *Hate speech on social media: Content moderation in context*, 52 Connecticut Law Review 1029 (2021).

114. See Evelyn Douek, *Governing online speech: From "posts-as-trumps" to proportionality and probability,* 121(3) Columbia Law Review 759(2021).

115. See *ibid*.

the likelihood that the post will cause harm, and how great that harm may be, rather than on the content of the post itself.

Overall, the goal of content moderation should be to find a balance between protecting society while also upholding the principles of free speech to promote a healthy online community. However, this can only be achieved through collaboration and cooperation between the public, the platforms, and the government.

6. *Conclusion*

The differences in the EU and U.S. approaches reflect the ones in values, caused by their unique historical and political backgrounds. The EU is more active in regulating harmful content, having passed the comprehensive Digital Services Act (DSA) governing online platforms, which aims to regulate the sharing of illegal content, online disinformation, or other societal risks. Along with numerous soft law instruments, the DSA has shaped the way content moderation is conducted within the EU and has fostered a culture of giving precedence to the safety of users, instead of enhancing free speech. While the DSA holds platforms liable if they do not remove harmful content that they are aware of, the U.S. grants them liability under the framework of Section 230 of the Communications Decency Act. The U.S. approach is more hands-free and places fundamental importance on free speech. The U.S. also highlights the fact that platforms are not bound to the 1st Amendment, which protects free speech, and have a certain level of sovereignty when deciding their moderation practices. However, this approach has been criticized, especially considering media monopolies and political censorship[116].

Issues stem from both the regulatory and the more liberal American model. The EU's regulatory approach poses problems due to fragmented legislation between the member states, leading to inconsistent moderation policies[117]. Furthermore, it upsurges a risk of

116. See Jonathan A. Obar and Anne Oeldorf-Hirsch, *The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services*, 23(1) Information, Communication & Society, 128 (2020).

117. See Céline Castets-Renard, *Algorithmic content moderation on social media in EU law: illusion of perfect enforcement*, 2 University of Illinois Journal of Law,

over-removal as platforms may begin to censor content that is not harmful solely to avoid potential fines[118]. The regulations established by the EU may also contradict those made by other countries, such as China, making it difficult for consistent moderation due to contradictory regulations. On the other hand, the U.S. model is characterized by unclear standards and guidelines, leading to confusion for both platforms and users. Additionally, social media platforms may result in under-removal to avoid accusations of censorship or biases. In the absence of clear regulations, platforms decide on their moderation policies by themselves, which often leaves users in the dark, due to a lack of platform transparency on their practices and methods[119]. The responsibility of content moderation falls on a small number of corporations, which presents an issue of potential monopolization[120].

To address these challenges, many reform proposals have been presented by scholars and policymakers. An increase in platform transparency is vital for a healthier online environment, as well as providing processes that allow users to appeal to or question moderation practices. This would ensure social media platforms' accountability for their moderation practices and transparency in their decision-making processes. Larger platforms should also invest in improved AI and in moderators' training to ensure effectiveness. In the U.S., protection of speech can still be ensured with legislation mandating platform Oversight Boards that monitor moderation practices.

Ultimately, the issue of content moderation is sensitive and is only gaining more significance in contemporary society. Social media networks have become forums and mediums of important conversation, and the responsibility to regulate it is too great for platforms to be left to deal with it alone. Collaboration between society, platforms, and governments is crucial for adopting a healthier online environment.

Technology & Policy 283(2020).

118. See Amélie Heldt, *Borderline speech: caught in a free speech limbo? Leibniz Institute for Media Research*, Hans-Bredow-Institut, Hamburg, Germany (2020).

119. See Edward Lee, *Moderating Content Moderation: Framework for Nonpartisanship in Online Governance*, 70(3) American University Law Review 913 (2021).

120. See Kyle Langvardt, *Regulating Online Content Moderation,* 106(5) Georgetown Law Journal 1353 (2018).