NYU | LAW

Journal of Intellectual Property & Entertainment Law

VOLUME 15 NUMBER 2

Statement of Purpose

Consistent with its unique development, the New York University Journal of Intellectual Property & Entertainment Law (JIPEL) is a nonpartisan periodical specializing in the analysis of timely and cutting-edge topics in the world of intellectual property and entertainment law. As NYU's first online-only journal, JIPEL also provides an opportunity for discourse through comments from all of its readers. There are no subscriptions or subscription fees; in keeping with the open-access and free discourse goals of the students responsible for JIPEL's existence, the content is available for free to anyone interested in intellectual property and entertainment law.

The New York University Journal of Intellectual Property & Entertainment Law is published up to three times per year at the New York University School of Law, 139 MacDougal Street, New York, New York, 10012. In keeping with the Journal's open access and free discourse goals, subscriptions are free of charge and can be accessed via www.jipel.law.nyu.edu. Inquiries may be made via telephone (212-998-6101) or e-mail (submissions.jipel@gmail.com).

The Journal invites authors to submit pieces for publication consideration. Footnotes and citations should follow the rules set forth in the latest edition of The Bluebook: A Uniform System of Citation. All pieces submitted become the property of the Journal. We review submissions through Scholastica (scholasticahq.com) and through e-mail (submissions.jipel@gmail.com).

The Journal is also available on WESTLAW, LEXIS-NEXIS and HeinOnline.

NEW YORK UNIVERSITY

JOURNAL OF INTELLECTUAL PROPERTY AND ENTERTAINMENT LAW

VOLUME 15 SPRING 2026 NUMBER 2

SOCIAL MEDIA PLATFORM REGULATION IN THE US AND THE EU: TOWARDS A DIVIDED INTERNET?

IOANNA TOURKOCHORITI*

This article analyses the systems of legal regulation of online social media platforms as regards to hate speech and misinformation in the US and the EU. It engages with the constitutional rights doctrines that allow platform regulation in Europe. It explains how the relationships between the government, private actors, the owners of social media platforms and the users of the platforms put to the test the constitutional doctrines that exist in the United States. It analyses the sophisticated system created by the Digital Services Act (DSA) in the EU and compares it with the regulation that exists in the US. The DSA enhances users' rights and imposes important transparency requirements on platforms, while strengthening their obligations to address hate speech and misinformation. It engages with Moody v. NetChoice*, the latest decision by the American Supreme Court on the topic and the regulatory options it allows. It discusses the differences in social media regulation in the US and the EU and explores whether they are likely to lead to a division of the internet. It examines the state of the art technology the platforms are using to detect hate speech and misinformation and evaluates the legal responses in the US and the EU in this respect. It also makes suggestions for further research that is needed to address hate speech, incitement to hatred and misinformation online.*

*Associate Professor of Law, University of Baltimore School of Law. The author would like to thank Gianmaria Ajani, Peter Danchin, Mark Graber, Eric Heinze, Fernanda Nicola, Martha Larson, Mortimer Sellers, Nathalie A. Smuha, Katherine Strandburg, James Whitman, Peter Yu, the participants and the audience in the Panel *AI And Digital Regulation in the United States, the EU, and China* held during the Annual Meeting of the American society of Comparative Law in 2024 at Texas A & M University School of Law, the Participants in the 9th Annual Texas A&M IP Scholars Roundtable held in November 2025, and the participants in the Information Law Institute research group seminar series at NYU in the spring of 2025 for suggestions and comments. Dawn Chukwurah, Emma Reid, Inga Brostek, and Savannah Long provided valuable research assistance.

Introduction

Online communication has created new challenges for contemporary democracies. These challenges are accentuated by the transnational nature of the internet and the diverging attitudes of states around the world in regulating speech. The US offers a more protective legal regime of free speech compared at least to European States. The comparison with Europe has become imperative recently since the enactment by the European Union in 2022 of the Digital Services Act. The Act creates a sophisticated system aiming to regulate online speech. On the other hand, several social media platforms have scaled back their content moderation practices in the US. X (formerly Twitter) drastically cut down its content moderation team following a change in ownership in 2022 and Meta's CEO announced in January 2025 that it is moving towards a model of content moderation based mostly on community oversight.[1] The divide between the EU and the United States on social media regulation raises the question of whether this is leading to a division of the internet.

Following the attack on the Capitol on January 6th, 2021, and the actions several social media platforms took to indefinitely block major political actors,[2] governments realized that it was important to regulate the ability of platforms to limit who has access to them. In the US, Texas and Florida, two states that enacted regulation of social media platforms, were concerned that the platforms are limiting too much conservative speech.[3] The EU officials, however, were concerned that

[1] Marianna Spring, *Charities' Dismay as Twitter Disbands Safety Group*, BBC (Dec. 13, 2022), https://www.bbc.com/news/technology-63907708 [https://perma.cc/XJ9R-XGAE]; Clare Duffy, *Meta is Getting Rid of Fact Checkers. Zuckerberg Acknowledged More Harmful Content Will Appear on the Platforms Now*, CNN (Jan. 7, 2025), https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation [https://perma.cc/MT62-WJL9].

[2] Bobby Allyn, *Facebook Bans President Trump From Posting For The Rest Of His Presidency*, NPR (Jan. 7, 2021), https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/07/954453630/facebook-bans-president-trump-from-posting-for-the-rest-of-his-presidency [https://perma.cc/8BBR-W7W7]. The ban was converted to a two-year suspension following a decision by Facebook's Oversight Board. *See* Shannon Bond, *Trump Suspended From Facebook For 2 Years*, NPR (June 4, 2021), https://www.npr.org/2021/06/04/1003284948/trump-suspended-from-facebook-for-2-years [https://perma.cc/44SK-9TZT]. The Oversight Board found the indefinite ban to be contrary to International Human Rights Standards and asked Facebook to re-examine the case and impose a proportionate penalty within six months.

[3] *See* Senator Ted Cruz, *Sen. Cruz Questions Mark Zuckerberg on Alleged Political Bias at Facebook*, (YouTube, Apr. 10, 2018), https://www.youtube.com/watch?v=-VJeD3zbZZI&t=73s [https://perma.cc/6QAT-WFTH]; Clay Calvert, *Anti-censorship Rhetoric v. First Amendment Realities: The Fight Over*

social media platforms as private actors may be limiting too much speech, much more than what is acceptable to limit in Europe, where limits to hate speech by the government are constitutionally permissible.[4] Texas and Florida legislated to affirm the First Amendment principles of protecting everyone's speech without discrimination, even hate speech.[5] The Europeans enacted legislation to strengthen the processes the platforms already have to limit hate speech and misinformation, and to provide to all users whose speech is limited avenues to complain if they believe that the platforms are limiting their speech unnecessarily.[6] They also passed legislation to protect the public from the spread of misinformation, disinformation, and manipulation.[7] The widespread misinformation during the pandemic and the perception that the platforms' responses to this problem were inadequate led EU officials to promulgate legislation that creates obligations for the platforms to address the relevant risks.[8]

State attempts to regulate the platforms' handling of hate speech in the US are failing. New York[9] and California[10] enacted legislation against hate speech on social media platforms. Two Circuit Courts have already found that the clauses are

Florida's Anti-deplatforming Statute and Some Thoughts About Speaker Autonomy, Compelled Expression and Access Mandates in Online Fora, 20 First Amend. L. Rev. 385, 393–94 (2022); Dawn Carla Nunziato, *The Old and the New Governors: Efforts to Regulate and to Influence Platform Content Moderation*, 22 First Amend. L. Rev. 348 (2024).

[4] *See* Eur. Comm'n, *Impact Assessment Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, SWD(2020) 348 final, pt. 1/2, at 19–20, Eur. Comm'n (Dec. 15, 2020), https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act [https://perma.cc/SEX5-V6UU].

[5] *See infra* notes 246, 271 and accompanying text.

[6] *See* Eur. Comm'n, *supra* note 4, at 26–27.

[7] Eur. Comm'n, *The Digital Services Act: Ensuring a Safe and Accountable Online Environment*, Eur. Comm'n, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act [https://perma.cc/HTK8-5WQF] (last visited July 21, 2025).

[8] *See* Eur. Comm'n, *supra* note 4, at 20.

[9] N.Y. Gen. Bus. Law § 394-ccc (McKinney 2024); *see* Volokh v. James, 656 F. Supp. 3d 431, 436 (S.D.N.Y. 2023) (granting preliminary injunction).

[10] Cal. Bus. & Prof. Code § 22677 (West 2025); X Corp. v. Bonta, No. 23-CV-1939, 2023 WL 8948286 (E.D. Cal. Dec. 28, 2023) (denying preliminary injunction), *rev'd and remanded*, 116 F.4th 888 (9th Cir. 2024). *See also* Cat Zakrzewski, *New California Law Likely to Set Off Fight Over Social Media Moderation*, Wash. Post (Sept. 14, 2022), https://www.washingtonpost.com/technology/2022/09/13/california-social-network-transparency/ [https://perma.cc/F9MV-6CRC].

content based and thus incompatible with the First Amendment.[11] Both courts are citing *Moody v. NetChoice*,[12] the US Supreme Court's latest decision in relation to social media platforms, which affirms their First Amendment right to curate speech.

There are to date few comparative studies of the DSA and the legal regime that exists in the US.[13] This article complements these studies by analysing the theoretical issues that underly the differences in regulation. The article explains how the relationships between the government, private actors, the owners of social media platforms and the users of the platforms put to the test the constitutional doctrines that exist in the United States. It engages with *Moody v. NetChoice* and explores possible avenues for regulation of social media in the US complementing the relevant scholarship. It discusses the differences in social media regulation in the US and the EU and explores whether they are likely to lead to a division of the internet. These differences matter, because social media platforms have emerged as extremely important and powerful social actors that define the public sphere and the quality of democracies world-wide.[14] It engages with the perspective of business and human rights and argues about the need for social media platforms to continue to self-regulate based on the standards that relate to the Code of Conduct they have agreed with the EU to abide by, even where they have no legal obligation to do so. It examines the state of the art technology the platforms are using to detect hate speech and misinformation and discusses which legal regime is more appropriate to address hate speech and misinformation risks. It also makes suggestions for further research that is needed to address hate speech and misinformation.

The divergence in the regulation of speech can be traced back to the foundation of democracies in Europe and the United States.[15] The American

[11] X Corp. v. Bonta, 116 F.4th 888 (9th Cir. 2024); Volokh v. James, 2025 U.S. App. LEXIS 19405 (2d Cir. 2025).

[12] Moody v. NetChoice, LLC, 144 S. Ct. 2383 (2024).

[13] Neil Netanel, *Applying Militant Democracy to Defend Against Social Media Harms*, 45 CARDOZO L. REV. 489 (2023); Dawn Carla Nunziato, *The Digital Services Act and the Brussels Effect on Platform Content Moderation*, 24 CHI. J. INT'L L. 115 (2023); Ioanna Tourkochoriti, *The Digital Services Act and the EU as the Global Regulator of the Internet*, 24 CHI. J. INT'L L. 129 (2023); Tiana Wang, *"Delicate Task": Content Moderation and Intermediary Liability in a Post-DSA World*, 39 BERKELEY TECH. L.J. 1507 (2024).

[14] *See* Eur. Comm'n, *supra* note 4, at 18–19.

[15] *See* IOANNA TOURKOCHORITI, FREEDOM OF EXPRESSION: THE REVOLUTIONARY ROOTS OF AMERICAN AND FRENCH LEGAL THOUGHT (2022).

constitutional system is characterised by distrust towards the government, whereas the French constitutional tradition has been characterised as a tradition of trust towards the government to help the citizens realise their liberties.[16] Where speech regulation is sparse, there is always the concern that political leaders and civil society may use their expressive freedoms in ways which can lead to undermining basic assumptions about facts and thus even legitimise dealing blows to constitutional democracy. Can too much freedom lead to unfreedom?[17] Does the danger of coming across too much "false" information run the risk to undermine democracy? Are democracies today facing the dangers that Isaiah Berlin was warning against in his famous critique of John Stuart Mill's argument?[18] Or does the state of the art technology at the service of speech regulation counsel a more cautious approach?

In the area of expressive freedoms, the US offers an extended area of protection of freedom of speech. In many European states, it is legitimate for the government to limit abuse of the same freedom to protect other citizens from the harm caused by hate speech.[19] It is also legitimate for European states to wield this power to prevent misinformation.[20] This regime is opposed to the absence of limits against hate speech, misinformation, and the lenient regime of incitement to violence that exists in the US. The emergence of online platforms of communication means that hate speech and incitement to violence can now escape the borders of the US. Several scholars have argued that online communication raises new challenges and that the legal regime in the US is no longer appropriate to address these challenges.[21] In their opinion, the risks to freedom and democracy that unregulated social media platforms entail necessitate the reference to the principles of militant democracy to justify regulating social media platforms.[22] The EU regulation in the DSA is inspired by a precautionary approach to the relevant risks. As analysed below, it requires social media platforms to elaborate important risk protocols to address situations of crisis when misinformation is

[16] *Id.* at 34.
[17] I am grateful to James Whitman for suggesting the importance of this question.
[18] Isaiah Berlin, *John Stuart Mill and the Ends of Life*, *in* Four Essays on Liberty 233 (2002).
[19] *See* Ioanna Tourkochoriti, Freedom of Expression, *supra* note 15, at 2–3.
[20] *Id.*
[21] *See infra* note 38 and accompanying text.
[22] *See* Neil Netanel, *supra* note 13.

likely to occur.[23] The first report that has been issued in application of the DSA by the EU indicates that the Act has given the opportunity to social media platforms to reflect on the possible risks of their practices in the area of misinformation and hate speech and to come up with ways to address them.[24]

Given the social power that social media platforms wield, the DSA serves as an important tool for addressing the power imbalances that exist between platforms and their users. Any attempt by platforms to moderate speech may escalate to removing users entirely. The DSA approach consists of multiplying the legal avenues available to users for challenging such exclusion and imposing specific timeframes for platform responses. The current regulatory approach in the United States does not seem to recognize that protecting the procedural rights of users removed from social media platforms is justified under the First Amendment. This means that the DSA protects users' rights more robustly than the US framework. The reluctance to protect these procedural rights means that the US system relies on excluded users finding alternative platforms to express themselves. EU law accepts the legitimacy of government involvement in regulating how civil society actors—including private platforms—affect users' speech rights. In contrast, the US approach emphasizes platforms' own speech rights and relies on market competition to provide alternative avenues for those excluded by one platform.

The DSA also obligates platforms to communicate how they use prioritization algorithms to deliver content to their users.[25] In this respect, the first report indicates that platforms have been compelled to consider how they can design prioritization algorithms that focus on content quality rather than engagement metrics alone.[26] In the United States, the use of prioritization algorithms by platforms is protected by their own free speech rights, which means that regulation in this area would likely not withstand First Amendment scrutiny.[27] The DSA has

[23] *See infra* Part II.A.2.

[24] Eur. Bd. For Digit. Services, *First Report in Cooperation with the Commission Pursuant to Article 35(2) DSA on the Most Prominent and Recurrent Systemic Risks as Well as Mitigation Measures*, Eur. Comm'n (Nov. 18, 2025), https://digital-strategy.ec.europa.eu/en/news/press-statement-european-board-digital-services-following-its-16th-meeting [https://perma.cc/535P-HPQ5].

[25] 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), art. 14 § 1, art. 35 § 1(d), art. 40 § 3, 2022 O.J. (L 277).

[26] Eur. Bd. for Digit. Services, *supra* note 24, at 36.

[27] Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2404–05 (2024).

stimulated a public conversation in Europe between social media platforms and civil society organizations that appears to be driving platforms to improve the design of their prioritization algorithms.

Until recently, the platforms had generalized their operation to abide by the strictest legal regime globally. In 2016, they signed with the EU a code of conduct, on the basis of which they took on the obligation to regulate hate speech and to engage in content moderation more broadly.[28] The code had a global effect, in that the platforms generalised their operations to abide by the code even where they had no legal obligation to do so, e.g., in the US.[29] In other words, the sparse government regulation of speech in the US led to a legal gap, which was filled by civil society actors. Facebook created the Oversight Board, a deontology committee which evaluates whether their content moderation practices are compatible with International Human Rights Law.[30] As analysed in this article, the policy changes announced by Meta in January 2025 will affect the content moderation practices and standards the platform applies around the world and will likely lead to a divergence between the US and the EU in handling misinformation, hate speech and incitement to hatred.[31]

In what follows, the article in Part I discusses how social media platform regulation puts standard constitutional doctrines in the EU and the US to the test. It engages with the constitutional doctrine of horizontal effect dominant in Europe, which legitimizes platform regulation, and the state action doctrine dominant in the US, which complicates this issue. It explores avenues of regulation in the US. In Part II, it then analyses the sophisticated legal regime that the DSA creates in Europe. The legal regime enhances users' rights and imposes important transparency requirements on platforms, while strengthening their obligations to limit hate speech and misinformation. It discusses how the DSA enhances the platform regulation that exists in some EU states. Part III examines attempts to regulate social media platforms across the US and the latest decision by the Supreme Court on the issue. Part IV examines the state of the challenges in

[28] *See infra* Part I.

[29] Anu Bradford described this phenomenon as the "Brussels Effect", *see* Anu Bradford, The Brussels Effect (2020); *see infra* Part II.A.9.

[30] *See, e.g.*, Oversight Board, *Former President Trump's Suspension*, Oversight Board (May 5, 2021), https://www.oversightboard.com/decision/fb-691qamhj/ [https://perma.cc/82CB-9PMP].

[31] See *infra* Part IV.A.

moderating hate speech and misinformation online. It examines the latest policy changes major platforms announced in the US and whether they are likely to lead to a division of the internet in the areas of misinformation, incitement to hatred, and limits on hate speech. It also analyses the state of the art technology detecting hate speech and misinformation and evaluates the legal responses in the US and the EU in this respect. Further, the paper explores some other avenues that may help towards combating hate speech and misinformation online, which relate to immunizing and empowering users.

I
Putting Constitutional Rights Theories to the Test

The major issues concerning free speech in regulating social media platforms are who has access rights to them, whether they should be limiting hate speech and other categories of speech they consider inacceptable, and how they should use artificial intelligence to engage in this task and to prioritize speech. These rights go together with legal avenues that provide procedural guarantees for their protection. If this has always been a controversial issue in relation to other media in the past, this is even more the case today. Online platforms largely define the public sphere and the possibilities for citizens to express themselves and to access the views of others.[32] Social media platforms are now providing the infrastructure that is necessary for citizens to engage with one another on topics of public interest.[33] The new techniques that they are using, such as prioritization algorithms, are affecting the very possibility that societies have the option to engage in a public dialogue towards discovering the public good. Scholars have been emphasizing that they are undermining the most important doctrines we hold dear as to the rules of public debate, and they require a new engagement with these theories to discover to what extent they can help point towards solutions that may improve the transparency and openness of the public debate.[34] If John Stuart Mill's emphasis in *On Liberty*[35] on the importance of being exposed to a variety of opinions in all

[32] *See* Jürgen Habermas, The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society (Thomas Burger trans., MIT Press, 1991).

[33] The Supreme Court has recognized itself that platforms are the "[most important places] to celebrate some views, to protest others, or simply to learn and inquire." Packingham v. N.C., 582 U.S. 98, 104 (2017).

[34] Vincent Blasi, *Is John Stuart Mill's On Liberty Obsolete?*, 5 J. Free Speech L. 151 (2024).

[35] John Stuart Mill, *On Liberty*, *in* On Liberty and Other Essays 34–140 (John Gray ed., Oxford University Press USA 1991).

domains of human action as a public good is always relevant, the ways through which people have access to information online raise new challenges. Although social media platforms play an important role in enabling more voices to be heard than ever before, they also have gained enormous ability to define this debate by the use of prioritization algorithms. The use of those algorithms leads to the creation of environments of limited exposure to ideas and beliefs that may be vital to distinguishing what is useful and what is not for some participants in the public debate. This means that regulation may be required in relation to whether and how prioritization algorithms give access to anyone to express themselves but also as regards the mechanisms by which algorithms make content available to users, especially whether there is proper oversight as to whether misinformation and hateful content is circulating. Traditionally, governments were considered the source of danger for expressive freedoms, but today the practices of privately held, multinational corporations also carry great weight. A transatlantic comparison illustrates how governments respond to this new challenge. The constitutional doctrine of the "horizontal effect" makes it easier for European governments and the EU to regulate social media platforms' practices of content moderation, while the state action doctrine dominant in the US complicates any attempt to regulate them.

A. The "Horizontal Effect" Of Constitutional Rights In The EU

In Europe, political theories on the role of the government legitimize government intervention within civil society to address social power imbalances.[36] It is legitimate for the government to intervene and limit strong social actors from protecting the rights of weaker social actors. The doctrine of "horizontal effect" is the legal translation of this conception of the role of the government. According to this doctrine, the Constitution applies not only to the vertical relationship between the state and its citizens, but also to the horizontal relationship between private parties within civil society.[37] This means that constitutional

[36] *See* Ioanna Tourkochoriti, Freedom of Expression, *supra* note 15, at 104.

[37] Constance Grewe & Helene Ruiz Fabri, Droits Constitutionnels Européens 181–83 (1995); Stephen Gardbaum, *The "Horizontal Effect" of Constitutional Rights*, 102 Mich. L. Rev. 387, 388 (2003); Claudia E. Haupt, *The Horizontal Effect of Fundamental Rights*, *in* The Oxford Handbook on Digital Constitutionalism (Giovanni De Gregorio, Oreste Pollicino & Peggy Valcke eds., forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4680759 [https://perma.cc/SM8D-SYX9]; Claudia E. Haupt, *Democratic Self-Defense*, 93 Fordham L. Rev. 1377 (2025).

rights are enforceable by their holders against private parties as well as against the government. The doctrine of "horizontal effect" also authorizes the state to enforce constitutionally protected rights against private parties. This means that governments have obligations under the Constitution to take action to prevent and address violations of constitutional rights, when these occur by other parties within civil society. In general, according to standard constitutional doctrine in several European states, governments have positive obligations to enable the citizens to exercise their rights.[38]

In the area of free speech, this means that the constitutionally protected right to free speech can be asserted against social media platforms when they decide to limit someone's speech through content moderation or to exclude them from using social media platforms altogether. Courts will examine the constitutional rights to free speech of the platforms' users and the rights of the platforms themselves when they are in conflict and will attempt to find a solution that balances the two towards a "practical concordance" of the two rights.[39] The German Federal Constitutional Court affirmed the doctrine in a dispute related to Facebook's attempt to exclude a neo-Nazi political group.[40] The Court accepted Facebook's ability to limit the group's speech when it is hateful, while ordering the reinstatement of the group's account.[41] Most Constitutional Courts of other EU member states also reason in ways that reflect the same ideas.[42] When courts examine cases of conflicts of rights, they usually refer to the constitutional rights of both parties, weigh them, and attempt to find solutions to resolve the conflict by harmonizing the interests of both sides to the greatest extent possible.[43]

---

[38] Ioanna Tourkochoriti, Freedom of Expression, *supra* note 15. *See also* Anu Bradford, Digital Empires: The Global Battle to Regulate Technology (2023) (discussing the EU's rights-based approach to digital regulation).

[39] *See* Haupt, *supra* note 37, at 1407.

[40] *See* BVerfG, May 22, 2019, 1 BvQ 42/19, https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2019/05/qk20190522_1bvq004219.html [https://perma.cc/P6LM-NE8Q].

[41] *Id.* at ¶¶ 5–10.

[42] Grewe, *supra* note 37, at 182. *See, e.g.*, Conseil Constitutionel [CC] [Constitutional Court] decision No. 2020-801DC, June 18, 2020, 1.

[43] On how European courts weigh rights in situations of conflicts towards finding solutions that try to harmonize the interests on both sides and the relevant criteria they use in this respect across several European states, *see* Ioanna Tourkochoriti, *Is Neutrality Possible? A critique of the Court of Justice of the European Union on Headscarves in the Workplace from a Comparative Perspective*, 71 Am. J. Comp. L. 444 (2023).

The European Commission in its preparatory documents before the enactment of the Digital Services Act (DSA) evaluated the importance of all relevant rights and elaborated regulation of social media platforms that reflects this mentality.[44] The European Charter has clauses protecting the right to free speech,[45] as well as the right to property[46] and the freedom to conduct a business.[47] The Charter also protects the right to dignity, and the right not to experience discrimination.[48] Conflicts of rights within civil society in the European Union are conceptualised as conflicts between two equally important rights under the Charter.[49] This means that the users' and the platforms' rights to free speech have equal value. Within the national legal systems, the same conflict of rights is conceptualized as a conflict between two rights of equal constitutional value, and it is the mission of the government to decide which one should be protected depending on the circumstances. The DSA complements this legal framework within each Member State. The Commission in its preparatory documents appears to weigh the rights of the platforms to conduct business and the expressive rights of the users.[50] It identifies important risks to the fundamental rights of the users that the platforms entail.[51] The relevant documents indicate that he European Commission aims to protect the users' expressive rights and their right to have access to information over the platforms' rights to conduct their business by regulating the circumstances when the platforms may remove a user.[52] The DSA makes enforceable against social media platforms hate speech offenses that already exist within the EU member states. When national courts apply the DSA, they will see it as complementing the doctrine of horizontal effect that already exists within the EU member states. Although the Commission does not refer explicitly to the

[44] *Id.*

[45] Article 11 of the European Charter of Rights protects: "Article 11. Freedom of expression and information. 1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. 2. The freedom and pluralism of the media shall be respected." 2012 O.J. (C 326) 391, 398.

[46] *Id.* at 399.

[47] *Id.*

[48] *Id.* at 400.

[49] See Eur. Comm'n, *supra* note 4, at 50, 61.

[50] *Id.* at 18.

[51] *Id.*

[52] *See infra* Part II.A.1.

constitutional doctrine of horizontality, the rationale it offers to regulate social media platforms reflects the idea that the fundamental rights of the users should be protected against the platforms. In this way, the Commission presupposes the horizontal effect of the fundamental rights that exist in the European Charter.[53]

Even before the EU took action, several European states had regulated social media platforms.[54] As analysed below, the DSA strengthens the legal avenues citizens have when their speech is limited by the platforms and obliges the platforms to engage in content moderation to prevent users from being exposed to hate speech and misinformation. The DSA is only one among a series of Acts the EU has enacted which reflect a conception of the role of the government as being to address social power imbalances. The same conception has led the EU to also legislate the Corporate Sustainability Due Diligence Directive which makes enforceable the "due diligence" responsibilities that private corporations have to prevent human rights violations by their subsidiaries and subcontractors around the world.[55] Under this Directive, private actors and corporations now have legal obligations to prevent and address any human rights risks in the operations of their subsidiaries and subcontractors abroad. The EU is also elaborating the AI Act which aims to provide the first comprehensive regulation of the sector.[56] The legislation aims to "ensure better conditions for the development and use of [AI]" by focusing on the risks that AI creates.[57] EU law ranks at the top of the hierarchy of legal rules for EU member states and it now co-defines the standards of Constitutional protection of rights.[58]

B. Versus the "State Action" Doctrine In The U.S.

By contrast to the doctrine of horizontal effect, in the United States the "state action" doctrine means that the protection of constitutional rights applies

[53] *See, e.g.*, Eur. Comm'n, *supra* note 4, at 18.

[54] *See infra* Part II.B.

[55] Directive (EU) 2024/1760 of the European Parliament and of the Council of 13 June 2024 on Corporate Sustainability Due Diligence and Amending Directive (EU) 2019/1937 and Regulation (EU) 2023/2859 (Text with EEA relevance), 2024 O.J. (L 1760) 1.

[56] Eur. Parliament, *EU AI Act: First Regulation on Artificial Intelligence*, Topics Eur. Parliament, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [https://perma.cc/HZZ7-RNG2] (last visited Mar. 7, 2026); *see also* 2024 O.J. (L 1689) 1.

[57] *Id.*

[58] On the role of EU constitutional standards within the Member States, see Giovani De Gregorio, Digital Constitutionalism in Europe 20–22 (Cambridge University Press 2022).

only against the government.[59] Citizens cannot enforce the protection of their constitutional rights against social media platforms through courts. The freedom of speech of social media platforms covers whether they will give access to any citizen to express themselves. In the absence of government regulation in this area, social media platforms have created modes of self-regulation to prevent the spread of hate speech, among others. They have appointed content moderators with linguistic and cultural expertise.[60] In addition, Facebook created a private body, the Facebook Oversight Board, with authority to review content that is removed and content that is kept up.[61] In *Moody v. NetChoice*,[62] its latest decision on the issue, the American Supreme Court emphasized the protection of free speech for the platforms in relation to their right to engage in content moderation. Concerning the legal avenues against the platforms citizens may have when their rights are limited, the Court deferred to the commercial speech doctrine.[63]

The emergence of social media and the new challenges inherent in online communication have led many scholars to advocate restrictions on extreme speech, even within the US, where such limits may conflict with national constitutional obligations. The First Amendment doctrine elaborated by the US Supreme Court during the 20th century accepts protection for speech that is much wider than other constitutional systems. Several scholars in recent years, however, have argued that new dangers that emerge in online communication and social networks necessitate government intervention to limit speech and to limit how online platforms

[59] Mark Tushnet, *The Issue of State Action/Horizontal Effect in Comparative Constitutional Law*, 1 Int'l J. Const. L. 79, 79–98 (2003); Charles L. Black Jr., *Foreword: "State Action," Equal Protection, and California's Proposition*, 81 Harv. L. Rev. 69, 69–262, (1967); Louis Michael Seidman & Marc V. Tushnet, *The State Action Paradox*, *in* Remnants of Belief, Contemporary Constitutional Issues 49, 49–71 (1996); Robert J. Glennon, Jr. & John E. Novak, *A Functional analysis of the Fourteenth Amendment "State Action" Requirement*, 1976 Sup. Ct Rev. 221, 221–261 (1976).

[60] *See* Meta, *How Review Teams Work*, Meta Transparency Center, https://transparency.meta.com/enforcement/detecting-violations/how-review-teams-work/ [https://perma.cc/97BU-KANR] (last updated Nov. 12, 2024).

[61] Katie Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L.J. 2418, 2418–99 (2020); Kenji Yoshino, *Reconsidering the First Amendment Fetishism of Non-State Actors: The Case of Hate Speech on Social Media Platforms and at Private Universities*, 76 Stan. L. Rev. 1755, 1755–85 (2024).

[62] *See supra* note 8.

[63] Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2399 (2024).

operate.[64] For these scholars, the 20th century's First Amendment doctrine of strong protection for free speech has become obsolete.[65] The communications technologies have changed since the First Amendment doctrine was elaborated.[66] While in the past the doctrine presupposed that citizens have time to evaluate the information available to them, attentional scarcity has become an important issue.[67] The internet, by altering the social epistemology of societies, necessitates a reconceptualization of doctrines articulated in reference to the media societies used in the past.[68] Other scholars emphasize the need to value speech even when it occurs in social platforms in relation with the social practice within which it occurs, to justify limitations when speech does not serve the public sphere.[69] Speech must serve the formation of a healthy public opinion towards ensuring state legitimacy, which means that the legal doctrine of free speech must move away from the current "libertarian, deregulatory agenda" that the current Supreme Court has been following.[70]

Justifying legislation to regulate online social media platforms in the United States is challenging under the First Amendment. A new paradigm is needed to address the impact of social media platforms on individual liberties. Jack Balkin proposes to consider platform regulation as part of the contemporary "pluralist system of speech regulation."[71] This means that speech may be regulated by the government. It may also be regulated by how the government regulates social media platforms. It may also be regulated by how private parties, the social media platforms, decide to regulate their users' speech via "community standards".

---

[64] Tim Wu, *Is the First Amendment Obsolete?*, Knight First Amend. Inst. (Sept. 1, 2017) (noting that the First Amendment was elaborated in a different era focusing exclusively on protecting speakers from government); Brian Leiter, *Free Speech on the Internet: The Crisis of Epistemic Authority*, 5 J. Free Speech L. 241 (2024); Brian Leiter, *The Epistemology of the Internet and the Regulation of Speech in America*, 20 Geo. J. L. & Pub. Pol'y 903 (2022). Several scholars are also emphasizing that content moderation may actually secure opportunities for more people to speak. *See generally* Danielle Keats Citron & Johnathan Penney, *Empowering Speech by Moderating It*, 5 J. Free Speech L. 171 (2024).

[65] *See generally* Wu, *supra* note 64.

[66] *Id.*

[67] *Id.*

[68] *See generally, e.g.*, Leiter, *Free Speech on the Internet, supra* note 64.

[69] Robert Post, *The Unfortunate Consequences of A Misguided Free Speech Principle*, 5 J. Free Speech L. 295 (2024).

[70] *Id.* at 308.

[71] Jack Balkin, *Moody v. NetChoice: The Supreme Court Meets the Free Speech Triangle*, 2024 Sup. Ct. Rev. 127, 130 (2025).

Balkin calls this situation "the free speech triangle."[72] It is composed, firstly, by "old-school" speech regulation, the government regulation of the speech of private individuals, secondly, by "new-school" speech regulation, the government's attempt to regulate the owners of private infrastructure "up and down the tech stack", in ways that affect their ability to regulate speech, and thirdly, by the regulation of the owners and operators of digital infrastructure of how users behave in the platforms.[73] Scholars emphasize that it is important to prevent the platforms from engaging in viewpoint discrimination, and to make sure that they give their users some "due process type protections" such as "advance notice" of the platforms content guidance "when their speech is censored or otherwise regulated," or "when the speaker is deplatformed," information on which rules were allegedly violated and "a meaningful opportunity to challenge content moderation in cases where such moderation restricts their exercise of free speech."[74] Viewpoint discrimination aside, the DSA regulates the platforms and ensures that all these "due process type protections" exist for users in the EU. The DSA obliges the platforms to continue their content moderation practices as far as hate speech is concerned.

The Supreme Court in *Murthy v. Missouri*[75] held that individual users of social media platforms do not have standing to complain when social media platforms are limiting posts they consider false or misleading following government encouragement around public health emergencies, as was the case during the COVID-19 pandemic.[76] Users whose speech was limited by the platforms following the government's encouragement do not have standing to request an injunction to stop "certain Government agencies and employees from coercing or encouraging the platforms to suppress speech."[77] *Moody v. NetChoice*[78] points in the direction of commercial speech for a doctrine that would allow the government to regulate how the platforms treat issues of access to users.

---

[72] *Id.*

[73] *Id.* at 129, 131.

[74] Dawn Carla Nunziato, *Protecting Free Speech and Due Process Values on Dominant Social Media Platforms*, 73 HASTINGS L.J. 1255, 1259–60 (2022).

[75] Murthy v. Missouri, 144 S. Ct. 1972 (2024).

[76] *Id.* at 1973.

[77] *Id.* at 1995. Furthermore, for the Supreme Court, the plaintiffs failed to prove that the content moderation by the platforms caused them identifiable harm. *Id.* at 1996.

[78] 144 S. Ct. 2383 (2024).

The Supreme Court in the past has held that private actors, when they are hosting others' speech, may not limit the variety of opinions that exist in society. In a series of cases related to owners of company towns and malls, *Marsh v. Alabama*,[79] *Amalgamated Food Employees Union Local 590 v. Logan Valley Plaza Inc.*,[80] and *PruneYard Shopping Ctr. v. Robins*,[81] the Court held that the owners-hosts are not allowed to exclude speakers. Those precedents would offer a more promising avenue for finding that online social media platforms should protect their users' speech and guarantee some procedural rights to them if they decide to exclude them. As analyzed below, *Moody v. NetChoice* distinguishes these cases from what is at stake in social media platforms' moderation as far as viewpoint discrimination is concerned.[82] For Justice Kagan, who wrote the opinion, these cases are to be distinguished because owners of malls are not engaging in expressive activity themselves, unlike social media platforms, who are curating content which is itself an expressive activity.[83] Justice Kagan distinguishes this case also from *Turner Broadcasting System, Inc. v. FCC* (*Turner II*)[84] which upheld a requirement for cable operators under the Cable Television Consumer Protection and Competition Act of 1992 to dedicate some of their channels to local broadcast television stations. For Justice Kagan, *Turner II* was "necessary to prevent the demise of local broadcasting."[85] She also clarified that even under this case, "a private party's collection of third-party content into a single speech conduct … is itself expressive and intrusion into that activity must be specially justified under the First Amendment."[86]

To work around the difficulties that emerge from the "free speech triangle," scholars suggested that the common carrier doctrine should apply in respect

[79] Marsh v. Alabama, 326 U.S. 501 (1946).

[80] Amalgamated Food Employees Union v. Logan Valley Plaza, 391 U.S. 308 (1968).

[81] PruneYard Shopping Ctr. v. Robins, 447 U.S. 74 (1980).

[82] *Moody*, 144 S. Ct. at 2401.

[83] *Id*.

[84] *Turner Broad. Sys., Inc. v. FC*C, 520 U.S. 180, 185, 189–90 (1997).

[85] *Moody*, 144 S. Ct. at 2400.

[86] *Id*. This argument resembles the rationale the court applied in *Red Lion Broad. v. FCC*, 395 U.S. 367 (1969), where the Court upheld the "fairness doctrine" for broadcast media, citing spectrum scarcity and a governmental interest in managing public airwaves.

to some of the features platforms are offering,[87] while others emphasize that the doctrine is contrary to the First Amendment.[88] The Common Carrier doctrine justifies regulating businesses that serve the public and imposing special obligations on them such as the duty to serve all customers without discrimination.[89] The principle ensures fairness and prevents unjust preferences or prejudice in the provision of services. The Supreme Court has suggested that the regulation of common carriers is justified when a business, "by circumstances and its nature … rise[s] from private to be of public concern."[90] Originally, the businesses covered were transportation and, more recently, communication networks. Volokh has suggested that the common carrier theory is relevant to the decisions platforms make as far as the speech they host, but that platforms do have a First Amendment right to choose what they affirmatively recommend to their users.[91] He argues that in some respects social media platforms serve as an infrastructure for communication, just like the US Postal Service, or private shipping and logistics companies such as the United Parcel Service (UPS), or phone companies such as Verizon.[92] And just like UPS and Verizon cannot refuse to carry electoral materials depending on their viewpoint, social media platforms should not refuse either.[93] Volokh also suggests that the hosting activities of the platforms—their letting users post materials on the users' own pages and to those who deliberately visit that page or subscribe to its feed—may be covered by the

[87] For arguing that the common carrier doctrine may be relevant in relation to some functions of the platforms, *see* Eugene Volokh, *Treating Social Media Platforms Like Common Carriers?*, 1 J. Free Speech L. 377 (2021). *See also* Nunziato, *supra* note 74, at 1286.

[88] James B. Speta, *Boden Lecture: The Past's Lessons for Today: Can Common-Carrier Principles Make for a Better Internet?*, 106 Marq. L. Rev. 741 (2023) (opinion arguing that the solution to problematic platforms is more platforms); Ashutosh Bhagwat, *Why Social Media Platforms Are Not Common Carriers*, 2 J. Free Speech L. 127, 151–52 (2022); Eric Goldman, *A Short Explainer of Why California's Mandatory Transparency Bill (AB 587) Is Terrible*, Tech. & Mktg. L. Blog (Aug. 9, 2022), https://blog.ericgoldman.org/archives/2022/08/a-short-explainer-of-why-californias-mandatory-transparency-bill-ab-587-is-terrible.htm [https://perma.cc/X9EM-CMGH].

[89] Scofield v. Lake Shore & M. S. Ry. Co., 43 Ohio St. 571 (Ohio 1885).

[90] *See* German All. Ins. Co. v. Lewis, 233 U.S. 389, 411 (1914) (affirming state regulation of fire insurance rates).

[91] Volokh, *supra* note 87, at 409.

[92] *Id.*

[93] *Id.* A concern about the private messaging infrastructure that platforms are offering is present in Justice Kagan's opinion in *Moody v. NetChoice*, 144 S. Ct. 2383. The decision remanded the case to the lower courts because they had not examined whether legislation limiting platform's content moderation practices covered the private messaging apps social media platforms are offering to the public. *Id.*

common carrier doctrine.[94] This relates to the platforms' decisions to remove posts and to remove users.[95] Some other of the features platforms offer, however, such as recommending certain posts, should be covered by the First Amendment.[96] Volokh also thinks that some content moderation of the comments others post on a user's page could be allowed under the First Amendment.[97]

In his concurring opinion in *Biden v. Knight First Amendment Inst. at Columbia Univ.*, Justice Thomas engaged with the analogy of social media platforms to common carriers and public accommodations and did not exclude the option of regulating platforms in reference to these doctrines.[98] The public accommodations doctrine could be another avenue towards regulating social media platforms in the US.[99] The public accommodations doctrine was developed to make sure everyone has access to the goods and services that private entities are offering to the public.[100] Social Media platforms today define the public sphere. In this respect they approximate the public interest requirement that the public accommodations doctrine serves. The public accommodations doctrine is compatible with the expressive rights of the social media platforms to engage in content moderation and to exclude the users they want to exclude. It can also help reasoning by analogy towards protecting procedural rights for the users of online social media platforms. In this respect, it may contribute to protecting users' access and procedural rights in a way that approximates the legal regime that exists in Europe. Under the public accommodations doctrine, businesses are allowed to exclude customers if they pose a "direct threat" to the health or safety of other customers.[101] The platforms should be free to define to what extent they believe that speech articulated by some users is contrary to their community standards. The doctrine allows room for protecting some procedural rights for those who have

[94] Volokh, *supra* note 87, at 409.

[95] *Id.*

[96] *Id.* at 451.

[97] *Id.* at 411–12.

[98] Biden v. Knight First Amend. Inst. at Colum. Univ., 141 S. Ct. 1220, 1226 (2021).

[99] For a criticism of this opinion, *see* Christopher Yoo, *The First Amendment, Common Carriers, and Public Accommodations*, 1 J. Free Speech L. 463 (2021).

[100] *See* Joseph William Singer, *No Right to Exclude: Public Accommodations and Private Property*, 90 Nw. U. L. Rev. 1283 (1996).

[101] Lockett v. Catalina Channel Express, Inc., 496 F.3d 1061, 1065 (9th Cir. 2007). State legislation foresees the same possibility, *see* Ariz. Rev. Stat. § 41-1492.02; Fla. Stat. § 509.142.

been excluded by online social media platforms.[102] If public accommodations may exclude customers, the latter also enjoy legal protection. Under the doctrine, social media platforms would be obliged to include everyone who wants to have access to them, while they would be allowed to exclude those who violate their community standards just like any public accommodation or store owner is allowed to remove customers who harm other customers.

As analyzed below, *Moody v. NetChoice*[103] moves away from the common carrier and the public accommodations doctrines as far as viewpoint discrimination is concerned in favor of recognizing expressive rights for platforms.[104] The decision examined legislation enacted in Texas and Florida which aimed to prevent social platforms from engaging in content moderation which impacted particular conservative viewpoints.[105] For the Court, the platforms' curating of content approximates the editorial discretion newspapers have enjoyed under *Miami Herald Co. v. Tornillo*.[106] The decision allows the platforms to pursue their content moderation policies even if it requires engaging in viewpoint discrimination. Although in other past cases, the Supreme Court's analysis has recognized the important role social media platforms play today in providing the infrastructure for the public sphere,[107] the Court does not seem willing to weigh the expressive rights of the users against the ones of the platforms in the way European Courts do thanks to the doctrine of "horizontal effect." The editorial discretion social media platforms have by "curating" speech means that they can decide whom to include and whom to exclude. Any legal avenues against the platforms that users as speakers might have cannot be justified on the basis of their free speech rights, as is the case

[102] Such as the right to a notice, the right to file a complaint to the platform, the right to some extra-judicial independent mechanism, and the right to receive a response within a reasonable timeframe, if they believe their speech has been unfairly limited in reference to a platform's "community standards."

[103] For an analysis, *see infra* Part IV.B.

[104] *Supra* Part II.B.

[105] Moody v. NetChoice, LLC, 144 S. Ct. 2383, 2396, 2407 (2024).

[106] *See* 418 U.S. 241 (1974).

[107] *See, e.g.*, Lindke v. Freed, 601 U.S. 187 (concerning when a public official's social media activity constitutes state action under section U.S.C. 1983. Section 1983, which provides a cause of action against "[e]very person who, under color of any statute, ordinance, regulation, custom, or usage, of any State" deprives someone of a federal constitutional or statutory right. The court held that social media activity by public officials may constitute state action when the official has actual authority to speak on the state's behalf and purports to exercise that authority when they speak on social media platforms. If, in these circumstances, they delete users' comments to their posts, they are violating the users' First Amendment Rights.)

in the EU. The Court does not recognize any legal foundation for procedural rights that platform users might have if they are excluded by the platforms.

Some scholars believe that the reluctance to regulate social media platforms in the United States as compared to regulation in the EU is likely to continue after *Moody v. NetChoice*.[108] For others, the decision does not preclude any regulation whatsoever, and more sophisticated regulation than that enacted in Texas and Florida may survive constitutional scrutiny.[109] Several scholars have emphasized that European-style positive action of the government is necessary to ensure access rights to the users of large social media platforms.[110] Although the First Amendment restricts only government action, legislation requiring platforms to host speech "might advance First Amendment interests" and the government may need to ensure access to privately owned channels of communication.[111] Although scholars emphasize that the amount and the type of control that platforms have over their users' speech "rivals or exceeds government power," which makes government regulation necessary to advance the interests that the First Amendment itself protects,[112] the current First Amendment doctrine as applied to social media platforms in *Moody* does not allow for these concerns to prevail.

The current state of the law in relation to social media platforms means that Europeans have more legal avenues to claim the protection of their rights when social media platforms limit them. They have more legal avenues to protect their right to have access to social media platforms. And although platforms may engage in more extended limitations of speech in the area of hate speech due to different standards of content moderation that are emerging between Europe and the US, European citizens have more avenues to litigate these limits as well.

---

[108] *See* Balkin, *supra* note 71.

[109] Evelyn Douek & Genevieve Lakier, *Lochner.com?*, 138 Harv. L. Rev. 100 (2024); Kyle Langvardt & Alan Z. Rozenshtein, *Beyond the Editorial Analogy: First Amendment Protections for Platform Content Moderation After Moody v. NetChoice*, 6 J. Free Speech L. 1 (2025).

[110] Langvardt & Rozenshtein, *supra* note 109 at 35.

[111] *Id.*

[112] *Id.*

II
Legal Developments in the European Union

A. From the Code of Conduct to the Digital Services Act (DSA)

The European Union has taken the lead in regulating the internet. Following its very influential regulation of online data privacy,[113] it elaborated in the first months of 2022 the Digital Services Act (DSA).[114] The DSA has entered into force progressively since January 2024. It foresees a sophisticated mechanism for content moderation of online platforms. It builds upon a previous regime that had already been elaborated by the European Union in 2016: a Code of Conduct on countering illegal hate speech online.[115] The Code was agreed upon between the EU and Facebook, Microsoft, Twitter and YouTube and led them to develop community codes and to start performing content moderation. The code obliged IT companies to put in place Rules or Community Guidelines clarifying that they prohibit the promotion or incitement of violence and hateful conduct. The code defines illegal hate speech as "publicly inciting to violence or hatred directed against a group of persons or a member of such group defined by reference to race, color, religion, descent or national or ethnic origin."[116] On the basis of the code, the Companies were obliged to create clear and effective processes for reviewing notifications regarding illegal hate speech on their platforms and to remove or disable access to such content.[117] The code created soft law which had great impact. It led these major platforms to modify their operation and to create local mechanisms of content moderation within each state where they provide their

[113] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 2016 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/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[114] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), O.J. (L 277) 1 [hereinafter DSA].

[115] Eur. Comm'n, *EU Code of Conduct on Countering Illegal Hate Speech Online*, Eur. Comm'n, 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 [https://perma.cc/9KA8-DPHD].

[116] For this definition, the Code refers to Council Framework Decision 2008/913/JHA of 28 Nov. 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 2008 O.J. (L 328) 55.

[117] Eur. Comm'n, *supra* note 115.

services. The implementation of the code was monitored annually in collaboration with a network of organizations located in various EU countries.[118] Using a commonly agreed methodology, these organizations tested how the IT companies were implementing the commitments in the Code.[119]

The DSA creates a protective regime for users of online media. It is part of the "New School" speech regulation, to follow Jack Balkin's terminology.[120] Unlike "Old School Regulation," which aimed to regulate the speakers themselves, "New School" regulates the infrastructure for speech.[121] It aims to ensure freedom of speech by regulating the social media platforms. It attempts to strike a balance between protecting free speech and limiting "illegal" hate speech.[122] To some extent it enhances free speech rights by creating a process for limiting speech and timelines for the duration of these limits. It regulates the circumstances in which online platforms may exclude users.[123] It enhances the obligation to implement a 'notice and action mechanism' (which already existed in the code of conduct) to alert about the presence of content or information they consider to be illegal. It enhances the current internal complaint-handling system that some platforms maintain. In addition, it foresees the possibility for out-of-court dispute settlements. It provides for the creation of new national and European bodies which will oversee its application.[124] It also authorizes the Member states to enact significant penalties related to the violation of its clauses.[125] It enhances transparency for the process by creating reporting obligations for the platforms.[126]

---

[118] On the monitoring of the implementation of the code, the European Commission has periodically published reports. *Id*.

[119] *See, e.g.*, Didier Reynders, *Countering Illegal Hate Speech Online 7th Evaluation of the Code of Conduct*, Eur. Comm'n, https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en?filename=Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf [https://perma.cc/EQJ3-S8VV].

[120] *See* Jack M. Balkin, *Old School/New School Speech Regulation*, 127 Harv. L. Rev. 2296 (2014); Jack M. Balkin, *Free Speech in the Algorithmic Society: Bid Data, Private Governance, and New School Speech Regulation*, 51 U. C. Davis L. Rev. 1149 (2018).

[121] Balkin, *supra* note 120, at 2306.

[122] See *infra*, Part II.A.1.

[123] *See Id*.

[124] *See infra*, Part II.A.5.

[125] *See infra*, Part II.A.4.

[126] *See infra*, Part II.A.6

The Act largely aims to address systemic problems that emerge in the platforms' content moderation processes. It was motivated by the need to set a standard of transparency and accountability on how the major platforms moderate content and use algorithms.[127] It obliges them to develop appropriate risk management tools. As explained in the Memorandum, the Act aims to mitigate risks of erroneous or unjustified blocking of speech, to address the chilling effects on speech, to stimulate the freedom to receive information and to hold opinions, and to reinforce users' redress possibilities.[128] It recognizes that it may be the case that some groups or persons may be vulnerable or disadvantaged in their use of online services because of their gender, race or ethnic origin, religion or belief, disability, age or sexual orientation.[129] It also recognizes that these users may be disproportionately affected by restrictions and removal measures following from unconscious or conscious biases, potentially embedded in the notification systems by users and third parties, as well as replicated in automated content moderation tools used by platforms.[130] It aims to create mandatory safeguards for the removal of users' information, which include the provision of explanatory information to the user, complaint mechanisms and external out-of-court dispute resolution mechanisms.[131]

1. Users' Procedural Rights

The Act enhances free speech rights by creating processes which online platforms should follow if they wish to exclude users. They can only suspend users found to frequently provide illegal content after having issued a prior warning.[132] This exclusion may last only for a reasonable period of time.[133] The Act creates an obligation for platforms to assess—in a timely, diligent and objective manner—whether a recipient, individual entity, or complainant engages in misuse.[134] In this evaluation, they shall take into consideration the numbers of items of manifestly

[127] Eur. Comm'n, Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM (2020) 825 final (Dec. 15, 2020).
[128] *Id.*
[129] *Id.* at 12.
[130] *Id.* at 12.
[131] *Id.* at 12.
[132] DSA, *supra* note 114, art. 23(1).
[133] *Id.*
[134] *Id.* art. 23(3).

illegal content or unfounded notices or complaints submitted, their proportion to the total number of items of information provided or notices submitted, the gravity of the misuses and its consequences, the intention of the recipient, or complainant.[135] The platforms must have a clear and detailed policy in respect of misuse.[136] They are obliged to notify the law enforcement or judicial authorities of the Member State or Member States where they are suspicious that a serious criminal offence has taken place.[137] This measure responds to concerns that have been raised about content moderation, given that platforms in the past intervened based on the behavior of groups or accounts and the actors or associations behind them.[138] In the past, behavioral content moderation was opaque, because giving notice and reasons to users was seen as undermining the effectiveness of rules rather than promoting compliance.[139]

The Act also enhances the system of complaints that platform users have when their speech is limited. Online platforms are obliged under the act to provide to recipients of the service, for a period of at least six months, access to an effective internal complaint-handling system if they decide to remove or disable access to information, if they decide to suspend or terminate the provision of service in whole or in part to recipients, and if they decide to suspend or terminate the recipients' account.[140] The platforms are obliged to handle complaints in a timely, diligent, and objective manner.[141] They are also obliged to agree to out-of-court dispute settlements to resolve disputes related to these decisions, including complaints that could not be resolved by means of the internal complaint-handling system.[142] These out-of-court dispute settlement bodies will be certified by the Digital Services Coordinators of the EU Member States.[143] These Coordinators must certify that the body is impartial and independent of online platforms and recipients, that they have the necessary expertise in relation to areas of illegal content, that they are easily accessible online through electronic communication

[135] *Id.* art. 23(3).
[136] *Id.* art. 23(4).
[137] *Id.* art. 18(1).
[138] *See* Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 526, 539–40 (2022).
[139] *Id.* at 540.
[140] DSA, *supra* note 114, arts. 17(3)(f), 20(1).
[141] *Id*. art. 20(4).
[142] *Id*. art. 21(1).
[143] *Id*. art. 21(3). On the Digital Services Coordinators, *see infra* Part II.A.5.

technology, and that they are capable of settling disputes in a swift, efficient, and cost-effective manner.[144]

2. Protection Against Hate Speech and Misinformation-related Systemic Risks

The Act also aims to protect users against exposure to hate speech. It creates due diligence obligations for a transparent and safe online environment.[145] It obliges online platforms to implement notice and action mechanisms.[146] Based on these mechanisms, any individual or entity may notify them of the presence on their service of specific items of information that they consider to be illegal hate speech.[147] The platforms are obliged to process the notices that they receive and decide in a timely, diligent, and objective manner—whether a piece of information should be removed or not.[148] If in acting on these decisions, a platform decides to remove or disable access to specific items of information, the platform should inform the recipients of the service about the decision to remove content or disable access, and they will provide a clear statement of the reasons for their decision.[149] Platforms should also provide information on the redress possibilities available to the recipients of their service in respect of the decision, in particular as this relates to internal complaint-handling mechanisms, out-of-court dispute settlement, and judicial redress.[150] Platforms also may suspend, for a reasonable period of time and after having issued a prior warning, the processing of notices and complaints submitted through the notice and action mechanisms and internal complaints-handling systems if they estimate that there is misuse of these mechanisms.[151]

The Act creates additional obligations for very large online platforms to manage systemic risks. The Preamble to the DSA emphasizes that the systemic risks the platforms pose may have a disproportionately negative impact in the Union when the number of recipients of a platform reaches a significant share of the Union population.[152] This reach exists where the number of recipients exceeds a threshold

[144] DSA, *supra* note 114, art. 21(3).
[145] *Id.* ch. 3.
[146] *Id.* art. 16.
[147] *Id*.
[148] *Id*. art. 16(6).
[149] *Id*. art. 17.
[150] *Id.* pmbl. ¶ 64, art. 17(3)(f).
[151] *Id*. art. 23(2).
[152] *Id.* pmbl. ¶ 76.

of 45 million, that is the 10% of the Union population.[153] Corporations that have such an impact should, under the Act, bear "the highest standard of due diligence obligations."[154] Platforms which provide their services to over 45 million users in the Union (a number to be constantly reevaluated in reference to the increase or decrease of the population of the Union) must identify, analyze and assess at least once a year any systemic risks stemming from the functioning and the use of their services in the Union.[155] The risks the platforms should be diligent about include the dissemination of illegal content through their services, any negative effects for the exercise of the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination, and the rights of the child in conformity with the articles of the Charter of Fundamental Rights in the Union.[156] They should also analyze the risks of intentional manipulation of their service, including by means of inauthentic use or automated exploitation of the service with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security.[157] The platforms must put in place reasonable mitigation measures tailored to the specific systemic risks they pose. These measures may include adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services and of their terms and conditions.[158] They should also include targeted measures aimed at limiting the display of advertisements in association with the service they provide.[159] They must also reinforce the internal processes or supervision of their activities in particular as regards detection of systemic risk.[160] One of the concerns that motivated this need for a systemic response is the concern that frequently it is groups of accounts that violate rules.[161]

The platforms are also obliged to put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks

[153] *Id.* arts. 33(1)–(2).
[154] *Id.* pmbl. ¶ 76.
[155] *Id.* art. 34(1).
[156] *Id.*
[157] *Id.* art. 34(1)(c).
[158] *Id.* art. 34(2).
[159] *Id.* art. 35(1).
[160] *Id.* art. 34(1).
[161] *See* Douek, *supra* note 138, at 539–40.

they identify.[162] These measures may include adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions. They may also include targeted measures aimed at limiting the display of advertisements in association with the service they provide. Several of these obligations imposed by the DSA would be considered contrary to the First Amendment in the US under the "editorial discretion" doctrine.[163]

One of those risk mitigation measures is signing up to the strengthened Code of Practice on Disinformation elaborated by the EU in 2022.[164] The Code was elaborated by the EU in response to the fact that platforms rely on third-party fact-checkers' judgments to guide content moderation.[165] In the area of political advertising, this Code of Practice requires the elaboration of proportionate and appropriate identity verification systems in place for sponsors and providers of advertising services.[166] It requires a commitment to provide transparency information to users on the political or issue ads they see on their service.[167] The platforms who decide to sign on to this code also commit to providing users with clear, comprehensible and comprehensive information about why they see a political or issue ad.[168] They also commit to maintaining repositories of political or issue advertising and ensuring their correctness, completeness, usability and quality, such that they contain all political and issue advertising along with the necessary information to comply with their legal obligations and with transparency commitments under the Code.[169] The Code also recommends that platforms should strengthen their efforts to better equip users to identify disinformation.[170] The participants commit to facilitating user access to tools for assessing the factual accuracy of sources through fact-checks from fact-checking organizations that have flagged potential disinformation and warning labels from other authoritative

[162] DSA, *supra* note 114, art. 27(1).
[163] Eric Goldman, *The Constitutionality of Mandating Editorial Transparency*, 73 Hastings L.J. 1203, 1220 (2022).
[164] Eur. Comm'n, *2022 Strengthened Code of Practice on Disinformation* (June 16, 2022).
[165] *See* Douek, *supra* note 138, at 543.
[166] Eur. Comm'n, *supra* note 164, Commitment 7.
[167] *Id.* Commitment 8.
[168] *Id.* Commitment 9.
[169] *Id*. Commitment 10.
[170] *Id*. Commitment 21.

sources.[171] In this respect, they should develop and apply policies, features, or programs to help users benefit from independent fact-checkers or authoritative sources by means of labels, such as labels indicating fact-checker ratings, notices to users who try to share or previously shared the rated content.[172] The platforms also commit to develop and apply tools or features to inform users, through measures such as labels and notices that independent fact-checking has taken place.[173] The platforms are obliged to report on the independent fact-checkers they have used.[174] They also commit to provide users with tools to help them make more informed decisions when they encounter online information that may be false or misleading, and to facilitate user access to tools and information to assess the trustworthiness of information sources, such as indicators of trustworthiness for informed online navigation, particularly relating to societal issues or debates of general interest.[175]

The Code also foresees that the signatories will provide details of the basic criteria they use to review information source and disclose relevant safeguards put in place to ensure that their services are apolitical, unbiased, and independent.[176] It foresees that the platforms should integrate, showcase, and consistently use fact-checkers' work in their services, processes, and contents across Member States.[177] They commit to creating a repository of fact-checking content that will be governed by the representatives of fact-checkers.[178] They also commit to operating on the bases of strict ethical and transparency rules.[179] These rules will comply with the requirements of instruments such as the International Fact-checking Network (IFCN) Code of Principles or the future Code of Professional Integrity for Independent European fact-checking organizations.[180] This requirement was integrated because it became obvious that the platforms were the ones who decided in the past which fact-checkers to trust, what content fact-checkers will have access to, and how fact-checking would lead to decision-making.[181]

---

[171] *Id*.
[172] *Id.* Measure 21.1.
[173] *Id.*
[174] *Id.*
[175] *Id*. Commitment 22.
[176] *Id*. QRE 22.4.1.
[177] *Id*. Commitment 31.
[178] *Id*. Measure 32.2.
[179] *Id*. Commitment 33.
[180] *Id*. Measure 33.1.
[181] *See* Douek, *supra* note 138, at 544.

The Act also asks the platforms to develop crisis protocols for addressing crisis situations in case of extraordinary circumstances affecting public security or public health.[182] The Commission will also collaborate with the online platforms towards the drawing up, testing and application of these crisis protocols.[183] These protocols will include numerous measures such as displaying information on the crisis situation provided by the Member states' authorities or at Union level, ensuring that a point of contact is responsible for crisis management.[184]

The first report on the implementation of these protocols, issued by the European Board for Digital Services, indicates that the platforms have taken measures to mitigate systemic risks stemming from the design, features, or interfaces of their systems.[185] Platforms mitigate systemic risks by limiting social functionalities and user interactions (e.g., chat), restricting the dissemination of user-generated content, and disabling content personalization by default.[186] They also employ blocking features to prevent users from seeing or interacting with content, or they control the conditions under which others may view or engage with content deemed misinformative or harmful.[187] Additionally, platforms blur images and videos and include warning messages for users.[188]

The report also notes that platforms are creating early detection models to identify dangerous viral trends, such as harmful challenges.[189] They are using wordlists—including anti-circumvention measures to adapt to new slang terms and emojis—especially for content posted by minors, as well as early detection models to identify emerging trends such as dangerous challenges.[190] Some platforms employ automated transcription of audio into text and automated analysis of the resulting text to improve content moderation.[191] They use human reviews for content reaching certain thresholds of popularity and implement age-based content

[182] DSA, *supra* note 114, art. 48.
[183] *Id*. art. 48(2).
[184] *Id*.
[185] Eur. Bd. For Digit. Services, *supra* note 24.
[186] *Id*. at 32.
[187] *Id*. at 33.
[188] *Id*.
[189] *Id*. at 34.
[190] *Id*.
[191] *Id*.

filters for minors.[192] Platforms also employ contextual text moderation measures, such as explaining to users why certain content or comments may not be posted due to content moderation filters.[193] Additionally, they collaborate with independent fact-checkers.[194]

3. Algorithmic Content Prioritization

The DSA recognizes the role of algorithms in prioritizing and presenting information to facilitate and optimize access to information for the recipients of the service.[195] It recognizes the significant impact that recommender systems can have on the ability of recipients to retrieve and interact with information online.[196] It also recognizes the important role that they play in the amplification of certain messages, the viral dissemination of information and the stimulation of online behavior.[197] It notes that very large online platforms should clearly present the parameters for such recommender systems in an easily comprehensible manner to ensure that the recipients understand how information is prioritized for them.[198] It emphasizes that the platforms should ensure that the recipients enjoy alternative options for the main parameters, including options that are not based on profiling the recipient.[199]

The DSA also notes that the advertising systems used by very large online platforms pose particular risks and require further public and regulatory supervision on account of their scale and ability to target and reach recipients of the service, based on their behavior within and outside of the platforms.[200] It notes that the very large online platforms should ensure public access to repositories of advertisements they display to facilitate supervision and research into emerging risks related to illegal hate speech, or manipulative techniques and disinformation

---

[192] *Id*.
[193] *Id*. at 34–5.
[194] *Id*. at 35.
[195] DSA, *supra* note 114, pmbl. ¶ 70.
[196] *Id*. pmbl. ¶ 70.
[197] *Id*.
[198] *Id*.
[199] *Id*. pmbl. ¶ 94.
[200] *Id.* pmbl. ¶ 68.

with a real and foreseeable negative impact on public health, public security, civil discourse, political participation, and equality.[201]

The platforms are obliged to collaborate with the national Digital Services Coordinators and provide them access to data necessary to assess the risks and possible harms brought about by the platforms' systems.[202] That data may be related to the accuracy, functioning, and testing of algorithmic systems for content moderation, recommender systems, advertising systems, data related to the processes and outputs of content moderation, or of internal complaint-handling systems within the meaning of the regulation.[203] The regulation provides a framework for compelling access to data from very large online platforms to vetted researchers especially for the identification of systemic risks.[204]

The first report on the application of the DSA indicates that platforms have adapted their recommender systems to demote harmful or potentially harmful content, such as illegal content, inauthentic content, content from repeat violators of terms and conditions, or disinformation.[205] They are also restricting the dissemination of potentially illegal content that has reached a certain level of popularity but has not yet undergone human review.[206] Additionally, platforms use algorithms to restrict content on high-risk topics or during critical periods, such as elections, or crises such as natural disasters or public health emergencies.[207] They also employ information banners or interstitials to limit the spread of content that occurs as a spillover from offline coordination or incidents from other platforms, such as viral trends.[208]

The platforms are designing their algorithmic systems to recommend content that is diverse, not narrow or repetitive, and they source content items for potential recommendation from reliable sources and trusted experts.[209] They also claim that their recommender systems optimize not only for view time but also for

[201] *Id.* pmbl. ¶ 95, art. 39(1).
[202] *Id.* pmbl. ¶ 97.
[203] *Id.* pmbl. ¶ 96, art. 40(4).
[204] *Id.* art. 40(4).
[205] Eur. Bd. For Digit. Services, *supra* note 24, § 4.4.
[206] *Id*. at 36.
[207] *Id*. at 36.
[208] *Id*. at 36.
[209] *Id*. at 36.

content quality.[210] Platforms give users the option to reset their recommendation profile to emulate a new user experience or allow them to turn off or opt out of recommender systems entirely.[211] They also allow their users to indicate that they are "not interested" in specific content, signaling to the system to reduce exposure to similar content.[212] Furthermore, platforms permit users to select keywords for content display or filtering.[213]

4. Penalties

The act authorizes the member states to enact penalties in relation to the infringements of the regulation by providers of intermediary services under their jurisdiction. These penalties should be effective, proportionate, but also serious enough to dissuade from violating the Act.[214] The Member States should ensure that the maximum amount of penalties imposed for a failure to comply with the obligations laid down in the Act should not exceed 6% of the annual income or turnover of the provider of intermediary services concerned.[215] As regards the periodic penalty payments, they should not exceed 5% of the average daily turnover of the provider of intermediary services concerned in the preceding financial year per day, calculated from the date specified in the decision concerned.[216] For the very large platforms, the Commission may impose fines of 6% of their total turnover in the preceding financial year, if it finds that they have intentionally or negligently failed to meet their obligations under the Act.[217]

5. Institutions for Supervision and Transparency

The DSA creates a system aiming to control the moderation procedures that already exist within the major social media networks.[218] These moderation procedures were enhanced by these platforms following the Code of Conduct on

[210] *Id*. at 36.
[211] *Id*. at 37.
[212] *Id*. at 37.
[213] *Id*. at 37.
[214] DSA, *supra* note 114, art. 52.
[215] *Id*. art. 52(3).
[216] *Id*. art. 52(4).
[217] DSA, art. 74(1).
[218] *See supra* note 85 and accompanying text.

countering hate speech agreed between them and the EU in 2016.[219] This system consists of creating national bodies, the Digital Services Coordinators within the member states, and national codes of conduct. These national bodies have the mission to ensure the effective and consistent application and enforcement of the regulation across the Union.[220] The Act also creates a European Board for Digital Services, which will be an independent advisory group for the national Digital Services Coordinators.[221] The National Digital Services Coordinators, the European Board for Digital Services, and the European Commission all collaborate toward enforcing the provisions of the Act upon the major social media platforms.[222]

The National Digital Services Coordinators should be authorities independent from the government and private bodies, and free from any external influence.[223] They have significant powers to impose fines in accordance with the regulation for failure to comply with its clauses, as well as periodic penalty payments to ensure that infringements are terminated.[224] They also can adopt interim measures to make sure that serious harm does not arise.[225] The Digital Services Coordinators also have the authority to request that the judicial authorities of Member States impose a temporary restriction on access to the service for its recipients when a violation of the DSA persists and results in serious harm or involves a serious criminal offence that threatens the life or safety of persons.[226]

The DSA creates also an independent advisory group for Digital Services Coordinators, the European Board for Digital Services.[227] It will be composed by high-level officials of the National Digital Services Coordinators.[228] It will, among others, support and promote the development and implementation of European standards, guidelines, reports, templates and code of conducts, and it will

[219] Eur. Comm'n, *Code of Conduct on Countering Illegal Hate Speech Online* (2016), https://ec.europa.eu/newsroom/document.cfm?doc_id=42985 [https://perma.cc/6Q3D-AVLD].
[220] DSA, *supra* note 114, art. 49.
[221] *Id*. art. 61.
[222] *Id*. arts. 62, 63, 64.
[223] *Id.* art. 50.
[224] *Id.* art. 51(2).
[225] *Id*. art. 51(2)(e).
[226] *Id*. art. 51(3)(b).
[227] *Id*. art. 61.
[228] *Id*. art. 62(1).

contribute to the identification of emerging issues, with regard to matters covered by the Regulation.[229] The Board shall be chaired by the European Commission.[230]

The Act contains several clauses related to enhancing transparency in the moderation activities of several platforms. The DSA has special clauses for the supervision, investigation, enforcement and monitoring in respect of very large online platforms. The Commission, or the Board either on their own initiative or upon request of at least three Digital Services Coordinators, may recommend to a national Digital Services Coordinator to investigate a suspected infringement.[231] The Commission has also the authority to monitor the effective implementation and compliance with the Act.[232] The DSA states that where the infringement of a provision that applies to very large online platforms is not effectively addressed, the Commission may on its own initiative or upon advice of the Board investigate the infringement concerned and the measures taken.[233] It should be able to issue decisions finding an infringement and imposing sanctions in respect of very large online platforms. The Act gives the Commission strong investigative and enforcement powers to enforce and monitor several rules.[234] The Commission may itself adopt a non-compliance decision in relation to very large online platforms.[235] It may request any information it needs in this respect and even conduct onsite investigations.[236]

All these institutions will play an important role in defining the standards of protection of online speech. Although the role of all these independent authorities is very important, Courts will have the last word in Europe about whether social media users' freedom will be protected adequately or not. Courts will apply existing constitutional standards for the protection of hate speech within each one of the member-states. They will also apply the definitions of hate speech that are acceptable within each one of the EU Member States.[237] The contribution of courts is already very important within the EU member-states legal systems. Judicial and

[229] *Id*. art. 61(2).
[230] *Id*. art. 62(2).
[231] *Id*. art. 14 (1).
[232] *Id*. art. 57.
[233] *Id.* pmbl. ¶ 138.
[234] *Id.* pmbl. ¶ 140.
[235] *Id.* art. 73.
[236] *Id.* pmbl. ¶ 141.
[237] *See infra* Part IV.A.

extra-judicial mechanisms will be applying the standards of protection in relation to hate speech, which vary across the European states. The role of the standards articulated by the European Court of Human Rights in this area will also play an important role.[238]

6. Reporting

The Act creates transparency reporting obligations for online platforms. They are obliged to publish detailed reports on the content moderation they engaged in, at least once a year.[239] These reports should include the number of orders they received from member states' authorities, the number of notices they received in relation to allegations of illegal content and any action they took.[240] They should also include the content moderation that they engaged in at their own initiative and the number and type of measures they took that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information.[241] They should also include the number of complaints received through their internal complaint-handling system.[242] The reports should also include information on the number of disputes submitted to the out-of-court dispute settlement bodies certified by the National Digital Services coordinators, the outcomes of the dispute settlement, and the average time needed for completing the dispute settlement procedures.[243] The reports should also include the number of suspensions to users' accounts imposed explaining whether these suspensions were enacted for the provision of manifestly illegal content, the submission of manifestly unfounded notices and the submission of manifestly unfounded complaints.[244] They should also report the use that they have made of AI for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards they applied.[245] The platforms are also obliged

[238] *Id.*
[239] DSA, *supra* note 114, art. 15(1).
[240] *Id.* art. 15(1)(a)–(b).
[241] *Id.* art. 15(1)(c).
[242] *Id.* art. 15(1)(d).
[243] *Id.* art. 24(1)(a).
[244] *Id.* art. 24(1)(b).
[245] *Id.* art. 15(1)(c), (e).

to communicate to the Digital Services Coordinator their establishment of the information on the numbers of their active users.[246]

7. Compliance Officers

The DSA creates an obligation for the very large online platforms to appoint compliance officers who should monitor the compliance with this Regulation within the platform's organization.[247] These compliance officers are involved in all issues that relate to their obligations under the regulation. The compliance officers must be able to perform their tasks in an independent manner.[248] The platforms have an obligation to support them in the performance of their tasks and provide them with the resources necessary to adequately carry out those tasks.[249]

8. Civil Society

The regulatory system created by the DSA gives an important role to civil society organizations active in the area of eliminating hate. It complements the online moderation practiced by major platforms with "trusted flagger organisations"[250] which will notify the platforms about suspect content. The national Digital Services Coordinators are tasked with the mission to award the status of "trusted flagger" to organizations which have demonstrated that they have particular expertise and competence in "detecting," "identifying," and "notifying" illegal content, if they represent collective interests and are independent from any online platform and if they carry out their activities for the purposes of submitting notices in a timely, diligent, and objective manner.[251]

The DSA foresees a system of "independent" auditing for very large online platforms.[252] This system of independent auditing should ensure that large online platforms abide by their obligations as regards the notice and action mechanisms, their obligation to provide reasons when they decide to remove or disable access to items of information, the measures they take against misuse and illegal behavior

[246] *Id.* art. 24(2)–(3).
[247] *Id.* pmbl. ¶ 99, art. 41(1)–(3).
[248] *Id.* pmbl. ¶ 92.
[249] *Id.* art. 41(5)–(7).
[250] *Id.* pmbl. ¶ 87.
[251] *Id.* pmbl. ¶ 61, art. 22(2).
[252] *Id.* pmbl. ¶ 92.

more generally, and the dispute settlement systems that they implement.[253] The auditors should write a report to be transmitted to the Digital Services Coordinator where the platform is established and the Board, together with the risk assessment and the mitigations measures and the platform's plans for addressing the audit's recommendation.[254] The auditors should include in their report their opinion on whether the platforms are abiding by the regulation.[255] The platforms must collaborate with the auditors and make available to them all relevant data.[256]

9. Transnational Impact

The DSA foresees that it applies to providers of intermediary services irrespective of their place of establishment or residence, "in so far as they offer services in the Union, as evidenced by a substantial connection to the Union."[257] This means that it covers the services that social media platforms are offering to anyone within the EU, even if the companies themselves are based outside of the EU. But even beyond that, companies have strong operational and reputational incentives to adapt to the DSA even if they do not have a legal obligation to do so.[258] Insofar as the major companies are transnational and must therefore follow European rules as well as American law, several companies have already modified their operation practices to abide by the EU's code of conduct countering illegal hate speech online.[259] Companies tend to modify their behavior so as to meet the most stringent legal regimes in order to be able to offer their services everywhere. Navigating across different legal regimes poses several logistic challenges, while following one single set of rules is operationally easier.[260] By engaging in regional regulation of online speech, the EU may be transforming itself into a global regulator. This transformation into a "global regulatory hegemon" helps the EU

[253] *Id.* art. 37(1).
[254] *Id.* pmbl. ¶ 93.
[255] *Id.*
[256] *Id.* art. 37(2).
[257] *Id.* pmbl. ¶ 7.
[258] *See also infra*, Part IV.A.
[259] *Supra* Part II.A.
[260] Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, Hoover Working Grp. on Nat'l Sec., Tech., and L., Aegis Series Paper No. 1902 (Jan. 29, 2019), 8, https://www.lawfaremedia.org/article/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech [https://perma.cc/J52E-QE3M].

obtain greater legitimacy for its rules and enhances its soft power.[261] The EU can thus perpetuate the ideology that an extensive regulatory system is needed to preserve global public goods.[262] This influence also affirms the EU's global standing during recent crises of constitutional democracy around the world.

In the area of data privacy, the EU has transformed itself into a global regulator through the adequacy decisions it requires for data transfers outside of the EU. The General Data Protection Regulation (GDPR) enacted by the EU has clauses which affect the legal regime of states beyond its territory.[263] In order for data transfers of anyone within the EU to take place outside the EU, the EU is requiring a decision by a Data Protection Authority of any Member state which affirms that the third country is affording an "adequate" level of protection of data privacy.[264] These "adequacy" decisions mean that the EU standards of privacy protection apply extraterritorially. Third countries must afford standards of protection considered by any EU Data Protection Authority at least adequate to the GDPR.[265] They must enact legal regimes that approximate EU privacy standards. There is a longstanding clash of values between the EU and the US in the area of enforcement of EU privacy law. Several agreements between the US and EU related to transatlantic privacy transfers have been invalidated by the Court of Justice of the European Union.[266] The latest agreement in this area was reached

[261] Bradford, *supra* note 29, at 24.

[262] *Id*.

[263] *See* GDPR, *supra* note 113, arts. 45–47 (detailing the requirement that other countries participating in data transfers with people in the EU must meet minimum protections).

[264] *Id*. art. 45.

[265] *Id.* arts. 45–47.

[266] For other examples of international values clashing in privacy law, *see, e.g.*, Paul M. Schwartz, *The EU-U.S. Privacy Collision: A Turn to Institutions and Procedures*, 126 Harv. L. Rev. 1966, 1978–79 (2013) (addressing the United States' tendency to let business experiment with data processing, while the EU prioritizes consumer protection); Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and European Union*, 102 Cal. L. Rev. 877, 879–80 (2014) (noting the EU's broader definition and thus regulation of personally identifiable information compared to that of the United States); Ioanna Tourkochoriti, *The Transatlantic Flow of Data and the National Security Exception in the European Data Privacy Regulation: In Search for Legal Protection Against Surveillance*, 36 U. Pa. J. Int'l. L. 459, 467–69 (2014) (asserting that the United States permits the processing of personal data by default, while the opposite is true in the EU); Ioanna Tourkochoriti, *The Snowden Revelations, the Transatlantic Trade and Investment Partnership and the Divide Between U.S.-E.U. in Data Privacy Protection*, 36 U. Ark. Little Rock L. Rev. 161, 163–64 (2014) (addressing the United States' and EU's different approaches to enforcing data protection laws).

in the spring of 2022, under the Biden presidency.[267] Following the recent policy changes in several major social media companies, it is very likely that some clash of values will emerge in the area of regulating hate speech online, too.[268] The EU's Code of Conduct asked the platforms to sign on to a global impact code, but some major platforms have decided not to abide by EU rules where they do not have a legal obligation to do so, as analyzed below.[269] The uncertainty about the enforcement of the Act has also affected EU-US relations. Due to concerns about the transatlantic enforcement of the DSA, Marco Rubio announced that the US government will refuse visas to officials of EU member states that are limiting Americans' speech.[270]

B. EU Member States' Legislation

The DSA now supersedes national legislation existing in several EU Member states on regulating hate speech and fake news online, e.g. France and Germany.[271] Germany enacted the Network Enforcement Act in 2017.[272] This legislation aimed to enable the enforcement of criminal code provisions and other regulations relating to online social media platforms.[273] It aimed to enforce clauses related to the dissemination of propaganda material of unconstitutional organization, the preparation of a violent offence endangering the state, and the encouragement of a serious violent offence endangering the state.[274] Under the DSA, the Network Enforcement Act applies only to the extent that it is compatible with it.

France has legislated an even more restrictive framework related to hate speech online which now complements the DSA. Legislation enacted already in

---

[267] For the latest agreement, see European Commission Press Release IP/22/2087, The Commission, European Commission and United States Joint Statement on Trans-Atlantic Data Privacy Framework (Mar. 24, 2022).

[268] *See infra* Part IV.A.

[269] *Id.*

[270] Agence France Press in Washington, *US Will Refuse Visas to Foreign Officials Who Block Americans' Social Media Posts*, Guardian (May 28, 2025), https://www.theguardian.com/us-news/2025/may/28/us-refuse-visas-foreign-officials-social-media [https://perma.cc/PC6C-SNSX].

[271] Jörn Reinhardt: *"Fake News", "Infox", Trollfabriken: Über den Umgang mit Desinformationen in den sozialen Medien*, 225/226 Vorgänge 97-108 (2019) (Ger.); Claudia E. Haupt, *Regulating Speech Online: Free Speech Values in Constitutional Frames*, 99 Wash. U. L. Rev. 751 (2021).

[272] Gesetz zur Verbesserung der Rechtsdurchsetzung in den sozialen Netzwerken [NetzDG] [Network Enforcement Act], Sept. 1, 2017 BGBl. I at 3352 (Ger).

[273] Haupt, *supra* note 271, at 761.

[274] *Id*. at n. 53.

2004 related to protecting online speech,[275] and it has been amended several times since. One of its most recent amendments, in 2020, related to adapting the speech offenses foreseen in the main statute which relates to freedom of the press in France enacted in 1881.[276] Proclaiming the protection of "the freedom to communicate online," the amended law foresees that this freedom can be limited if this is necessary to respect "human dignity," the freedoms and property of others, pluralism, or the public order, among others.[277] The law defines hate speech as speech which offends a person or a group "based on their origin or membership or non-membership in an ethnicity, a nation, a race or a religion, their sex, their sexual orientation, their gender identity or disability."[278] In France, illegal hate speech also covers the comments of those who incite discrimination, hatred, or violence against a person based on one of the features.[279] It further covers the denial of crimes against humanity as defined by French law, challenging the historical occurrence of genocide and other crimes against humanity, as well as challenging the enslavement or exploitation due to enslavement, and challenging the accuracy a war crime condemnation pronounced by a French or an international court.[280]

The legislation aiming to combat hate speech online also foresees that the platforms maintain procedures and human and technological means to cooperate with judicial and administrative authorities when the latter ask them to identify users who engage in illegal hate speech. They must provide elements of identification to public authorities and they must preserve content that has

[275] *See* Loi 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (1) [Law No. 2004-575 of June 21, 2004 for Confidence in the Digital Economy], Journal Officiel de la République Française [J.O.] [Official Gazette of France], June 22, 2004 (setting out narrow situations in which freedom of speech may be limited online).

[276] *See* Loi 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (1) [Law 2020-766 of June 24th 2020 Aiming to Fight Hate Speech on the Internet], Journal Officiel de la République Française [J.O.] [Official Gazette of France], June 25, 2020 (incorporating provisions relating to registered users of online platforms); Conseil constitutionnel [CC] [Constitutional Court] Décision n° 2020-801 DC, June 18, 2020 (Fr.) (imposing liability on online platform operators).

[277] Loi 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique (1) art. 1.

[278] *Id.* art. 1-1; *see also* Loi du 29 juillet 1881 sur la liberté de la presse [Law of July 29, 1881 on Freedom Of The Press] arts. 32–33 Journal Officiel de la République Française [J.O.] [Official Gazette of France], July 20, 1881.

[279] Loi du 29 juillet 1881 sur la liberté de la presse arts. 23–24.

[280] Loi du 29 juillet 1881 sur la liberté de la presse, art. 24(a).

been signaled as illegal hate speech, even when they have withdrawn from public view.[281]

France has also enacted legislation in relation to misinformation. Based on this law, online platforms must take measures to combat fake news that may threaten the public order or alter the sincerity of electoral processes up to three months leading up to elections.[282] The legislation authorizes judges to issue preliminary injunctions via an urgent applications process ordering the removal of allegations of facts during electoral campaigns, which are "inexact or deceiving" and likely to alter the sincerity of the voting process.[283] The judges must decide within 48 hours.[284] Furthermore, according to the same law, the French *Conseil Supérieur de l'Audiovisuel* (CSA), which has now been succeeded by the French Regulatory Authority for Audiovisual and Digital Communication (*Autorité de régulation de la communication audiovisuelle et numérique – ARCOM*) can unilaterally terminate a contract with a broadcasting corporation when the latter transmits "false information."[285] The law applies to corporations broadcasting in France that are controlled by foreign states.[286] Legislation also obliges online platforms to take measures against the transmission of "false information" which can disturb the public order.[287] The law obliges platforms to take measures to allow users to signal anything they consider fake news.[288] The platforms should also communicate to the public their activities toward ensuring the transparency of the algorithms they use, their action towards combating misinformation.[289]

Examining the constitutionality of legislation, the French *Conseil Constitutionnel* held that the law was necessary to prevent the risk of citizens

[281] Loi 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (1) [Law 2020-766 of June 24, 2020 Aiming to Fight Hate Speech on the Internet] art. 2(I), JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], June 25, 2020.

[282] Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1) [Law 2018-1202 of Dec. 22 2018 on the Fight Against the Manipulation of Information], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], Dec. 23, 2018.

[283] Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1) art. 1.

[284] *Id.*

[285] *Id.* art. 8.

[286] *Id.* art. 10.

[287] *Id.* art. 11.

[288] *Id.* art. 11(I).

[289] *Id.*

being manipulated in exercising their vote by the mass dissemination of misinformation.[290] The *Conseil Constitutionnel* also upheld the possibility the law gave to urgent applications judges's abilities to order the removal of certain incorrect or misleading information from the internet.[291] The judges were asked to evaluate whether the time frame of 48 hours was appropriate given the risks in intervening within the political debate ahead of a future election.[292] According to the *Conseil*, the legislation concerns cases where the "incorrect" or "misleading nature" of a publication "is apparent."[293] The legislation does not capture opinions, parodies, partial inaccuracies, or simple exaggerations.[294] The legislation thus is compatible with the right to a fair trial and the "constitutional value" of properly administrating of justice and the guaranteeing rights.[295] The same *Conseil* also held that the authorities given to the CSA—those authorities now held by ARCOM—are also compatible with freedom of expression, given that those concerned possess judicial avenues to ensure their rights are protected against the misuse of power by a public authority.[296]

France has enacted several other statutes to complement the application of the DSA. The law to secure and regulate the digital space, known as the SREN law, designates the national French authority responsible for monitoring its application.[297] The law designates ARCOM as the digital services coordinator in France.[298] The law introduces a new supplementary penalty which allows for the suspension of social media accounts used to commit certain offenses such as harassment, defamation, or invasion of privacy, for a maximum period of six months to a year.[299] Failure to comply with these obligations regarding the design and organization of online platforms is a criminal offense, as are the violations of

[290] Conseil constitutionnel [CC] [Constitutional Court] Décision n° 2018-773 DC, Dec. 20, 2018, ¶ 18 (Fr.).
[291] *Id.* at ¶ 23.
[292] *Id.*
[293] *Id.*
[294] *Id.* at ¶ 21.
[295] *Id.* at ¶ 26.
[296] *Id.* at ¶¶ 52, 64.
[297] LOI n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique [Law 2024-449 of May 21, 2024 to secure and regulate the digital space], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], May 22, 2024.
[298] *Id.* art. 51.
[299] *Id.* art. 16.

the DSA clauses which oblige very large platforms to appoint compliance officers, to create repositories with advertisements they make available, and to provide access to researchers who investigate systemic risks that may occur within the platforms.[300]

ARCOM's authorities have been increased to also regulate foreign audiovisual services available in France if they transmit online. It requires online public communications providers and hosts to remove or cease broadcasting content from entitles that are covered by the DSA within 72 hours upon formal notice from ARCOM.[301] It also foresees penalty enhancements for distributing AI-generated visual or audio content depicting a person without their consent. Sharing deepfakes carries a penalty of two years of imprisonment maximum and a fine of 45,000 euros.[302] It obliges major online platforms to adopt charters for the monitoring and support of content moderators, including measures for training, psychological support, and well-being at work.[303]

The SREN law is also important because it reinforces digital citizenship education. It makes it compulsory for pupils to obtain a digital skills certificate at the end of their first and last years of secondary school.[304] This digital citizenship training includes awareness of the risks associated with artificial intelligence and the fight against disinformation.[305] It also enhances the protection of minors from accessing online pornographic content by blocking, delisting and imposing finds for sites that do not verify user age.[306]

III
US: Reluctance to Regulate

The DSA is likely to lead to a transatlantic clash in the standards of protecting freedom of expression between the US and the EU. Federal legislation in the US protects the platforms from liability for content posted by others.[307] It also exempts

[300] *Id.* art. 52.
[301] *Id.* art. 14.
[302] *Id.* art. 15.
[303] *Id.* art. 25.
[304] *Id.* arts. 7, 8.
[305] *Id*. art. 7.
[306] *Id*.
[307] 47 U.S.C. § 230(c)(1) ("No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider").

them from civil liability for any action they take "in good faith" to moderate content posted by others which they consider to be "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[308] During his first term, President Trump issued an executive order according to which, "it is the policy of the United States to ensure" that this clause "is not distorted to provide liability protection" when the platforms "engage in deceptive or pretextual actions … to stifle viewpoints with which they disagree."[309] When the President himself was excluded following the events at the Capitol in January 6th 2021, he sued Facebook, Twitter, and Google/YouTube, alleging viewpoint discrimination and due process-type violations, such as lack of clear standards/guidelines, and lack of notice on the relevant terms of service.[310] The lawsuits argued that the platforms are state actors for First Amendment purposes and asked the courts to find Section 230 unconstitutional.[311] The cases against the platforms were dismissed for failing to plausibly allege a First Amendment claim.[312] Several attempts to enact federal legislation to regulate the platforms and to protect some due process rights for their users did not succeed.[313]

[308] *Id.* § 230(c)(2).

[309] Exec. Order No. 13925, 85 Fed. Reg. 34079 (2020).

[310] *See* Class Action Complaint ¶59, Trump v. Facebook, Inc., No. 1:21-cv-22440 (S.D. Fla. July 7, 2021); Class Action Complaint ¶¶37–38, Trump v. Twitter, Inc., No. 1:21-cv-22441 (S.D. Fla. July 7, 2021); Class Action Complaint ¶8, Trump v. YouTube, LLC., No. 1:21-cv-22445 (S.D. Fla. July 7, 2021). For an analysis of the cases *see* Nunziato, *supra* note 74, at 1269–71.

[311] Class Action Complaint ¶151, Trump v. Facebook, Inc.; *see* Nunziato, *supra* note 74, at 1269–70.

[312] *See, e.g.*, Trump v. Twitter, Inc., 602 F.Supp.3d 1213, 1225 (N.D. Cal. May 6, 2022). The court also dismissed Trump's request for declaratory judgment that CDA 230 is unconstitutional. *Id.*; *see also* Nunziato, *supra* note 74, at 1271 n.97.

[313] Then-Senator Marco Rubio's proposal DISCOURSE Act aimed to protect due process rights for those excluded and to create civil liability for moderating content in a way that advantages certain viewpoints by amending § 230(c)(2). Senator Bill Hagerty introduced the 21st Century FREE Speech Act, which called social media platforms "common carrier technology companies" and abrogated § 230 immunity when they failed to comply with the Act. Senators John Thune and Brian Schatz introduced the Platform Accountability and Consumer Transparency (PACT) Act, which would create due process-type requirements but no liability for viewpoint or speaker discrimination. Senator Roger Wicker introduced the PRO-SPEECH Act to prohibit discrimination in the platforms and to authorize the FDC to regulate unfair methods of competition. A few more senators introduced a bill to repeal § 230 entirely. *See* Nunziato, *supra* note 74, at 1277–81.

A. Texas's And Florida's Attempts To Regulate

Following the moderation practice adopted by the platforms in conformity with the deontology code elaborated by the EU, Texas and Florida in the US enacted legislation to prevent large social media platforms from engaging in viewpoint discrimination.[314] To a great extent, the motivation behind this legislation was the same as the one behind the DSA: it was important to ensure access to social media platforms which now define the public sphere and to limit their ability to remove users.[315] There are important differences between the DSA and how these statutes go about it. The legislation enacted in Florida and Texas aimed to protect conservative viewpoints and ideas from being removed from online social media.[316] In this respect, the legislation was based on a problematic motivation, as the Supreme Court found.[317] The DSA aims to protect all kinds of users from being removed from platforms. Furthermore, unlike the DSA, the legislation in Florida and Texas would consider hate speech to be viewpoint discrimination and would limit the ability of the platforms to engage in content moderation in Texas and Florida.[318]

The question that emerges is whether the platforms exercise editorial judgment which is protected under the First Amendment, or whether the common carrier doctrine should apply to them, which means that they are not allowed to engage in viewpoint discrimination. The prohibition to discriminate means that the content moderation the platforms are engaging in thanks to the EU deontology code and the limits they impose upon hate speech and incitement to hatred are not acceptable under the legislation existing in Texas and Florida. If content moderation is to be considered as part of the platforms' editorial discretion, *Miami Herald v. Tornillo* dictates that it is protected under the First Amendment.[319] Following a split in circuit courts on the matter, the Supreme Court, in *Moody v. NetChoice*, made some important clarifications as to the meaning of the First Amendment for online social media platforms.[320] It also vacated and remanded the

[314] *See* FLA. STAT. § 501.2041 (2025), TEX. BUS. & COM. CODE ANN. § 120.001.
[315] *See infra* notes 323–328, 349–357, and accompanying text.
[316] *See id.*
[317] Moody v. Netchoice, LLC., 144 S. Ct. 2383, 2407 (2024).
[318] *See infra* notes 323–328, 349–357, and accompanying text.
[319] 418 U.S. 241 (1974).
[320] 603 U.S. 707 (2024).

case to the Circuit Courts because, in its opinion, these courts did not perform the facial analysis they should have performed as to all the possible applications of the statutes.[321] The Supreme Court, by affirming that the freedom of the platforms is protected by the First Amendment, appears to be reluctant to accept the possibility for government regulation of the platforms. The decision implies that according to the First Amendment, government may not regulate these platforms in either direction. Jack Balkin described Texas's and Florida's attempts to regulate the platforms as attempts to de-constitutionalize the matter—that is, to get courts to accept legal theories that turn conflicts over speech claims into non-constitutional questions of statutory or administrative law or technological design.[322]

Florida enacted SB 7072 which prevents social media platforms from "unfairly censoring," "shadow banning," "deplatforming," or applying post-prioritisation algorithms to political candidates, users, or residents.[323] It defined "shadow ban" as "actions by social media platforms which limit or eliminate the exposure of a user or content or material posted by a user to other users of the social media platform," including acts "which are not readily apparent to a user."[324] The law was inspired by the common carrier doctrine.[325] In relation to candidates for office in particular, the law foresaw that social media platforms may not deplatform them, beginning on the date of qualification and ending on the date of election or the date the candidate ceases to be a candidate.[326] The text's focus on political candidates was due to concerns that Big Tech engaged in viewpoint discrimination against conservatives.[327] Social media platforms may remove users only after having given written notice within seven days with a "thorough rationale" for their decision and they must explain how they used any algorithms to identify or flag the user's content or material as objectionable.[328]

---

[321] See *infra* Part III.B.

[322] Jack Balkin, *Moody v. Netchoice: The Supreme Court Meets the Free Speech Triangle*, SUP. CT REV., 127, 137 (2025).

[323] FLA. STAT. § 501.2041(1)(g) (2023). "Shadow-banning" is defined as any action that limits exposure to content. *Id.* § 501.2041(1)(f).

[324] *Id.* § 510.2041(1)(e).

[325] *See also* Nunziato, *supra* note 3, at 367.

[326] Fla. S.B. 7072, § 2 (2021) (amending FLA. STAT. § 106.072).

[327] Upon elaboration of the law, Florida's Governor DeSantis emphasized the need to protect against viewpoint discrimination for conservative voices, Calvert, *supra* note 3, at 396.

[328] FLA. STAT. § 501.2041(2)(d)–(3)(d) (2023).

The legislation also contained important elements towards protecting users from being unfairly excluded from the platforms. The law further created a private right of action and foresaw statutory damages of $100,000 and punitive damages against the platform.[329] Platforms that willfully deplatform political candidates would face fines of $250,000 per day for statewide candidates and $25,000 per day for other candidates.[330] The law created a further obligation for social media platforms to provide notice annually to users on the post-prioritisation algorithms for content and material posted by or about a user who is known to be a candidate for office.[331] The law also provided that a social media platform may not take any action to censor, deplatform, or shadow ban a journalistic enterprise based on the content of its publication or broadcast.[332]

The Eleventh Circuit found that the law violated the First Amendment rights of social media platforms.[333] For the court, the social media platforms "express themselves for better or worse through their content-moderation decisions."[334] "When a platform selectively removes what it perceives to be incendiary political rhetoric … it conveys a message and thereby engages in 'speech' within the meaning of the First Amendment."[335] Content moderation is editorial judgment that is inherently expressive.[336] The platforms express themselves by announcing their community standards and by enforcing them. They exercise editorial judgment by removing posts that violate their terms of service or community standards, e.g. hate speech or violent content.[337] They also exercise editorial judgment "by choosing how to prioritize and display posts – effectively selecting which users' speech the viewer will see and in what order."[338] According to the court, these practices are "curating" users' posts into collections of content that they then disseminate to others, an activity which is close to the editorial control and judgment of the press protected by *Miami Herald Publishing Co. v. Tornillo*.[339]

[329] *Id*. § 501.2041(3)(d).
[330] Fla. Stat. § 106.072(3) (2024).
[331] Fla. Stat. § 501.2041(3)(d) (2023).
[332] *Id.* § 501.2041(1)(j).
[333] NetChoice, LLC v. Attorney General, Florida, 34 F.4th 1196, 1203 (11th Cir. 2022).
[334] *Id.*
[335] *Id.*
[336] *Id.* at 1227.
[337] *Id.* at 1210.
[338] *Id.*
[339] Miami Herald Pub. Co. v. Tornillo, 418 U.S. 241, 258 (1974).

Miami Herald held that a Florida statute giving a right to reply to those facing "criticism" violates the First Amendment because it interferes with the newspaper's exercise of editorial control and judgment.[340] Forcing the paper to print what otherwise it would not, intrudes into editorial discretion.[341]

For the Eleventh Circuit, "a private entity's decisions about whether, to what extent and in what manner to disseminate third-party-created content to the public are editorial judgments protected by the First Amendment."[342] For the court, the common carrier doctrine is not relevant to social media platforms because platforms require users to accept their terms of service and to abide by their community standards.[343] They also make individualised content- and viewpoint-based decisions about whether to publish particular messages or users.[344] The court also notes that Congress has distinguished internet companies from common carriers because the Telecommunications Act of 1996 explicitly differentiates "interactive computer services" from "common carriers or telecommunications services."[345]

The court applied strict scrutiny and did not find any compelling interest that could justify restrictions on editorial judgment.[346] It found that there is no "governmental interest in levelling the expressive playing field. Nor is there a substantial governmental interest in enabling users – who, remember have no vested right to a social-media account – to say whatever they want on privately owned platforms that would prefer to remove their posts."[347] "Preventing 'unfairness' to certain users or points of view isn't a substantial governmental interest; rather, private actors have a First Amendment right to be 'unfair' – which is to say, a right to have and express their own points of view."[348] The decision means that platforms should be free to regulate how their users should behave and to enforce this regulation by engaging in content moderation or not.

---

[340] *Id.*

[341] *Id.*

[342] *NetChoice*, 34 F.4th at 1212.

[343] *Id.* at 1220.

[344] *Id.*

[345] *Id.* at 1221, citing 47 U.S.C. § 223(e)(6) ("Nothing in this section shall be construed to tread interactive computer services as common carriers or telecommunications carriers.").

[346] *Id.* at 1228.

[347] *Id.*

[348] *Id.*

Texas enacted HB 20 which prohibited large social media platforms from censoring speech based on the viewpoint of its speaker.[349] The Texas legislature in enacting HB 20 found that the platforms "function as common carriers, [they] are affected with a public interest, are central public forums for public debate and have enjoyed governmental support in the United States."[350] It also found that "social media platforms with the largest number of users are common carriers by virtue of their market dominance."[351] Texas Governor Greg Abbott emphasized the role social media play in forming "the public square" to justify the need to "protect Texans from being wrongfully censored."[352] The legislation foresaw that a social media platform "may not sensor a user, a user's expression, or a user's ability to receive the expression of another person based on the viewpoint of the user or another person, the viewpoint represented in the user's expression or another person's expression or a user's geographic location in this state or any part of this state."[353] Although the DSA aims to provide guidance as to the content moderation of hate speech, HB 20 aimed to prohibit content moderation of hate speech altogether. Nevertheless, concerning users' access rights to the platforms, the law contained several clauses which are very close to the content of the DSA. The law obliged the platforms to provide written notice to the users when they decide to remove content and to explain the reason why the removal took place.[354] It also obliged the platforms to create a complaint system in relation to the decision to remove the content.[355] It contained disclosure requirements as to the moderation policy, obliging platforms to publish a biannual transparency report in relation to their moderation activities.[356] Users and the state attorney could seek injunctive relief for violations of viewpoint neutrality.[357]

Unlike the Eleventh Circuit, the Fifth held that the common carrier doctrine is relevant and that the legislation is compatible with the First Amendment.[358] The

[349] Tex. Bus. & Com. Code Ann. §§ 120.001(1), 120.002(b) (West 2023).
[350] H.B. 20, 87th Leg., 2d Spec. Sess. § 1(3) (Tex. 2021).
[351] *Id.* § 1(4).
[352] Nunziato, *supra* note 3, at 373.
[353] Tex. Civ. Prac. & Rem. Code Ann. § 143A.002(a) (West 2021).
[354] *Id.* § 120.103(b)(1).
[355] *Id.* §§ 120.103(b)(2), 120.104.
[356] *Id.* § 120.053.
[357] *Id.* §§ 143A.007–008.
[358] NetChoice L.L.C. v. Paxton, 49 F.4th 439, 448 (5th Cir. 2022).

decision focuses on the law as protecting other people's speech and regulating the platforms' conduct, not their speech.[359] The court ruled that the State of Texas can regulate conduct in a way that requires private entities to host, transmit or otherwise facilitate speech.[360] The court applied the common carrier doctrine, which vests States with the power to impose non-discrimination obligations on communication and transportation providers that serve the public.[361] The doctrine vests the Texas Legislature with the power to prevent the platforms from discriminating against Texas users.[362] The court held that the platforms are common carriers because they are holding out their communications medium for the public to use on equal terms and because they are understood as having the social and economic role to facilitate other people's speech.[363]

For the same court even if the legislation burdens the platforms' First Amendment rights, it does so in a content-neutral way and is thus subject to intermediate scrutiny.[364] The Fifth Circuit disagrees with the Eleventh Circuit that the Supreme Court has recognized 'editorial discretion' as an independent category of First-Amendment-protected expression.[365] It also disagrees with the conclusion that the platforms' censorship is akin to "editorial judgment" that has been mentioned in Supreme Court doctrine. For the Fifth Circuit, the platforms do not exercise editorial control because they use algorithms to screen out certain obscene and spam-related content.[366] They also disclaim any reputational or legal responsibility for the content they host.[367] They merely engage in viewpoint-based censorship with respect to expression they already have disseminated.[368] The court cites in this respect 47 U.S.C. § 230 which provides that platforms "shall [not] be treated as the publisher or speaker" of content developed by other users.[369]

[359] *Id.* at 455.
[360] *Id.*
[361] *Id.* at 469.
[362] *Id.*
[363] *Id.* at 479–80.
[364] *Id.* at 480.
[365] *Id.* at 490.
[366] *Id.* at 459–60.
[367] *Id.* at 464.
[368] *Id.*
[369] *Id.* at 480.

B. The Supreme Court's Delimitation Of These Efforts

1. Content Moderation and Editorial Discretion

The Supreme Court examined these cases together on appeal. In its recent ruling *Moody v. NetChoice*, a 9-0 opinion written by Justice Kagan, the Supreme Court vacated the judgments of the Courts of Appeals for the Fifth and Eleventh Circuits and remanded the cases for further proceedings.[370] In its opinion, neither Court considered the scope of the application of the laws and weighed the unconstitutional applications as against the constitutional ones.[371] The courts merely concentrated on whether a state law can regulate the content-moderation practices used in Facebook's News Feed.[372] They treated these cases as applied claims and not as facial ones.[373] The Circuit Courts did not assess the scope of the statutes and in particular what activities the laws prohibit or regulate.[374] They did not examine whether the laws affect other services the platforms have like direct messaging or event management.

Further, the Supreme Court articulated some criteria for the lower Courts that are relevant to the transatlantic difference in the protection of freedom of expression and the scope of the DSA. Its analysis is leaning closer to the Eleventh Circuit. It rejected the common carrier doctrine and analogized the platforms' activities to the legal regime which covers editorial discretion. It held that though social media platforms are new, the essence of their project is something the Court "has seen before."[375] The platforms "include and exclude, organize and prioritize and produce their own distinctive compilations of expression."[376] Thus, they resemble "traditional publishers and editors" who "also select and shape other parties' expression into their own curated speech products."[377] This means

[370] Moody v. Netchoice, LLC., 144 S. Ct. 2383, 2393 (2024).
[371] *Id.* at 2397–98.
[372] *Id.* at 2397.
[373] *Id.* at 2398.
[374] *Id.*
[375] *Id.* at 2393.
[376] *Id.*
[377] *Id.*

that "laws curtailing their editorial choices must meet the First Amendment's requirements."[378]

Justice Kagan's opinion criticises the Fifth Circuit's opinion which held that the content choices the major platforms make for their main feeds are "not speech" at all, so states may regulate them free of the First Amendment's restraints. For the same justice, the Fifth Circuit was wrong in concluding that Texas's restrictions on the platforms' selection, ordering, and labelling of third-party posts do not interfere with expression.[379] Per Justice Kagan, it was also wrong to treat as valid Texas's interest in changing the content of the platforms' feeds.[380] Presenting a curated compilation of speech originally created by others is expressive activity, so *Miami Herald Publishing Co. v. Tornillo*,[381] the case protecting editorial discretion, is relevant.[382] The First Amendment provides protection when an entity engaging in expressive activity, including compiling and curating others' speech, is directed to accommodate messages it would prefer to exclude.[383] Justice Kagan emphasized that the editorial function is an aspect of speech[384] and that an entity exercising editorial discretion in the selection and presentation of content is engaged in Speech activity.[385] These activities are protected expressive activities because they involve "[d]eciding on the third-party speech that will be included in or excluded from a compilation—and then organizing and presenting the included items."[386] This

[378] *Id.* The Court provides more clarification on how platform's activities are to be considered here as compared to a previous case, *Twitter, Inc, v. Taamneh*. 598 U.S. 471 (2023). In *Taamneh*, where the Court decided whether major platforms may be considered as "aiding and abetting" terrorism, the Court did not take a clear position on this issue. It held there that "these platforms appear to transmit most content without inspecting it." *Id.* at 499. The court held that these platforms are generally available to the public and that they do not appear to carefully screen content before allowing users to upload it onto their platforms. *Id.* at 498. On the contrary, they "appear to transmit most content without inspecting it." *Id.* at 499. The "recommendation" algorithms the platforms use to make available to the public material related to their interests "appear agnostic as to the nature of the content" they transmit. *Id.* The platforms "are global in scale and they allow hundreds of millions (or billions) of people to upload vast quantities of information on a daily basis." *Id.* at 500.

[379] *Moody*, 144 S. Ct. at 2399.

[380] *Id.*

[381] 418 U.S. 241 (1974).

[382] *Moody*, 144 S. Ct. at 2400.

[383] *Id.* at 2401.

[384] *Id.* at 2402 (citing Denver Area Educ. Telecomms, Consortium, Inc. v. FCC, 518 U.S. 727, 737 (1996)).

[385] *Id.* at 2402 (citing Ark. Educ. Television Comm'n v. Forbes, 523 U.S. 666, 674 (1998)).

[386] *Id.*

does not change "just because a compiler includes most items and excludes just a few."[387]

Justice Kagan further notes that although the new media differ from the old, analogies to old media, even if imperfect, can be useful. "The government may not, in supposed pursuit of better expressive balance, alter a private speaker's own editorial choices about the mix of speech it wants to convey."[388] Facebook offers a personalized collection of stories to every user as soon as they open their account via algorithms, the algorithms being based on each user's history.[389] The fact that platforms use algorithms and not human beings to implement those standards, to prefer content deemed trustworthy or to suppress content viewed as deceptive, does not alter the fact that they engage in curating speech, because the platforms apply their community standards to engage in content moderation.[390] Facebook removes content in cases of violence and criminal behaviour, to ensure safety online and protect against suicide and self-injury, sexual exploitation, bullying and harassment.[391] They also remove objectionable content (hate speech, violent and graphic content), and content that raises concerns as to its Integrity and Authenticity (false news, manipulated media).[392] YouTube's guidelines also target videos falling within hate speech, misinformation, violent or graphic content, and child safety.[393] These policies rest, for Justice Kagan, on a series of choices about whether and how to convey posts having a certain content or viewpoint. Their expressive quality lies in that they "rest on a set of beliefs on which messages are appropriate and which are not."[394] Justice Kagan clearly states that enforcing Texas' legislation means that the platforms cannot disfavor posts because they "support Nazi ideology; advocate for terrorism; espouse racism, Islamophobia, or anti-Semitism glorify rape or other gender-based violence; encourage teenage suicide and self-injury; discourage the use of vaccines; advise phony treatments for diseases; advance false claims of election fraud."[395] In other words, content

[387] *Id.*
[388] *Id.* at 2403.
[389] *Id.*
[390] *Id.* at 2403–04.
[391] *Id.* at 2404.
[392] *Id.*
[393] *Id.*
[394] *Id.* at 2405.
[395] *Id.* at 2405.

moderation as currently practiced by the platforms and under the influence of EU law and other market imperatives the platforms may be facing—such as their own need to protect their users from being exposed to speech they may consider harmful—is protected under the First Amendment.

For Justice Kagan, because these are all inherently expressive activities, Texas' statute which targets them does not pass even intermediate scrutiny.[396] Under that standard, a law must further a "substantial governmental interest" that is "unrelated to the suppression of free expression."[397] Although many possible interests relating to regulating social media can meet that test, "the interest Texas has asserted cannot carry the day," because "it is very much related to the suppression of free expression."[398] This is the case, because it aims to "correct the mix of speech that the major social-media platforms present" in a way that advances the government's vision on what is the proper ideological balance.[399] Justice Kagan repeated a standard approach to speech rights by the Court, according to which "[the government] cannot prohibit speech to improve or better balance the speech market."[400] She cites in this respect *Buckley v. Valeo*,[401] the decision related to political campaign funding, according to which "the government may not 'restrict the speech of some elements of our society in order to enhance the relative voice of others.' "[402] And the government may not pursue it consistently with the First Amendment. In other words, she repeats a standard First Amendment principle on the basis of which government intervention is not legitimate in order to redistribute speech rights within civil society.[403]

Justice Kagan engages with *Turner Broadcasting System, Inc. v. FCC* (Turner II)[404] to emphasize that "a private party's collection of third-party content into a single speech product is itself expressive and intrusion into that activity must be specially justified under the First Amendment."[405] Justice Kagan engaged with

[396] *Id.* at 2391, 2405–07.
[397] *Id.* at 2407 (citing United States v. O'Brien, 391 U.S. 367, 377 (1968)).
[398] *Id.* at 2407.
[399] *Id.*
[400] *Id.*
[401] 424 U.S. 1, 48–49 (1976).
[402] *Moody*, 144 S. Ct. at 2407.
[403] *See* Tourkochoriti, *supra* note 15, at 171–72, 181–82, 232.
[404] 520 U.S. 180, 185, 189–90 (1997).
[405] *Moody*, 144 S. Ct. at 2401.

PruneYard Shopping Center v. Robins,[406] where the Court rejected a shopping mall's First Amendment challenge to a California law requiring it to allow members of the public to distribute handbills on its property, noting that the mall owner did not claim that they were engaged in any expressive activity.[407] For Justice Kagan, in the latter there was little prospect of misattributing the opinions of the third party's speech to the host.[408]

Although Justice Kagan emphasizes the platforms' expressive rights, a good number of Justices believe that the use of AI in content moderation may not be covered by the First Amendment. Justices Barrett and Alito wrote two concurring opinions which bring some variation to Justice Kagan's reasoning. They distinguish between human content moderators and the use of AI in content moderation. Justice Barrett questions whether the use of AI in moderating hate speech should be covered by First Amendment rights.[409] The reliance on large language models to determine what is "hateful" may not, for Justice Barrett, necessarily be an inherently expressive choice.[410] Given that the use of AI leads to results that even the researchers and programmers that came up with them do not understand, Justice Alito believes (and Justice Thomas and Gorsuch joined his opinion) that it should not be covered by editorial discretion.[411] Justice Alito also appears concerned by the "enormous power" exercised by platforms like Facebook and YouTube as a result of "network effects."[412] Justice Jackson notes that "not every potential action taken by a social media company will qualify as expression protected under the First Amendment" and advises that courts must examine "how the regulated activities actually function before deciding if the activity in question constitutes expression and therefore comes within the First Amendment's ambit."[413]

2. Users' Procedural Rights

Justice Kagan appears to be rejecting the common carrier doctrine as to viewpoint discrimination. This has implications as to whether platforms should

[406] 447 U.S. 74 (1980).
[407] *Moody*, 144 S. Ct. at 2401.
[408] *Id.* at 2406.
[409] *Id.* at 2410.
[410] *Id.*
[411] *Id.* at 2438–39.
[412] *Id.* at 2439.
[413] *Id.* at 2411–12.

be allowed to entirely remove users as well. This is important because platforms may escalate in the measures they take against users. They may moderate their speech, or they may decide to remove users entirely. For the Justice, "[d]eciding on the third-party speech that will be included in or excluded from a compilation—and then organizing and presenting the included items—is expressive activity of its own."[414] This does not appear to leave room for protecting users' procedural rights in the way these rights are protected by the DSA.

Furthermore, Justice Kagan's analysis in relation to the motivation behind Texas' law is relevant not only to content moderation but also to users' rights. She emphasizes that the motivation behind the Texas statute, which was to protect "conservative viewpoints and ideas" from being silenced,[415] is impermissible because the government in Texas was dictating the inclusion of a specific viewpoint.[416] For Justice Kagan, Texas (under the pretext of preventing viewpoint discrimination) is actually dictating the inclusion of a specific view point, which is unacceptable under the First Amendment.[417] "On the spectrum of dangers to free expression," she notes, "there are few greater than allowing the government to change the speech of private actors in order to achieve its own conception of speech nirvana."[418] This ruling creates difficulties because a platform may allow users to have accounts and engage in content moderation of their speech. They may also engage in more radical measures such as removing users entirely as part of a series of escalating sanctions against them. Although Justice Kagan did not directly address this problem, the response may be in her treatment of Florida's requirement to provide rationales for exclusion.

Justice Kagan engaged with Florida's rules (which approximate the DSA) requiring the platforms to provide a "thorough rationale explaining the reason that the social media censored the user"[419] and a "precise and thorough explanation of how the social media platform became aware of the censored content or material,

---

[414] *Id.* at 2402.

[415] *Id.* at 2407 (quoting Texas Governor Greg Abbott's comments in signing the legislation: "[S]ilencing conservative views is un-American, it's un-Texan and it's about to be illegal in Texas").

[416] *Id.* at 2407. For an analysis of this point, *see* Evelyn Douek & Genevieve Lakier, *Lochner.com?*, 138 HARV. L. REV. 100, 138 (2024).

[417] *Moody*, 144 S. Ct. at 2407.

[418] *Id.* at 2407.

[419] FLA. STAT. § 501.2041(3)(c) (2023).

including a thorough explanation of the algorithms used, if any, to identify or flag the user's content or material as objectionable."[420] For Justice Kagan, the appropriate test in this case is asking whether this requirement unduly burdens expression in reference to *Zauderer v. Office of Disciplinary Counsel of Supreme Court*.[421] *Zauderer* was a case on commercial speech, which required disclosures by attorneys advertising contingent fee arrangements to avoid misleading customers. The Court articulated the criterion that "[C]ommercial speech that is not false or deceptive and does not concern unlawful activities … may be restricted only in the service of a substantial governmental interest and only through means that directly advance that interest."[422] And restrictions on commercial speech should also be narrowly crafted to serve the State's purposes. Commercial speech is subject to intermediate scrutiny, and it is difficult to anticipate whether regulation protecting rights of access to social media platforms may be upheld under the doctrine. As Jack Balkin also notes, the procedural guarantees are very important from the perspective of protecting access to social media platforms.[423] These refer to the fact that platforms must apply the criteria of excluding users consistently and non-arbitrarily. It is not clear whether Justice Kagan's reliance on the theory of commercial speech for these procedural guarantees is adequate to protect these rights. The difficulties that platforms face in performing this kind of reporting at scale may be seen as burdening platforms unduly and thus as being in tension with the First Amendment under *Zauderer*.[424] It is not clear whether the type of procedural rights guaranteed by the DSA may be justified under the compelled commercial speech doctrine.

If, according to Justice Kagan, Texas's law had a problematic motivation under the façade of imposing viewpoint neutrality, the DSA shows that it is possible to create procedural rights to ensure that social media platforms respect users' rights. The DSA enacted in the EU offers a model on the legal avenues users should have to complain against social media platforms if they have been excluded unfairly.[425] It shows that it is possible to impose upon the platforms viewpoint

[420] *Id.* § 501.2041(3)(d).

[421] *Moody*, 144 S. Ct. at 2398. *See also* Zauderer v. Off. of Disciplinary Couns. of Sup. Ct., 471 U.S. 626, 657–58 (1985).

[422] *Zauderer*, 471 U.S. at 638.

[423] *See also* Balkin, *supra* note 71, at 158.

[424] *See also id.* at 164.

[425] *See supra* Part II.A.1.

neutral requirements to justify why they are limiting users' speech and if the measures they take against users culminate to excluding them, why they were excluded, and how long.[426]

In this respect, the DSA offers a more comprehensive system of protection to anyone based in the EU, because it provides that platforms and governments must make available several legal avenues, in parallel to having access to courts, for users to complain if they think their speech was unfairly moderated. In the European Union, the legal framework is more robust in this respect, because, as analysed earlier, citizens can claim the protection of constitutional rights to free speech against platforms too before courts.[427] This means that users in the United States do not have the same legal avenues that users have in Europe to complain against the platforms. All they have are the avenues the platforms themselves are making available to them, e.g. the Facebook Oversight Board. The Board is very selective in the cases it examines, which means that most users who might want to complain about unfair content moderation of their speech are unable to do so.[428] If the Supreme Court is in favor of protecting users' access to social media platforms, it may need to engage with a different theory which will allow governments to impose these requirements. Or, platforms may actually decide to abide by the requirements the EU has posed in this respect too on their own, in order to protect their reputational interests which are monetizable interests.[429]

C. New York's And California's Attempts To Limit Hate Speech.

New York and California each enacted legislation aiming to limit hate speech online, which contain some elements that also exist in the DSA.[430] Two circuit courts have found that both these attempts to dictate to the companies that they should be moderating hate speech are contrary to the First Amendment.[431] The

---

[426] *Id.*

[427] *See supra* Part I.A.

[428] *See Appeal to the Oversight Board*, Oversight Board, https://www.oversightboardappeals.com/login/?locale_redirect=1&redirect_url=https%3A%2F%2Foversightboardappeals.com%2Fsubmit [https://perma.cc/MV73-E8UH] (last visited Mar. 27, 2026).

[429] *See infra* Part IV.A.

[430] For California, *see* Cal. Bus. & Prof. Code § 22677 (West 2022). For New York, *see* N.Y. Gen. Bus. Law § 394-ccc (McKinney 2024).

[431] X Corp. v. Bonta, 116 F.4th 888, 903 (9th Cir. 2024); *see also* Eugene Volokh, Locals Tech. Inc. v. James, 148 F.4th 71, 83 (2d Cir. 2025).

most comprehensive of the two is California's law.[432] It requires social media companies to submit to the Attorney General a report on a semi-annual basis, which will include the terms of service of their platforms and any changes, and a statement of whether and how they define "hate speech" or "racism," "extremism or radicalization," "disinformation or misinformation," "harassment," "foreign political interference and controlled substance distribution."[433] The report should also include a detailed description of the content moderation practices used by the social media company for their platforms, how automated content moderation systems enforce terms of service, and when they involve human review.[434] The same report should indicate how the company responds to user reports of violations of the terms of service and how it takes broader action against individual users or against groups of users that violate the terms of service.[435] It should also indicate the content that was flagged by the social media company as belonging to any of the above categories, such as the total number of flagged and actioned items of content, the number of items of content that were removed, demonetized or deprioritized by the social media company, the number of times the same items were viewed and shared before they were removed and the number of times users appealed company actions, and the numbers of reversals of these actions.[436]

The Ninth Circuit, which examined the legislation, found that the content category report provisions facially violate the First Amendment.[437] The Court applied strict scrutiny to the provisions because in its opinion they are "content-based" and they "compel the platform's non-commercial speech."[438] The Reports are not commercial speech, because they "do not propose a commercial transaction."[439] They require each company to "recast its content moderation practices in language prescribed by the State, implicitly opining on whether and how certain controversial categories of content should be moderated."[440] Further,

[432] Cal. Bus. & Prof. Code § 22677 (West 2022).
[433] *Id.* § 22677(a)(1)–(3).
[434] *Id.* § 22677(a)(4)(A)–(B).
[435] *Id.* § 22677(a)(4)(C)–(D).
[436] *Id.* § 22677(a)(5)(A)(i)–(vii).
[437] X Corp. v. Bonta, 116 F.4th 888, 898 (9th Cir. 2024).
[438] *Id.* at 899–900.
[439] *Id.* at 901 (citing United States v. United Foods, Inc., 533 U.S. 405, 409 (2001) and IMDb.com Inc. v. Becerra, 962 F.3d 1111, 1122 (2020)).
[440] *Id.* at 901.

the Court of Appeals held that the provisions fail strict scrutiny because they are not narrowly tailored to serve California's goal of requiring social media companies to be transparent about their policies and practices so that consumers can make informed decisions about where they consume and disseminate news and information.[441] The Court remanded to the district court to determine whether the Report provisions are severable from the remainder of the legislation and if so, which, if any, of the remaining challenged provisions should be subject to the preliminary injunction.[442]

New York's legislation obliges social media networks to have "a clear and concise policy readily available and accessible on their website and application" which will indicate how the network will respond and address reports of incidents of hateful conduct on their platform.[443] The legislation contains a definition of "hateful conduct" as "the use of a social media network to vilify, humiliate, or incite violence against a group or a class of persons on the basis of race, color, religion, ethnicity, national origin, disability, sex, sexual orientation, gender identity or gender expression."[444] It also creates an obligation for social media platforms to create a mechanism for users to report incidents of hateful conduct and to provide a response to any individual reporting hateful conduct.[445] The legislation creates civil responsibility for platforms which "knowingly" fail to comply with the requirements of the section.[446] The legislation contains an interpretive clause requiring that it will not be construed as adversely affecting the right to free speech under the First Amendment.[447]

The Second Circuit held that the constitutionality of the law depends on how it is interpreted.[448] If the law is interpreted to mean that the platforms must adopt the state's definition of "hateful conduct," then the statute fails under intermediate scrutiny.[449] If the law is interpreted to only impose a requirement to have a content moderation policy and a general mechanism for reporting content-related

[441] *Id.* at 903.
[442] *Id.* at 904.
[443] N.Y. Gen. Bus. Law § 394-ccc(3) (McKinney 2024).
[444] *Id.* § 394-ccc(1)(a).
[445] *Id.* § 394-ccc(2).
[446] *Id.* § 394-ccc(5).
[447] *Id.* § 394-ccc(4).
[448] Eugene Volokh, Locals Tech. Inc. v. James, 148 F.4th 71, 83 (2d Cir. 2025).
[449] *Id.* at 94.

complaints, then the statute survives constitutional scrutiny under *Zauderer.*[450] The Court engages with *Moody* and reaffirms the First Amendment rights of the platforms, which engage in "their own expressive activity."[451] Following *Moody's* suggestion, it considers *Zauderer* as offering the governing standard when evaluating the constitutionality of a law related to social media platforms.[452] For the court, if the law merely requires social media networks to publicly disclose their content moderation policies and contains no requirement that these policies address "hateful conduct" as defined by the statute, then the statute requires "purely factual and uncontroversial information about the terms under which [commercial] services will be available."[453] A "neutral disclosure requirement" would survive *Zauderer* scrutiny, provided that it is "truly agnostic" about the substance of the content moderation policy.[454] For the Court, *Zauderer* requires protection for commercial speech, because of its value in informing consumers and thus allowing them to choose the platform of their choice, a moderated or an unmoderated one.[455] However, if the law requires platforms to accept the state's definition of hateful conduct, then it is requiring "social media networks to 'provide a forum for someone else's views' – in this case the State's."[456] In this case, it would not survive under either strict or intermediate scrutiny.

IV
The State of the Challenges in Moderating Hate Speech and Misinformation Online

A. Platforms' Latest Policy Changes In The US—Towards A Divided Internet?

Moody means that platforms are allowed to decide on their own whether they will regulate hate speech and misinformation or not in the United States. They may decide to abide by the obligation that they have within the European Union to engage in content moderation in the United States as well, although they do not have a legal obligation to do so. As discussed earlier, platforms usually abide by

[450] Zauderer v. Off. of Disciplinary Couns. of Sup. Ct. of Ohio, 471 U.S. 626, 651 (1985).
[451] *James*, 148 F.4th at 84.
[452] *Id.* at 85.
[453] *Id.* (citing Zauderer v. Off. of Disciplinary Couns. of Sup. Ct. of Ohio, 471 U.S. 626, 651 (1985)).
[454] *Id.* at 90–91.
[455] *Id.* at 86 ("[T]he extension of First Amendment protection to commercial speech is justified principally by the value to consumers of the information such speech provides." (citing *Zauderer*, 471 U.S. at 651)).
[456] *Id*. at 94 (citing Moody v. Netchoice, LLC., 144 S. Ct. 2383, 2400 (2024)).

the strictest legal regime that exists in one part of the world, and then generalize their behavior due to operational reasons.[457] Or they may decide to forgo content moderation altogether. X(Twitter)'s practices since 2022 and the declarations by Facebook CEO Mark Zuckerberg in the beginning of 2025 point in the direction of scaling back content moderation in the United States.[458] In making these decisions, platforms need to evaluate also whether they are interested in offering to their users an environment where they feel safe to interact with others to maintain their popularity among the public, and by extension among investors and advertisers.

In general, the platforms have a strong reputational interest to maintain a safe atmosphere which protects users from being exposed to hatred. This reputational interest is monetizable. It relates to retaining customers and funding for the publicly owned corporations. Scholars discuss the business case for corporate social responsibility, which associates the value of a company with the social value of the policies it maintains.[459] Especially in recent years, the corporate world has moved more generally towards Corporate Social Responsibility and Environmental Social Governance policies which aim to prevent creating human rights risks across their planet.[460] Global corporations are interested in affirming their standing as good Global Citizens to keep raising revenue, if they are publicly held, and to keep their customers. Social Media companies' reputational interests are related to losing income from advertisers if they do not protect their users from exposure to content they are not happy with. X(Twitter)'s value as a company dropped dramatically in the past following its cuts in content moderation.[461] The same changes also led to significant losses in advertising revenue.[462] The

[457] *See supra* Introduction.

[458] *See generally* Spring, *supra* note 1; Duffy, *supra* note 1.

[459] Gordon L. Clark, Andreas Feiner & Michael Viehs, From The Stockholder To The Stakeholder: How Sustainability Can Drive Financial Outperformance 10 (2015).

[460] *See generally, e.g.*, Kishanthi Parella, *International Law in the Boardroom*, 108 Cornell L. Rev. 839, (2023); Luca Enriques et al., *How the EU Sustainability Due Diligence Directive Could Reshape Corporate America,* 78 Stan. L. Rev. 241, (2026).

[461] Adam Baggatt, *Value of X has Fallen 71% Since Purchase by Musk and Name Change from Twitter*, Guardian (Jan. 2 2024, at 09:24 AM ET), https://www.theguardian.com/technology/2024/jan/02/x-twitter-stock-falls-elon-musk [https://perma.cc/2S6L-73P3].

[462] Aisha Counts & Eari Nakano, *Twitter's Surge in Harmful Content a Barrier to Advertiser Return*, Bloomberg (July 19, 2023, at 6:00 AM ET), https://www.bloomberg.com/news/articles/2023-07-19/twitter-s-surge-in-harmful-content-a-barrier-to-advertiser-return [https://perma.cc/R29T-8ND7].

platform rebounded only thanks to the political alliances of its current owner.[463] And more recently, thanks to the purchase of X by xAI, also owned by the same businessman.[464] Therefore, even in the absence of regulation, platforms have an interest in regulating themselves by moderating hate speech and ensuring safe environments to their users. Adapting to the standards elaborated by the DSA, even in parts of the world where there is no legal obligation to do so, may offer a good avenue towards protecting the platforms' reputational interests. This will likely confirm the Brussels effect in this area as well.[465] Alternatively, platforms may opt to come up with self-regulation modes that combine elements from the DSA and end-user empowerment, as discussed later in this paper.[466]

As regards misinformation, X appears to have transitioned to a system of community notes and to have reduced its content moderation team, a proposal which Meta also follows since January 2025.[467] Other platforms have also announced that they are transitioning to the same system for fact checking purposes.[468] As analysed below, several studies indicate that the Community Notes System is not sufficient to address misinformation and disinformation. Meta announced in January 2025 several other policy changes such as "Simplifying" content policies by removing certain restrictions on topics like immigration and gender, changing enforcement approach for policy violations, focusing automated filters only on illegal and high-severity violations, requiring user reports before taking action on lower-severity violations, increasing the confidence threshold required before removing content, reintroducing civic and political content into recommendation systems on Facebook, Instagram, and Threads, and relocating

[463] Mark Sweney, *Value of Elon Musk's X 'rebounds to $44bn purchase price'*, Guardian (March 19, 2025, at 07:33 AM ET), https://www.theguardian.com/technology/2025/mar/19/value-elon-musk-x-rebounds-purchase-price [https://perma.cc/HXW6-63ND].

[464] Greg Bensinger, *Musk's Social Media Firm X Bought by his AI Company, Valued at $33 Billion*, Reuters (March 29, 2025, at 1:25 AM ET), https://www.reuters.com/markets/deals/musks-xai-buys-social-media-platform-x-45-billion-2025-03-28/ [https://perma.cc/7WNP-G6FH]. The purchase valued X at $33 billion, not including $12 billion in debt, which is already lower from the price Elon Musk bought the company in 2023; *see* Kurt Wagner & Katie Roof, *Musk's xAI Deal Offers Unexpected Win for X Investors*, LA Times (March 31, 2025, at 4:19 PM PT), https://www.latimes.com/business/story/2025-03-31/musks-xai-deal-offers-unexpected-win-for-x-investors [https://perma.cc/3S4E-NBJF].

[465] *See generally* Bradford, The Brussels Effect, *supra* note 29.

[466] *See infra* Part IV.A.2–3.

[467] *See infra* Part IV.A.1.

[468] *See id.*

trust and safety and content moderation teams from California to Texas.[469] Meta has also announced that they are deleting all videos posted on their platforms after 30 days.[470] This is a measure which aims to save it time from its content moderation practices. Several of these changes are likely to lead to a divided internet between the US and the EU.

1. Misinformation

In the area of misinformation, since 2023, X cut down on its content moderation team and deployed a crowd-sourced fact-checking system under the name Community Notes.[471] The system was piloted in the US and used in other countries as well.[472] Several other platforms such as YouTube, TikTok and Meta are also transitioning to this system.[473] Meta's CEO announced in January 2025 that Meta is implementing changes in the platform's operations in the United States, which includes, among other changes, introducing Community Notes.[474] This system is very different from the one that the DSA requires platforms to apply in Europe, although X has applied it in several European states too. As analysed earlier,[475] the DSA requires the platforms to prevent systemic risks and to maintain content moderation by professionals, which are bound by the Code of Practice on Disinformation elaborated by the EU.[476]

Several studies have shown that the system of community notes has a number of shortcomings which make it unreliable alone to combat misinformation and

---

[469] "This will help remove the concern that biased employees are overly censoring content," according to Mark Zuckerberg. *See* Justin Hendrix, *Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations*, TechPolicy.Press (Jan. 7, 2025), https://www.techpolicy.press/transcript-mark-zuckerberg-announces-major-changes-to-metas-content-moderation-policies-and-operations/ [https://perma.cc/MPT8-7P9Q].

[470] *Updating Our Facebook Live Video Storage Policy*, Meta (2025), https://about.fb.com/news/2025/02/updating-our-facebook-live-video-storage-policy/amp/ [https://perma.cc/GTM9-LZPC].

[471] Paul Bouchaud & Pedro Ramaciotti, *Algorithmic Resolution of Crowd-Sourced Moderation on X in Polarized Settings Across Countries*, arXiv, at 2 (June 18, 2025), https://arxiv.org/abs/2506.15168v1 [https://perma.cc/8YQJ-9R5T]. The study included United States, United Kingdom, Japan, Spain, France, Brazil, Canada, Germany, Argentina, Israel Australia, Poland and Mexico.

[472] *Id.*

[473] *Id.*

[474] *See* Hendrix, *supra* note 469.

[475] *See supra* Part III.B.2.

[476] Eur. Comm'n, Strengthened Code of Prac. on Disinformation (2022), https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation [https://perma.cc/5SMQ-ZGYM].

disinformation. A study by media scientists in France on how the system was implemented in 14 countries on X shows some of these shortcomings towards combatting misinformation.[477] The transition to the Community Notes system was motivated by criticisms of the use of expert fact-checking.[478] Although expert moderation was shown to successfully decrease the extent to which people agree with misleading claims, it faced criticisms for bias and for its likelihood to further radicalize polarized individuals.[479] The Community Notes system allows self-selected contributors to attach contextual notes to posts they consider "misinformed or potentially misleading," and it asks other contributors to rate the notes as *Helpful*, *Somewhat Helpful*, or *Not Helpful*.[480] The system infers ideological positions for both notes and raters and then predicts *Helpful* ratings on the basis of their ideological alignment.[481] A note rated *Helpful* by raters of diverse ideological backgrounds is given a higher *ßn* value in this system.[482] The system then creates an algorithm using these parameters to resolve conflicting ratings by selecting notes with the highest *ßn* value. Comments are removed or kept depending on whether diverse raters find them helpful or not. In the system's pilot phase in the US, there was high agreement between Community Notes and expert fact-checker classifications, and the system was also found to generate greater user trust and to be less prone to entrench individuals.[483] While the system was developed and tested in the United States, where political disagreement is modeled using a single ideology dimension, it turned out to be less promising for other parts of the world where political competition is organized along multiple ideological currents.[484] Posts discussing politically divisive issues are not likely to be rated as *Helpful* and thus moderated, while others reporting scams and misleading practices are likely to

[477] Bouchaud & Ramaciotti, *supra* note 471, at 2.

[478] *Id.*

[479] *Id.*

[480] *Id.* If sufficient contributors from diverse viewpoints rate a note as *Helpful*, X assigns it *Helpful* status, while it removes from public view notes marked as *Not Helpful*. *Id.* at S1. Contributors who systematically author notes assigned *Not Helpful* status may lose their ability to contribute notes, and notes that have not received sufficient ratings from contributors with diverse perspectives remain invisible to users outside the Community Notes program. *Id*.

[481] *Id.* at 2.

[482] *Id.*

[483] *Id.*

[484] *Id.* at 2.

reach that status.[485] This means that Community Notes cannot properly moderate the most polarizing issues related to national politics.[486] Allowing polarizing, political content to remain unresolved is a design choice by the platform.[487] The researchers conclude that the Community Notes system, which substitutes crowd-sourced moderation that relies on consensus for expert fact-checking, needs a risk assessment for polarizing content, especially during election periods.[488] Posts that are uninformative or misinformative will not necessarily be caught in time.

A second study, conducted by academics in computer science from all over the world, also points out that because there are important problems with how the system works, it cannot be used without being supplemented by expert fact-checkers.[489] Community Notes is effective for those cases where professional content moderators have already fact checked claims.[490] The system is not successful as a work-around to biases because Community Note users carry their own cognitive biases.[491] Users overestimate the truthfulness of claims and are overconfident in their ability to judge statements.[492] Few have the time or motivation to participate, and the personalization and polarization caused by prioritization algorithms means that knowledgeable users may not encounter misleading posts at all.[493] Ideological diversity is only measured for political bias and not for other cultural, linguistic or conspiratorial biases.[494] Furthermore, the system may be attacked by malign bots or groups of users who "inflate social reputations," "amplify specific narratives," and engage in other harmful behavior.[495] Such groups may "deceive the community notes algorithm" by "creating an artificial perception of internal disagreement through

[485] *Id.* at 4.
[486] *Id.* at 7.
[487] *Id.*
[488] *Id.*
[489] Isabelle Augenstein et al., *Community Moderation and the New Epistemology of Fact Checking on Social Media*, arXiv, at 1 (May 26, 2025), https://arxiv.org/pdf/2505.20067v1 [https://perma.cc/3KCE-SNBC].
[490] *Id.* at 7.
[491] *Id.* at 6.
[492] *Id.*
[493] *Id.* at 3.
[494] *Id.* at 6.
[495] *Id.*

collusive inorganic activity on benign content."[496] The study concludes that only a collaborative approach between professional fact checkers and the community model may bring about reliable results,[497] something that the DSA also encourages the platforms to do.

Further, research by NGOs indicates that misinformation is being made widely available on X without being checked through the Community Notes system and without any other effort by the platform to contain it.[498] Another positive element in the DSA is that it encourages platforms to consider seriously notes made by "trusted flaggers."[499] As analyzed earlier, these trusted flaggers are NGOs which self-register as the defenders of human rights against hate speech.[500] They may also contribute to the credibility of the notes system if the platforms take their warnings seriously, as the DSA requires.

The DSA, which encourages professional fact-checking performed by journalists who are trusted to have received ethics training, appears to offer a more reliable system for combating misinformation and disinformation. As the first report on its application has shown, the DSA has been successful in bringing together representatives from platforms and civil society actors to collaborate on tools and methods of content moderation.[501] Civil society organizations emphasized the need for human moderators to be well-trained and aware of local contexts, languages, and cultures.[502] At the operational level, the report mentions that this collaboration led to recognition of the need to create close links between content moderation and legal teams to escalate borderline cases, establish partnerships with anti-bias researchers and civil society organizations, and develop market-specific moderation processes to capture local nuances in languages, slang terms, and cultural references while ensuring content moderation enforcement by human reviewers.[503]

---

[496] *Id.*

[497] *Id.* at 7.

[498] *X Posts Claiming European Leaders Took Cocaine Reach 135M Views with no Community Notes*, Ctr. for Countering Digit. Hate (May 15, 2025), https://counterhate.com/research/x-posts-claiming-european-leaders-took-cocaine-reach-135m-views-with-no-community-notes/ [https://perma.cc/M5XQ-M9SB].

[499] *See supra* Part II.A.8.

[500] *See Id.*

[501] *See* Eur. Bd. For Digit. Services, *supra* note 24, at 34; *see also supra* Part II.A.3.

[502] *See* Eur. Bd. For Digit. Services, *supra* note 24, at 34.

[503] *Id.* at 35.

In general, as analyzed earlier, the DSA obliges the platforms to create crisis protocols in case of emergencies, i.e. waves of misinformation or disinformation that occur during a pandemic.[504] If the platforms do not generalize these protocols across their operations around the world, systemic risks that occur outside of the EU may not be addressed in time. As the COVID-19 crisis showed, public health emergencies are difficult to manage due to the evolving nature of expert guidance in response to the fast pace of scientific discoveries about the best ways to respond to unknown viruses.[505] Even if obtaining accurate information during a pandemic has its own challenges, the EU approach is to err on the side of caution and to prevent information that is likely misleading from being shared quickly around the platforms. As Evelyn Douek reminds us, the race to combat misinformation online during the recent COVID-19 pandemic led to frequent changes in public guidance by the relevant authorities, which undermined the public's trust towards the platforms in the US.[506] The problem, however, seems to have been that the platforms did not clearly explain to the public how they communicated with the government and reliable expert sources to define the guidance they made available to the public.[507] Evelyn Douek argues for "an affirmative vision for the role of platforms in policing medical claims," such as limiting claims that are directly related to physical harms while letting be claims that may not be entirely accurate but are not likely to lead to such harm.[508] This view is compatible with the liberty *Moody* recognized that platforms have. Having a protocol, according to what the DSA requires, may be a useful way for the platforms to proceed with their global operations. Or it could be the point of departure for them to elaborate a policy. The protocol required by the DSA allows the platforms to find for themselves the best way to engage in content moderation in this area, which allows them to follow some of the criteria Douek suggests.[509] To address this difficulty, other legal scholars in the US are arguing in favor of amending Section 230 of the Communications

[504] *See supra* Part II.A.2.

[505] *See* Wendy E. Parmet & Jeremy Paul, *COVID-19: The First Posttruth Pandemic*, 110 Am. J. Pub. Health 945, 945 (2020); Claudia E. Haupt & Wendy E. Parmet, *Lethal Lies: Government Speech, Distorted Science, and the First Amendment*, 2022 U. Ill. L. Rev. 1809, 1814 (2022).

[506] *See* Evelyn Douek, *The Politics and Perverse Effects of the Fight Against Online Medical Misinformation*, 134 Yale L.J. F. 237, 243–45 (2025) (describing how the changing guidance by authorities during the COVID 19 guidance led to concerns about how the platforms were using content moderation).

[507] *Id.* at 246.

[508] *Id.* at 262–66.

[509] *Id.*

Decency Act, which excludes from civil liability platforms for the content that others provide.[510] In the area of regulating emerging technology, scholars suggest that the "maximin" principle might offer guidance.[511] According to the principle, regulators must choose the policy with "the best worst-case outcome."[512] The crisis protocol may be a DSA requirement that imposes upon the liberty of the platforms. At the same time, it seems to be appropriate in the area of confronting crises, when the correct solutions might be uncertain. Erring on the side of caution may well be the best worst-case outcome in times of crises. Protecting their users could be an important imperative for the platforms to generalize these protocols and use them also outside of the EU, where they do not have a legal obligation to do so.

Additional difficulties arise from the use of prioritization algorithms, which address users on the basis of their previous history. These algorithms may enhance biases that users already have that are likely to radicalize them further. This can only inhibit the public dialogue thereby preventing users from potentially exposure to information that challenges their beliefs —- thus strengthening confirmation biases.[513] Studies have shown that recommendation algorithms detect users' interests very quickly, and adapt over time in a way that narrows exposure to new topics and perspectives.[514] In particular, research on YouTube and Twitter shows that "certain topics (e.g., political content) are reinforced more strongly than others."[515] And, recent research on X indicates that "right-leaning accounts are often more prominently featured in algorithmic curation" for both "right-leaning" and "neutral" accounts.[516] Although recent political science research indicates that the concerns with "echo-chambers" may have been exaggerated due to social

[510] *See, e.g.*, Leiter, *Free Speech on the Internet*, *supra* note 64, at 255.

[511] Cass R. Sunstein, *Maximin*, 37 Yale J. on Reg. 940, 950–51 (2020); Noam Kolt, *Algorithmic Black Swans*, 101 Wash. U. L. Rev. 1177, 1239 (2024).

[512] Sunstein, *supra* note 511, at 966.

[513] Vincent Blasi, *Is John Stuart Mill's on Liberty Obsolete?*, 5 J. Free Speech L. 151, 161 (2024).

[514] *See* Fabian Baumann et al., *Dynamics of Algorithmic Content Amplification on TikTok*, arXiv, at 1 (2025), https://arxiv.org/abs/2503.20231 [https://perma.cc/6WGM-8B3D].

[515] *Id.* at 20.

[516] Jinyi Ye, Luca Luceri & Emiliio Ferrara, *Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X During the 2024 U.S. Presidential Election*, Proc. 2025 ACM Conf. on Fairness, Accountability, & Transparency, 2349, 2352, 2355 (2025), https://doi.org/10.1145/3715275.3732159 [https://perma.cc/UHK4-2BA2].

media not functioning as viewers' only source of daily news[517] — nonetheless, more research in this area is needed to explore potential risks and responses. As the first report on the application of the DSA indicates, social media platforms, at least within the EU, are demoting or restricting potentially harmful content and using information banners or interstitials to limit the spread of such content.[518] The report also noted that some platforms design their recommender system to optimize for content quality not just view time.[519] This shows that the Act has encouraged platforms to reflect on the design of their prioritization algorithms if they have not done so by themselves.

2. Incitement to Hatred

The use of prioritization algorithms that are demoting harmful content is even more important in the area of incitement to hatred and violence. An important difference between European legal systems and the US is the definition of incitement. The DSA refers to the Code of Conduct, which platforms have agreed to abide by, and to the national legislation existing within EU Member States to define hate speech. The Code's definition which reflects the standards existing within EU States covers "public incitement to violence or hatred."[520] This standard is much broader than the one existing in the US. According to the criteria posed by *Brandenburg v. Ohio*, speech may be outlawed only when it encourages imminent lawless action that will very likely occur.[521] The criterion is motivated by the need to preserve social peace. Within EU legal systems, incitement to hatred

[517] Andrew Guess et al., *Avoiding the Echo Chamber About Echo Chambers: Why Selective Exposure to Like-Minded Political News Is Less Prevalent than You Think*, Knight Foundation, 1, 3 (2018), https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf [https://perma.cc/VB2P-EA6L]. The study covering the U.S. indicates that 43% of conservative Republicans and 26% of liberal Democrats consult opposing media sources for their information. *Id.* at 7. This leaves a large percentage of the public limited in the source of information it consults, leaving room for risks in the area of incitement to hatred and violence. The study also details that only about 20% of U.S. adults claim they regularly get their news from social media. *Id.* at 8. The study concludes that a subset of the most politically engaged and vocal members of the public is living in an echo chamber. *Id.* at 16. These are small numbers, but the risks are worth investigating.

[518] Eur. Bd. For Digit. Services, *supra* note 24, at 36; *See supra* Part II.A.3.

[519] Eur. Bd. For Digit. Services, *supra* note 24, at 36.

[520] Eur. Comm'n, *supra* note 219, at 1; *see also* Council Framework Decision 2008/913/JHA of 28 Nov. 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 2008 O.J. (L 328) 55.

[521] Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).

or violence entails legal responsibility independently of whether it encourages to imminent lawless action which is likely to occur or not.[522] This means that in the area of incitement, the difference in the legal standards will likely lead to a division of the internet. The *Brandenburg* Court follows John Stuart Mill, who famously defines incitement in his important book *On Liberty*, in reference to an agitator inciting an angry mob.[523] In the current context of social media operating on the basis of prioritization algorithms which may enhance hatred, the Mill-*Brandenburg* definition of incitement may no longer be adequate.[524] Whether a person is likely to engage in violent acts may well depend on their own circumstances in combination with targeted content due to prioritization algorithms. Users prone to radicalization may well be "captive audiences," which means that the EU definition of incitement to hatred may be preferable to the one of *Brandenburg*.

The constant exposure to hateful speech along with misinformation and the speed with which it is shared may be altering our understanding of how incitement to hatred and violence occurs. At the very least, it is imperative to investigate through interdisciplinary studies involving behavioral sciences (psychology) whether the way prioritization algorithms work is likely to contribute to incitement to hatred and to committing violence. Further research is required to identify how users prone to radicalization respond to constant exposure to materials that may constitute incitement. Further research is needed in this area to study the psychological impact of constant exposure to hate speech and how legislation related to bullying and harassment should apply to these behaviors when they occur online. The DSA's broader definition of incitement shows a preference for erring on the side of caution in the area of online incitement. As analyzed earlier, the first report on the application of the DSA indicates that platforms have adapted their recommender systems to demote harmful or potentially harmful content, which includes content that constitutes incitement to hatred, at least within the EU.[525] This measure, which follows the precautionary approach, is a very positive development in the absence of data related to how human beings react to being targeted with speech that constitutes incitement to hatred.

---

[522] *See* definition of incitement to hatred, *supra* Part II.A.

[523] *See* John Stuart Mill, On Liberty 52 (Batoche Books: Kitchener, 2001).

[524] *See* Leiter, *Free Speech on the Internet*, *supra* note 64, at 248.

[525] Eur. Bd. For Digit. Services, *supra* note 24, at 36. *See supra* Part II.A.3.

3. Hate Speech

Divergence is likely to occur between the US and the EU in the area of defining and limiting hate speech. While some social media platforms, Facebook included, engage in content moderation,[526] a division of the internet is likely to emerge in the interpretation of what constitutes hate speech and misinformation. National standards within each of the EU member states will play a role in this respect, and the same will occur in the US as well. The DSA leaves plenty of room for national understandings of what constitutes hate speech because it does not contain a definition itself. Instead, it refers throughout to "illegal hate speech."[527] This reference includes the definition of hate speech which exists in the Code of Conduct the EU asked major Social Media platforms to sign.[528] It also aims to allow the Member States to enforce their national standards in this respect. In Europe, the European Court of Human Rights (ECtHR) plays in important role in harmonizing the standards for the protection of free speech across the 46 Member States of the Council of Europe.[529] 27 of those are also EU members, where the DSA applies.[530] This means that the ECtHR will continue to play an important role in defining the acceptable standards for free speech and what constitutes hate speech. National institutions within the EU Member States are likely to play an even more important role daily for defining hate speech and whether users have been properly removed from the platforms or not.

Meta updated their hate speech policy in January 2025.[531] The new policy has replaced references to "hate speech" with references to "hateful conduct."[532] The platform has removed the reference to "statements of inferiority, expressions of contempt or disgust; cursing; and calls for exclusion or segregation" as well

---

[526] *See Detecting Violations*, Meta, https://transparency.meta.com/enforcement/detecting-violations/ [https://perma.cc/HH8J-9PHW] (last visited Mar. 6, 2026).

[527] *See, e.g.*, DSA, *supra* note 114, pmbl. ¶¶12, 62, 80, 87, 106, art 12.

[528] *See supra* Part II.A.

[529] *See* Natalie Alkiviadou, Hate Speech and the European Court of Human Rights 79 (Taylor & Francis Group 2025) (ebook).

[530] *See The Council of Europe and the European Union: Partners in Promoting Human Rights and Democracy*, Council of Europe (2024), https://edoc.coe.int/en/different-roles-shared-values/6332-leaflet-the-council-of-europe-and-the-european-union-partners-in-promoting-human-rights-and-democracy.html [https://perma.cc/9QUY-ATHQ].

[531] *See Hateful Conduct*, Meta, https://transparency.meta.com/policies/community-standards/hateful-conduct/ [https://perma.cc/J54M-VW4C] (last visited Mar. 6, 2026).

[532] *Id.*

as to "slurs that are used to attack people on the basis of their protected characteristics."[533] This alone is different from the law that exists in some EU states. Several of the removed phrases would be captured by French hate speech legislation, for instance.[534] French law punishes defamation and insult and foresees penalty enhancements when defamation and insult occur on the basis of someone's membership or non-membership of a specific ethnic group, nation, race or religion, their sex, their sexual orientation or gender identity or their disability.[535] The French *Cour de Cassation* (Supreme Court of the Ordinary Jurisdiction in France) has held that the penalty enhancement for group defamation applies to statements posted on Twitter, against a religious group falsely accusing it of having committed crimes in the past.[536]

Some other policy measures Facebook announced in January 2025 are compatible with the current content moderation practices it already follows under the DSA. Increasing the confidence threshold before removing content is something that the DSA itself is trying to achieve, thanks to the increased role it foresees for user reporting of hateful materials and the increase in legal avenues it requires the platforms and European states to make available to users. With the exception of removing restrictions on topics like immigration and gender, the policy changes announced do not appear to change much to the current content moderation practices. Increasing the confidence threshold before removing content, which Facebook also announced,[537] will most likely require some expert fact-checking, a requirement under the DSA.[538]

Furthermore, Facebook's new policies, interesting as they are, do not seem to create more legal avenues for users to complain about policies such as the ones

---

[533] *Id.*

[534] Loi du 29 juillet 1881 sur la liberté de la presse [Law of July 29, 1881 on the Freedom of the Press], ch. IV, arts. 23–35.

[535] *Id.* arts. 32–33.

[536] Cour de cassation [Cass.] [supreme court for judicial matters] crim., Oct. 15, 2019, 18-85.368, Inédit, ECLI:FR:CCASS:2019:CR01825 (Fr.) (finding that a statement on Twitter according to which "Jewish people are the ones responsible for the massacre of thirty million Christians in USSR between 1917 and 1947" constitutes group defamation under article 32, because it attributes to the Jewish people a fact which may be proven otherwise, and which violates their honor and their consideration).

[537] *See* Hendrix, *supra* note 469.

[538] *See supra* Part II.A.2.

users have in the EU under the DSA, as analyzed earlier.[539] In case of disagreement between a user and the platform, the only avenue that remains available is the Facebook Oversight Board.

Combatting hate speech and misinformation has always been challenging. The difficulties largely relate to defining what kind of speech should be considered hateful. Previously, arguments against limiting hate speech against social groups were based on the distinction between general claims against social groups and individualised expressions against persons. Scholars were arguing in the past that broad negative utterances against social groups should not trigger the enforcement mechanism against hate speech.[540] The reasons for this difference in treatment is that broad, hateful utterances against social groups are easily refutable by more speech. Individualised defamatory expressions against persons, on the other hand, arbitrarily associate a person with a social group and project upon them negative qualities associated with the same group. This kind of speech, the argument goes, should trigger the enforcement of anti-hate speech regulation. Hate speech frequently attributes social stereotypes associated with groups to individuals. Persons are associated arbitrarily with social groups and with the negative characteristics that these social groups have. In this case, hate speech should be limited by government. Government intervention is necessary when a person experiences a verbal attack which attributes to her characteristics attributed arbitrarily to a social group she is classified as being a member of. The justification for limiting speech is the arbitrariness which exists in the association of some characteristics with a person and in the association of the person with the social group. The person is denied the possibility to define for herself her personality and show it to others. This distinction between broad defamatory utterances and individualised attacks against persons needs to be evaluated anew in the context of online communication. Algorithms trigger targeted content which may reproduce broad hateful claims against groups in ways that are persistent upon a user. Constant exposure to hate speech together with the inability of checking the accuracy of these claims runs the risk of deteriorating the social standing of members of social groups which have historically experienced discrimination. Users are "captive

[539] *See supra* Part III.A.

[540] *See* Ioanna Tourkochoriti, *Should Hate Speech be Protected? Group Defamation, Party Bans and the Divide Between Europe and the US*, 45 Colum. Hum. Rts. L. Rev. 552, 552–622 (2014).

audiences" of hateful utterances. Further research in psychology is urgently needed to evaluate how people react to being targeted with hate speech due to algorithmic prioritization.

Furthermore, new concerns constantly emerge from the ability of AI to generate hate speech itself. For example, an AI chatbot called Grok created by xAI shared on X a series of antisemitic posts.[541] This was later attributed to the AI's primitive nature and a need for better automated systems to ban hate speech before it is posted.[542] The comments were deleted and xAI published an apology,[543] but the very presence of the comments and the absence of any mechanism that would prevent them from being posted is very concerning about the ability and willingness of the platform to prevent similar material from being made available. AI progress is likely to lead to similar phenomena in the future, because the competition between tech companies leads them to overlook ethics issues.[544] The combination of prioritization algorithms with AI Chatbots that engage in hate speech are likely to lead to systemic risks of transmitting hateful information across the internet at a very high speed. The DSA, which obliges the platforms to create crisis protocols and authorizes national authorities and the EU to impose sanctions upon platforms should this situation occur, is offering a way that may protect users from being exposed to defamatory inaccurate information that is likely to incite to hatred and violence. The EU authorities were quick to respond to Grok's dissemination of hate speech and to ask X to discuss why this occurred.[545]

The DSA's success in fostering collaboration between platform representatives and civil society actors to collaborate on methods of content

[541] Kate Conger, *Elon Musk's Grok Chatbot Shares Antisemitic Posts on X*, N.Y. Times (July 8 2025), https://www.nytimes.com/2025/07/08/technology/grok-antisemitism-ai-x.html [https://perma.cc/48YQ-YXEK].

[542] Paresh Dave, *Elon Musk Unveils Grok 4 amid Controversy over Chatbot's Antisemitic Posts*, Wired (July 10, 2025), https://www.wired.com/story/grok-4-elon-musk-xai-antisemitic-posts/#intcid=_wired-article-bottom-recirc_c5ee0d61-16d2-49e8-9a32-1e3c4d524997_roberta-similarity1 [https://perma.cc/B6WD-AEX7].

[543] Kate Conger, *Grok Chatbot Mirrored X Users' 'Extremist Views' in Antisemitic Posts, xAI Says*, N.Y. Times (July 12 2025), https://www.nytimes.com/2025/07/12/technology/x-ai-grok-antisemitism.html [https://perma.cc/QC9G-VG7F].

[544] *See* Noam Kolt, *Algorithmic Black Swans*, 101 Wash. U. L. Rev. 1177, 1200 (2024).

[545] Elisa Gkritsi, *EU Calls in X to Talk Grok After Antisemitic Outbursts*, Politico (July 14, 2025), https://www.politico.eu/article/european-commission-x-artificial-intelligence-chatbot-grok-antisemitism/ [https://perma.cc/HQA5-E42Q].

moderation, with an emphasis on the need for human moderators to be well-trained and aware of local contexts, can also be very beneficial for detecting hate speech .[546]

B. Using AI To Filter Hate Speech And Misinformation.

Content moderation as practiced by major platforms themselves is highly problematic. Scholars are alert to the dangers of imposing excessive limits to freedom of expression by using AI in content moderation.[547] Moderation technology is not accurate and it is not sure that it has the level of nuance that is required to detect hate speech. Its use may be overinclusive and underinclusive at the same time. A 2024 study, commissioned by the NGO The Future of Free Speech, focusing on Germany, Sweden, and France, showed that the use of AI in content moderation returns too many false positives, even for those legal systems that limit hate speech.[548] The study brought together academics who examined whether speech that had been removed by Facebook and YouTube in Germany, Sweden, and France would fall under the scope of the applicable legislation in each one of these countries.[549] It involved using researchers to double-check speech that had been removed against the relevant legislation limiting hate speech. They found that only 10%–15% of the speech that had been removed fell under the scope of the relevant legislation.[550]

The challenges in detecting hate speech and misinformation threaten the democratic character of online communication itself. Distinguishing truth from falsity is particularly challenging for democracy because flagging and filtering content implies serious disempowerment for speakers and users of online information.[551] The state of the art lies in automated detection systems using large language models (LLMs). These are computer models that can be purposed to

---

[546] Eur. Bd. For Digit. Services, *supra* note 24, at 34.

[547] Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 528, 551 (2022).

[548] The Future of Free Speech, Preventing "Torrents of Hate" or Stifling Free Expression Online? 6 (May 2024), https://futurefreespeech.org/wp-content/uploads/2024/05/Preventing-Torrents-of-Hate-or-Stifling-Free-Expression-Online-The-Future-of-Free-Speech.pdf [https://perma.cc/M6LU-Y7VS].

[549] *Id.* The author of this article was among the academics involved in the study on France.

[550] *Id.*

[551] Sille O. Søe, *Algorithmic Detection of Misinformation and Disinformation: Gricean Perspectives*, 74 J. Documentation 309, 331 (2018).

recognize and filter or flag certain content that contains false information.[552] LLMs use datasets which are collections of works, e.g. news articles, which have been labeled as "false" or "true" to check information.[553] Once models are built and their performance evaluated, they can be implemented for real world use, where it is unclear how they decide on their labeling.[554] A type of these models are the natural language processing models, use of which can lead to problematic situations to the extent that they pick up cultural biases about gender, race, ethnicity, and religion.[555] This means that it is important to ensure that datasets must train models to do what they are intended to do and to avoid the accidental propagation of undesirable patterns in the data. Some scientists argue that linguistic data will always include pre-existing biases.[556] Those gender-based biases are due to word embedding.[557] Embedding consists of providing a dictionary for computer programs that use words. Words are associated with semantic meanings and with other words. These then create arithmetic models that capture a variety of relationships which may reflect sexist and other attitudes.[558] Computer programming has evolved towards debiasing algorithms. Nevertheless, this debiasing does not fully eliminate the presence of bias. In some attempts, 6% of new instances of word embedding were still judged as reproducing stereotypes.[559] All these difficulties complicate the task of detecting misinformation.

---

[552] *See generally* Lynn E.M. de Rijk, *Who Gets to Decide What Is True? The Free Speech Problem and the Importance of Datasets to False Information Detection Models* (2022) (MRes Linguistics and Communication Sciences Thesis, Radbound University, unpublished study on file with author). I am grateful to Lynn for the references to literature in Linguistic and Communication Sciences in this paper.

[553] *Id.* at 4.

[554] *Id.*

[555] Tolga Bolukbasi et al., *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*, arXiv, at 1 (July 21, 2016), https://arxiv.org/pdf/1607.06520 [https://perma.cc/42JE-H73K]; Robyn Speer, *Conceptnet numberbatch 17.04: Better, Less-Stereotyped Word Vectors*, ConceptNet Blog (Apr. 24, 2017), http://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/ [https://perma.cc/2LB6-2LEB].

[556] Emily M. Bender & Batya Friedman, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 Transactions Ass'n for Computational Linguistics 587, 604 (2018).

[557] *See* Speer, *supra* note 555.

[558] *See* Speer, *supra* note 555.

[559] *See* Bolukbasi, *supra* note 555, at 8.

Furthermore, researchers use two approaches to detect false information: one based on content, and one based on context.[560] The approaches based on content focus either on style or on knowledge. In the style-based approach, several elements in a publication are used to classify it as misinformation, such as textual or linguistic features, images, and the relation between headline and article.[561] The knowledge-based models presuppose a list of facts made available by their creators, against which they check arguments in a text under consideration.[562] A context approach uses elements in the context surrounding a post, such as metadata, user behavior, and how the post is propagated in the social media network (spreading patterns, reliability of the source).[563] All these approaches indicate that there is a great degree of subjectivity that is at stake. The model designer has virtually unlimited freedom to create a model that limits speech.

These models are created with some underlying assumptions that inform the data collection and labeling.[564] Fact-checking is done by journalists and researchers collecting data from sources that are deemed reliable or unreliable.[565] To evaluate the accuracy of information, the models use mainstream online newspapers and data labeled by journalists. Labelling is done by experts or researchers making use of journalist-managed sources.[566] Several difficulties arise in this process, in part because there are many cases which fall along the continuum between truth and falsity. The case of satire is particularly interesting because some of these models consider it as "false information," while others don't. The methods of data labeling are not always clear and the entire process depends entirely on the subjectivity of the evaluator. Although journalists are trained in fact-checking, their judgment is also subjective. Researchers by themselves do not seem to be able to trace the line between false information and the edge cases of misinformation.[567] AI models are likely to limit too much speech, unless human moderators also intervene. Researchers suggest that the solution to the problem of developing a false

[560] *See* de Rijk, *supra* note 552, at 5.
[561] *See id.*
[562] *Id.*
[563] *Id.*
[564] *Id.* at 11.
[565] *Id.*
[566] *Id.* at 21.
[567] *Id.* at 23.

information detection model should focus on where the model will be implemented to reduce the risk of false positives.[568]

The latest research on LLMs indicates that they are capable of impressive results. They are becoming better at capturing misinformation. They are able to "handle images, access timely and credible knowledge on the web, retrieve evidence that refutes or contextualizes the given content that may or may not be misinformation, and generate clear explanations with accurate and trustworthy references."[569] They even respond automatically to potential misinformation.[570] They can determine the credibility of web pages by looking up their publishers' factuality and bias ratings.[571] They can extract text from each web page they refer to as evidence to verify if misinformation exists.[572] The quality of their responses is estimated to be 8-10% better than the crowd-sourced misinformation analyses on X.[573] Nevertheless, they still are not able to recognize misinformation that is partially correct and even factual but may be misleading through the tactics of its use and across domains, and they are ill equipped to identify misinformation on social media.[574] Addressing such misinformation requires understanding content that is multimodal along with its context.[575] Furthermore, any errors LLMs make may actually contribute to misinformation. A study found that even when an LLM expressed uncertainty about the accuracy of false headlines, participants were still inclined to believe and share them.[576]

Other studies have shown that as promising as they are, LLMs present both opportunities and challenges.[577] They are prone to hallucinations may well create unreliable information themselves. They may well be used by malicious

---

[568] *Id.* at 24.

[569] Xinyi Zhou et al., *Correcting Misinformation on Social Media with a Large Language Model*, arXiv, at 3 (Jan. 11, 2026), https://doi.org/10.48550/arXiv.2403.11169 [https://perma.cc/59DZ-3CM4].

[570] *Id.*

[571] *Id.*

[572] *Id.*

[573] *Id.* at 5.

[574] *Id.* at 2, 7.

[575] *Id.* at 7.

[576] Matthew R. DeVerna et al., *Fact-Checking Information from Large Language Models Can Decrease Headline Discernment*, PNAS, at 6 (Dec. 4, 2024), https://www.pnas.org/doi/pdf/10.1073/pnas.2322823121 [https://perma.cc/4MF4-EYZR].

[577] Isabelle Augenstein et al., *Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking*, Nat. Mach. Intel. 852, 852 (2024).

disinformation actors to generate information that will be then used as training data for other LLMs.[578] They can be used by fake accounts to spread hateful and manipulative content, altering important parts of public opinion. Furthermore, generative AI tools can generate infinite variations of misinformation claims that may reach an enormous number of users, undermining efforts to fact-check.[579] State-of-the-art tools are not able to distinguish between legitimate Twitter/X accounts and those managed by ChatGPT. LLMs may be used in fact-checking only when their advantages are balanced with responsible and ethical practices.[580] LLMs may only support human moderators when they have first fact-checked claims by identifying sections of documents "that repeat a previously fact-checked claim or that make a claim semantically equivalent to a previously verified one."[581] This means that it is not possible to dispense with human content moderators altogether, biased as they may be.[582] It is also promising that as noted above, at least 4 Justices of the US Supreme Court are skeptical about extending expressive rights to the use of AI by social media platforms.[583]

Facebook's latest policy announcement indicates that they will focus automated filters only on illegal and high-severity violations.[584] This policy is promising and it will likely reduce the number of false positives. It also means that the differences in understanding illegality in the area of hate speech between the EU and the US will manifest themselves here as well. Because the DSA captures more speech than what is considered to be illegal in the U.S., e.g. hate speech, this will likely lead to a division of the internet as far as the definition of hateful content is concerned.

C. Alternative Solutions: "Immunizing And Empowering End Users"

Any solution in the area of misinformation must involve raising awareness. Studies have shown that raising awareness serves the role of "immunizing"

---

[578] *Id.* at 854, 856.
[579] *Id.* at 856.
[580] *Id.* at 858.
[581] *Id.*
[582] Mahi Kolla et al., *LLM-Mod: Can Large Language Models Assist Content Moderation?*, *in* Extended Abstracts of the CHI Conference on Human Factors in Computing Systems 1, 2 (2024).
[583] *See supra* Part III.B.1.
[584] *See* Hendrix, *supra* note 469 and accompanying text.

users.[585] One of the most difficult questions that arises in the area of hate speech is how past experience of individuals and societies affects how people perceive hate speech. Hate speech appeals to emotions. It is also addressed from a position of authority. How would humans judge an utterance as hateful and what aspects can we use to decide? It is important for research to devise ways of getting society involved in making these decisions. Facebook appointed content moderators with linguistic and cultural expertise.[586] Another solution is to enhance the role of NGOs which engage in direct interaction and communication in an attempt to change online users' attitudes towards hate speech. NGOs like *#IAmHere* are already interjecting themselves in the public debate and trying to affect the conversations occurring online which are perceived as hateful or demeaning by them.[587] The French legislation offering civic training in schools for minors on these issues is an important step in this direction.[588] The Reports that the European Board for Digital Services is obliged to draft on the application of the DSA is also an important opportunity to bring together civil society organisations and representatives by the platforms to reflect on the standards of content moderation policies.[589]

Another solution may come from the fact that receiving technology has also improved significantly. The receivers of information online can nowadays affect the content of information that is reaching them. This means that another option would be to suggest government regulation that requires large social media platforms to make available "middleware" companies which provide content moderation and recommendation services.[590] The receivers could use different companies to perform these tasks. These middleware options can allow users to delegate curation to providers of their choosing, by subscribing to providers of their choice. They can choose services they trust to align with their interests.[591] Bluesky, for instance, has

[585] Nico Grant & Tiffany Hsu, *Google Finds 'Inoculating' People Against Misinformation Helps Blunt Its Power*, N.Y. Times (Aug. 24, 2022), https://www.nytimes.com/2022/08/24/technology/google-search-misinformation.html [https://perma.cc/3J6R-66SM].

[586] *See How Review Teams Work*, Meta (Nov. 12, 2024), https://transparency.meta.com/enforcement/detecting-violations/how-review-teams-work/ [https://perma.cc/SV9D-SLN5].

[587] *See* Jessica Bateman, *'#IAmHere': The People Trying to Make Facebook a Nicer Place*, BBC News (June 9, 2019), https://www.bbc.com/news/blogs-trending-48462190 [https://perma.cc/48DA-KDSJ].

[588] *See supra* Part II.B.

[589] Eur. Bd. For Digit. Services, *supra* note 24.

[590] Shaping the Future of Social Media with Middleware 8, 10 (Luke Hogg & Renée DiResta eds., 2024); *see also* Balkin, *supra* note 71, at 149–50.

[591] Shaping the Future of Social Media with Middleware, *supra* note 590, at 14.

already given its users the ability to label content and prevent it from appearing in their feed.[592] The DSA encourages the use of middleware by requiring transparency in algorithmic processes and enhancing user control of the feed they receive.[593]

Furthermore, users can modify the content they receive in ways that reduce its hateful or insulting impact. It is possible to apply smart filters that reformulate hate speech or replace it with content that approximates its semantic value. Existing technology offers paraphrasing capabilities through machine-learning-based systems currently under development in text AI and natural language processing.[594] Technology enables solutions that are not binary: users can alter what they see, while speakers' expression is not entirely restricted. In such cases, platforms can claim to have removed a significant quantity of harmful content (e.g., 70%), while users retain the option to disable these filters. It is crucial to explore the philosophical and epistemological implications of this practice and to determine the most appropriate legal response. The use of this technology for moderating hate speech offers significant advantages, potentially creating a situation where platforms no longer need to make direct decisions about speech moderation. However, broader implementation of this technology on online platforms raises several concerns. Most notably, speakers remain unaware of the modifications made to their utterances. This technological development raises important questions regarding the protection of speakers' autonomy and self-definition.

Under International Human Rights Law, speech may be limited only when the principle of proportionality is respected.[595] This principle requires that any restriction on speech must be necessary, appropriate, and not excessive in relation

[592] *See Blueskey User FAQ*, Bluesky (May 19, 2023), https://bsky.social/about/blog/5-19-2023-user-faq [https://perma.cc/AW37-QPXX].

[593] *See supra* Part II.A.3.

[594] Jianing Zhou & Suma Bhat, *Paraphrase Generation: A Survey of the State of the Art*, *in* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 5075, 5075–86 (2021).

[595] European Convention on Human Rights, art. 10 (Nov. 4, 1950) ("Freedom of expression 1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises. 2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others,

to the legitimate aim pursued. The question that emerges is whether altering speech through automated filters constitutes a proportionate response to preventing harm to others. Freedom of expression protects several fundamental interests and core values.[596] These include individual autonomy, the pursuit of truth through open debate, and democratic self-governance.[597] Free speech has been characterized as "the most human right" due to its crucial role in conceptualizing and defending all other human rights.[598] When a person's speech is altered without their knowledge or consent, all these interests and values are seriously compromised. The use of automated paraphrasing filters raises significant concerns that warrant further research. While platforms might benefit from reduced moderation burdens, such technology fundamentally alters a speaker's expression without their awareness or ability to respond. The Digital Services Act (DSA) already offers an alternative solution: notice and action mechanisms that require platforms to investigate user reports of hateful or misinformative speech.[599] This approach may be preferable to automated paraphrasing because it preserves speaker autonomy while still addressing harmful content. Unlike automated filters, notice and action mechanisms allow speakers to know when their content is challenged and provide them an opportunity to defend their expression—a procedural safeguard essential to respecting freedom of speech.

Conclusion

The two systems of social media regulation are offering different approaches. The DSA seems to be offering a more robust system of protection of the users of social media platforms than the legal regime that exists in the US. The DSA reflects a conception of the government dominant in Europe according to which it is legitimate for the state to address social power imbalances. The social media platforms play an important role in defining the infrastructure of the public sphere and in enabling citizens to express themselves. If social media platforms have emerged as strong social actors that are threatening the expressive rights of the average citizen, then the government has legitimacy to protect the expressive

for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.").

[596] Joshua Cohen, *Freedom of Expression*, *in* 22 Philosophy & Public Affairs 207, 211 (1993).

[597] *See generally* Eric Heinze, Hate Speech and Democratic Citizenship (2016).

[598] *See generally* Eric Heinze, The Most Human Right: Why Free Speech is Everything (2022).

[599] *See supra* Part II.A.

rights of the citizens over those of the platforms themselves. The doctrine of the "horizontal effect" is the legal translation of this conception of the role of the government. It means that the constitutional rights of the users of social media platforms may be affirmed against the platforms themselves. Users of social media platforms may refer to their constitutional rights to free speech against social media platforms within EU member states. The European Commission was motivated by a similar conception to enact legislation to protect the procedural rights of the users of social media platforms to challenge limits to their speech imposed by the platforms.[600] The Commission sees as the "the problem drivers" which necessitate its regulation the fact that "private companies make fundamental decisions with significant impact on users and their rights."[601] It considers the way the companies are applying their terms of service and community standards as having no "checks and balances" and as being "opaque."[602] It is concerned that the tools the platforms are using for content moderation "are not consistently accurate" and that they lead to "overenforcement."[603] The rationale the Commission proposes to regulate social media platforms reflects the doctrine of the horizontal effect in the protection of the rights included in the EU Charter. The DSA enhances the procedural rights the users have across the EU within each EU member state. Furthermore, the DSA makes enforceable against social media platforms that offer services within the EU the legislation against hate speech that exists already within EU member states.

The US legal system, which relies on the state action doctrine, recognizes constitutional protection only for the free speech rights of the social media platforms themselves. This leads to a reluctance in developing regulation to protect users' free speech rights when these are limited by social media platforms. It also means that the platforms are allowed to decide on their own whether they will limit hate speech. The constitutional freedom of speech rights of the users of social media platforms do not apply against the platforms due to the state action doctrine, which means that these rights only cover users' ability to choose which social media platform to use and to decide whether they prefer a moderated or a non-moderated platform. The result in the US is that strong economic and political actors are able to create their own platforms and to ensure access to those who agree with them,

[600] Eur. Comm'n, *supra* note 4, at 26–27.
[601] *Id.* at 25.
[602] *Id.* at 25.
[603] *Id.* at 25.

while the Europeans are willing to regulate existing strong social media platforms to ensure everyone has access to them. The European system aims to guarantee pluralism within each social media platform – except for hate speech. In the area of hate speech, national legislation and standards will apply as informed by the case law of the ECtHR. This means that a division of the internet is likely to emerge even between EU member states as regards the standards of content moderation. A variation in the same standards is likely to occur between Europe and the US if platforms scale back their content moderation there.

The DSA is likely to have a global effect thanks to the regulatory gap that exists in the US. Recent declarations by social media platforms CEOs indicate some scaling back of content moderation in the US, while platforms are also maintaining several of their existing practices.[604] Platforms are likely to continue engaging in content moderation that meets the DSA requirements even where they have no legal obligation to do so, such as in the US. This has several advantages and disadvantages. Users based in the US will also be protected against hate speech and misinformation. However, they will not have the same procedural rights to challenge limits to their speech that users have within the EU. Whether large social media platforms are likely to abide by the DSA outside of the EU depends on whether they are interested in the reputational benefits that accompany responsible business practices. The fact that they maintain their global content moderation teams suggests they do.[605] In general, the movement towards responsible business practices seems to be gaining ground. If governments do not regulate platforms to protect users' procedural rights, platforms should self-regulate by providing avenues for users to complain similar to those that the DSA affords European users. Europeans are likely to experience more limits on their speech, due to legal regimes in Europe that restrict speech more extensively than in the US, but they also have more legal avenues to challenge such restrictions.

The uncertainty regarding the impact of algorithmic prioritization on human thinking has led several scholars to invoke the framework of militant democracy to justify regulating social media platforms.[606] These concerns are significant given

[604] *See supra* Part IV.A.

[605] *See, e.g.*, *Transparency Center*, Meta, https://transparency.meta.com [https://perma.cc/TX7T-J6YA] (last visited April 1, 2026) (information on global content moderation teams).

[606] *See generally* Netanel, *supra* note 13; Haupt, *supra* note 37.

the enormous economic and social power these platforms exercise over the public sphere, defining the rules of public debate. The EU approach certainly reflects a militant democracy mentality. It also expresses a greater willingness than the US to redistribute speech rights in order to ensure that everyone enjoys equally effective free speech, to the extent possible.[607] Nevertheless, there are still several issues worth exploring to ensure that the DSA is applied in a way that does not excessively limit personal liberties. There is always the danger of erring on the side of under-protecting speech in the enforcement of the DSA.

Although the dangers of misinformation and exposure to hate speech might not be acute for large parts of the population, who seem to receive information from multiple sources and not from social media platforms alone, there are some parts of the population that are exposed to the dangers of the echo-chambers.[608] Given that the most politicized and politically active parts of the population are likely to be exposed to discourses they are not willing to verify, it is important to conduct further research in order to investigate the impact of hate speech online. Some regulation that recognizes access rights to large social media platforms may also alleviate a feeling of alienation which is likely to lead to radicalization.

That said, both systems of regulation need further research in order to redefine their priorities in relation to a constantly-evolving, state-of-the-art of the technology used by very large social media platforms. It is important to investigate the challenges that regulating online extreme speech poses for governments and private actors (online platforms). It is important also to focus on the transnational enforcement of the DSA and how it affects the standards of protection of online users globally. It is necessary to recommend a balance between under-enforcement and over-enforcement of limits to hate speech. Research is needed to enlighten what should count as hateful, violent, dangerous, offensive, or defamatory expression online. That task will include evaluations of the positive legal obligations borne by governments under the DSA to counter hate speech and disinformation. It is important to investigate the regulation of extreme speech from dual perspectives, including the consequences that exposure to extreme speech can have upon vulnerable sections of the population, and the consequences that limiting speech can have upon the users whose freedoms are limited.

---

[607] *See* Tourkochoriti, *supra* note 15, at 137.

[608] *See generally* Guess et al., *supra* note 517.