The Digital Services Act and the EU as the Global Regulator of the Internet

Ioanna Tourkochoriti*

Abstract

This Essay discusses the Digital Services Act (DSA), the new regulation enacted by the EU to combat hate speech and misinformation online, focusing on the major challenges its application will entail. However sophisticated the DSA might be, major technological challenges to detecting hate speech and misinformation online necessitate further research in implementing the DSA. This Essay also discusses potential conflicts with U.S. law that may arise in the application of the DSA. The gap in regulating the platforms in the U.S. has meant that the platforms adapt to the most stringent standards of regulation existing elsewhere. In 2016, the EU agreed with Facebook, Microsoft, Twitter, and YouTube on a code of conduct countering hate speech online. As part of this code, the platforms agreed to rules or Community Guidelines and to practice content moderation in conformity with them. The DSA builds on the content moderation system by enhancing the internal complaint-handling systems the platforms maintain. In the meantime, some states in the U.S., namely Texas and Florida, enacted legislation prohibiting the platforms from engaging in viewpoint discrimination. Two federal courts of appeals that have examined the constitutionality of these statutes under the First Amendment are split in their rulings. This Essay discusses the implications for the platforms' content moderation practices depending on which ruling will be upheld.

---

* Associate Professor of Law, Baltimore Law School. I am grateful to Martha Larson (Professor in Artificial Intelligence, Machine Learning, Language & Communication at Radboud University in the Netherlands) for the reflections and bibliographical suggestions related to the state of the art in artificial intelligence and machine learning in detecting misinformation. I was able to elaborate legal commentary of the state of the art following fascinating conversations with Martha. I am also grateful to Eric Heinze, Jörn Reinhardt, Kristian Skagen Ekeli, Jan-Willem van Prooijen, and the participants of the symposium organized by the *Chicago Journal of International Law* on free speech for interesting discussions. Special thanks to Tori Keller, Christian Pierre-Canel, and Mike Antosiewicz, editors with the *Chicago Journal of International Law*, for excellent editing suggestions.

Table of Contents

I. INTRODUCTION

Extreme speech has become a major source of mass unrest throughout the world. Social media platforms magnify the conflicts that lie latent within many societies, which are often further fueled by powerful political actors. Similarly, widespread misinformation during the COVID-19 pandemic and the perceptions of these platforms' inadequate responses led the European Union (EU) to pass the 2022 Digital Services Act (DSA) to combat misinformation and extremist speech.[1] The EU also strengthened its Code of Practice on Disinformation.[2] Although these are important developments toward regulating hate speech online, the legislation will be difficult to implement. There are major technological challenges in monitoring online hate speech that necessitate further research. Furthermore, depending on legal developments in the United States (U.S.), the EU's new legal regime might lead to a conflict with U.S. law, which will complicate platforms' content moderation processes.

The DSA responds to concerns expressed about the shortcomings of the system of content moderation currently applied by major social media platforms. Although it offers a sophisticated regulatory model to combat hate speech and misinformation, further research is required in several areas related to detecting such content. The state of the relevant detection technologies raises several concerns, which relate to the difficulties in the current artificial intelligence (AI) models that have been developed to detect hate speech and misinformation. Research is also needed to determine the impact of exposure to hate speech online.

The U.S. offers extended protection for freedom of speech. In many European states, however, it is legitimate for the government to limit abuse of the same freedom to protect citizens from harm caused by hate speech. It is also legitimate to limit fake news. In the U.S., the sparse regulation of speech at the federal level has left a gap to be filled by states and civil society actors. Florida and Texas enacted legislation to limit online platforms' discretion to refuse to host others' speech.[3] More frequently, contractual terms limit speech rights in several private institutions in the U.S. The major U.S.-based social media companies (Facebook and Twitter) have created deontology committees to limit hate speech in the U.S. under pressure from the EU. Questions emerged recently among academics and political actors in the EU on whether these platforms are limiting too much speech as private actors. The concern emerged that the platforms may

---

1 Digital Services Act, 2022 O.J. (L 277) 1 [hereinafter DSA].

2 *2022 Strengthened Code of Practice on Disinformation*, EUR. COMM'N (June 16, 2022), https://perma.cc/5SMQ-ZGYM.

3 Fla. SB 7072 (2021); H.B. 20, 87th Leg. 2nd Special Sess. (TX. 2021).

be limiting even more speech than what is acceptable in Europe, where limits to hate speech by the government are acceptable.[4]

Courts have the last word in Europe about whether social media users' freedoms will be adequately protected. Citizens can bring claims before courts alleging violations of their constitutional rights by the platforms. The doctrine of horizontal effect of constitutional rights, dominant in European states, enables them to do so. 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 society.[5] The constitutionally protected right to freedom of expression justifies government intervention to ensure its protection against civil society actors too. In several EU member states, the DSA will supersede existing national legislation regulating hate speech and fake news online. France has enacted such legislation, the constitutionality of which was examined by the Constitutional Council.[6] Germany has also enacted legislation generating significant case law in this area.[7] The DSA will trump even U.S. free speech law insofar as the major companies are transnational and must therefore follow European rules as well as American law. However, depending on future court decisions, a conflict may emerge between U.S. law and the DSA. Should this conflict emerge, content moderation may become challenging for the platforms, as they will need to maintain different moderation standards in the U.S. and in the EU.

---

4 *See* Ioanna Tourkochoriti, *Should Hate Speech Be Protected? Group Defamation, Party Bans, Holocaust Denial and the Divide Between Europe and the United States*, 45 COLUM. HUM. RTS. L. REV. 552 (2014).

5 Stephen Gardbaum, *The "Horizontal Effect" of Constitutional Rights*, 102 MICH. L. REV. 387, 388 (2003).

6 For hate speech in France, see Loi 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet [Law 2020-766 of June 24, 2020 on Combatting Hate Speech Online], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], June 24, 2020, p. 1, https://perma.cc/GMR9-DKDS. *See also* Décision 2020-801 DC du 18 juin 2020 [Decision 2020-801 of June 18, 2020], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], June 24, 2020, p. 5. For misinformation during electoral political campaigns in France, see Loi 2018-1202 du 22 Décembre 2018 relative à la lutte contre la manipulation de l'information [Law 2018-1202 of December 22, 2018 on Combatting the Manipulation of Information], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE] Dec. 23, 2018, p. 3, art. 1 (modifying art. L. 163-2.-I. of the Electoral Code), https://perma.cc/849X-ZQ59; Décision 2018-773 DC du 20 décembre 2018 [Decision 2020-773 of December 20, 2018], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], Dec. 23, 2018, p. 79.

7 For hate speech in Germany, see Gesetz zur Verbesserung der Rechtsdurchsetzung in den sozialen Netzwerken [Act to Improve Enforcement of Law in the Social Networks], BGBl. I, S. 3352 of Sept. 1, 2017, (Netzwerkdurchsetzungsgesetz, "NetzDG"), https://perma.cc/KRM9-THKD. *See also* Jörn Reinhardt: "„Fake News", „Infox", Trollfabriken. Über den Umgang mit Desinformationen in den sozialen Medien", 225/226 Vorgänge 97–108 (2019); Claudia Haupt, *Regulating Speech Online: Free Speech Values in Constitutional Frames*, 99 WASH. U. L. REV., 751–86 (2021).

Social media companies are required to modify their operational practices to abide by the EU's Code of Conduct Countering Illegal Hate Speech Online.[8] Specifically, platforms are required to offer enhanced internal complaint-handling mechanisms. They must also meet several procedural requirements in investigating complaints. They must issue prior warnings before removing users.

The DSA applies to providers of intermediary services irrespective of their place of establishment or residence "in so far as they provide services in the Union, as evidenced by a substantial connection to the Union."[9] Social media companies modify their behavior to meet the most stringent legal regimes in order to be able to offer their services everywhere. So, by engaging in regional regulation of online speech, the EU is becoming a global regulator of the internet.

Part II of this Essay discusses the role platforms play in defining the public sphere today and the implications of that role for government regulation. Part III presents how the DSA complements existing codes of practice in countering illegal hate speech. Part IV investigates the challenges that regulating online extreme speech and misinformation pose for governments and platforms. These challenges relate to the state of the relevant detection technologies. Part V focuses on transnational enforcement of the Act and discusses possible areas of conflict with U.S. law. Further research is needed to establish guidelines for establishing what counts as hateful, violent, dangerous, offensive, or defamatory expression, insofar as these forms of expression are subject to DSA regulation.

II. THE CONTEMPORARY PUBLIC SPHERE AND ITS PROBLEMS

Today, online platforms largely define the public sphere and the opportunities for citizens both to express themselves and access the views of others. Traditionally, governments were considered the source of danger for expressive freedoms, but today, the practices of privately held, multinational corporations also pose a great threat. A transatlantic comparison illustrates how governments respond to this new challenge. In Europe, the doctrine of horizontal effect of human rights authorizes the state to intervene and regulate platforms.[10] That doctrine also authorizes the state to enforce constitutionally protected (or other higher-order) rights against private parties as well.[11]

By contrast, in the U.S., the state action doctrine means that the protection of constitutional rights applies only against government actors.[12] The U.S.

---

8 *Code of Conduct on Countering Illegal Hate Speech Online*, EUR. COMM'N (June 30, 2016) [hereinafter *EU Code of Conduct*], https://perma.cc/72CG-NDMQ.

9 DSA pmbl. ¶ 7.

10 *See* Gardbaum, *supra* note 5.

11 *Id.*

12 *See generally* Mark Tushnet, *The Issue of State Action/Horizontal Effect in Comparative Constitutional Law*, 1 INT'L J. CONST. L. 79 (2003); Charles L. Black, Jr., *Foreword: State Action, Equal Protection, and*

Constitution does not provide protection against private actors. On the basis of this doctrine, citizens are not able to enforce through courts the protection of their constitutional rights against social media platforms. And social media platforms' *own* right to freedom of speech covers how they allow users 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 all around the world bodies of content moderators with language and regional expertise[13]. In addition, Facebook created a private body, the Facebook Oversight Board, with authority to review content that is taken down and content that is kept up.[14]

The emergence of social media and the new challenges inherent in online communication have led many scholars to advocate for restrictions on extreme speech, even within legal systems where such limits may conflict with national constitutional obligations. If in the past many scholars in the U.S. defended free speech as a value against government intervention, more recent discourse has emerged that argues that the government should limit hate speech. Several scholars in recent years have argued that new dangers emerging in online communication and social networks necessitate government intervention to limit speech and to limit how online platforms operate. For Tim Wu, the strong protections of free speech adopted by the U.S. Supreme Court in the 20th century have become obsolete.[15] Brian Leiter has emphasized that the internet, by altering the social epistemology of societies, necessitates a reconceptualization of doctrines

---

California's Proposition, 81 HARV. L. REV. 69 (1967); Louis Michael Seidman & Marc V. Tushnet, *The State Action Paradox*, *in* REMNANTS OF BELIEF, CONTEMPORARY CONSTITUTIONAL ISSUES 49–71 (1996); Robert Glennon & John E. Novak, *A Functional Analysis of the Fourteenth Amendment 'State Action' Requirement*, 1976 SUP. CT. REV. 221 (1976).

13 *See* Spandana Singh, *Everything in Moderation Case Study: Facebook*, NEW AMERICA (July 22, 2019), https://perma.cc/Y3JK-MTSU; Alexis C. Madrigal, *Inside Facebook's Fast-Growing Content-Moderation Effort*, ATLANTIC (Feb. 7, 2018), https://perma.cc/5WLX-MRTY.

14 *See* Katie Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418 (2020).

15 Tim Wu, *Is the First Amendment Obsolete?*, KNIGHT FIRST AMEND. INST. (Sept. 1, 2017), https://perma.cc/Z4R6-TEE4 (noting that the First Amendment was elaborated in an information-free world and focused exclusively on protecting speakers from government). Wu argues that the First Amendment must be adapted to promote healthy speech environments by addressing a number of speech control techniques that have arisen due to communications technologies. First Amendment doctrine presupposed that information is scarce, that few people would be willing to invest in speaking publicly, and that listeners have abundant time to evaluate the information available to them. All these assumptions, together with the idea that the government is the main threat to the "marketplace of ideas," are now obsolete. In our information-rich world, listeners are overwhelmed with information and attentional scarcity is an important issue. Furthermore, the government is no longer the only threat to free speech. Abusive online mobs, reverse censorship through counter programming, and the use of propaganda bots are also important threats.

articulated in reference to the media societies used in the past.[16] On issues related to knowledge, any society relies upon some epistemic authorities.[17] These authorities are sustained on the basis of "second-order norms" about whom to believe.[18] The internet has contributed to an epistemic crisis that has undermined existing epistemic authorities.[19] The negative unintended consequences of this phenomenon become particularly obvious in times of crisis, like the COVID-19 pandemic.[20] In response to such consequences, President Biden set up a task force to investigate problems arising from online harassment.[21] Similar concerns inspired the regulatory regime established by the DSA.

III. A CLOSER LOOK AT THE DSA

The DSA was motivated by the need to set a standard of transparency and accountability on how major platforms moderate content and use algorithms.[22] It requires 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, address the chilling effects on speech that the current moderation practices may have, enhance users' access to information, and reinforce users' redress possibilities.[23] It recognizes 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.[24] It also recognizes that these users may be disproportionately affected by restrictions and removal measures due to unconscious or conscious biases potentially embedded in the notification systems used by individuals and replicated in automated content moderation tools used by platforms.[25] The Act creates 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.[26]

---

16 *See* Brian Leiter, *The Epistemology of the Internet and the Regulation of Speech in America*, 20 GEO. J.L. & PUB. POL'Y 903 (2022).

17 *Id.* at 3.

18 *Id.* at 4.

19 *Id.* at 13.

20 *Id.* at 14.

21 *See Readout of the White House Task Force to Address Online Harassment and Abuse Launch*, THE WHITE HOUSE (June 17, 2022), https://perma.cc/FU46-5M7P.

22 DSA pmbl. ¶ 45.

23 *Id.*

24 *Id.* at 17, 26, 29.

25 *Id.*

26 *See id.* at 7, 11–12, 15–16, 82.

The Act foresees a sophisticated mechanism for content moderation of online platforms.[27] It builds upon a previous regime that had already been elaborated by the EU in 2016, a code of conduct on countering illegal hate speech online.[28] The Code defines illegal hate speech as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin."[29] The Code was agreed to by Facebook, Microsoft, Twitter, and YouTube. It led these major platforms to modify their operations and create mechanisms of content moderation within each state where they operate. The platforms assumed the obligation to have in place rules or community guidelines and to create clear and effective processes to review notifications regarding illegal hate speech.[30] Under the Code, the platforms were obliged to review the majority of valid notifications for removal of illegal hate speech in less than twenty-four hours and remove or disable access to such content.[31] The platforms also moderated content pursuant to the Code outside of the EU. The legal gap in regulation in the U.S. meant that the platforms adapted their practices worldwide to adhere to the most stringent legal regime. This is an example of the "Brussels effect," and signals that the EU has become a global regulator of the internet.[32]

Each platform's efforts to abide by the Code are monitored annually in collaboration with a network of organizations located in several states where platforms offer their services.[33] Using a commonly agreed-upon methodology, these organizations test how platforms are implementing the commitments in the Code.[34] During the latest assessment, a total of 3,634 notifications alleging instances of hate speech were submitted to platforms since October 2021.[35]

The DSA enhances this system. It creates a protective regime for users of online media and attempts to strike a balance between protecting free speech and limiting "illegal" hate speech. To some extent, it enhances free speech by enacting a process for imposing limitations on speech as well as by creating timelines for the duration of those limitations. It enhances access to social media platforms by

---

27 *See id.* at 12–13.

28 *EU Code of Conduct*, *supra* note 8.

29 *Id.* at 1. For this definition, the code refers to the 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.

30 *EU Code of Conduct*, *supra* note 8, at 2.

31 *Id.*

32 *See generally* ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD (2020) (discussing the "Brussels effect").

33 DIDIER REYNDERS, 7TH EVALUATION OF THE CODE OF CONDUCT 5 (Nov. 2022), https://perma.cc/5TFS-XDCD.

34 *Id.*

35 *Id.* at 1.

regulating the circumstances when these platforms may exclude users. These measures aim to respond to concerns that platforms so far have intervened based on the behavior of accounts or groups and the actors or associations behind them.[36] Evelyn Douek, in her comprehensive study on the subject, noted that 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.[37] Under the DSA, platforms must have already issued a warning to users who frequently post illegal content before those users may be suspended.[38] Moreover, such suspensions may only last for a reasonable period of time.[39]

The DSA also creates an obligation to implement "notice and action mechanisms" to alert platforms to the presence of content that the notifier considers to be illegal.[40] This mechanism must make it possible to identify the specific items of information thought to be illegal.[41]

The Act further enhances current internal complaint-handling systems that some platforms maintain.[42] In addition, it foresees the possibility for out-of-court dispute settlements.[43] It provides for the creation of new national and European bodies that will oversee its application. These are composed of independent administrative authorities, including Digital Services Coordinators, which will be created within each EU member state, and a European Board for Digital Services, which will be an independent advisory group for the national bodies.[44] The DSA enhances transparency for the process by creating reporting obligations for the platforms.[45]

The Act creates additional obligations for very large online platforms to manage systemic risks,[46] which seems to respond to the warnings of academics about the need to address this issue.[47] The preamble to the DSA emphasizes that platforms' systemic risks may have a disproportionately negative impact in the EU when the number of users of a platform reaches a significant share of the Union

---

36 *See* Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526, 540 (2022).

37 *Id.*

38 DSA art. 23.

39 *Id.* art. 23.

40 *Id.* art. 16.

41 *Id.* art. 16(1).

42 *Id.* art. 17.

43 *Id.* art. 18.

44 DSA arts. 38–49.

45 *Id.* art. 24.

46 *Id.* art. 33.

47 *See* Douek, *supra* note 36, at 598.

population:[48] specifically, where the number of users exceeds a threshold of 45 million, which is equivalent to 10% of the Union population.[49] Platforms of such scale should, under the Act, bear the highest standards of due diligence.[50] Platforms are also obliged to implement reasonable, proportionate, and effective mitigation measures tailored to the specific systemic risks they identify.[51] These measures may include adapting content moderation or recommender systems, decision-making processes, the features or functioning of their services, or their terms and conditions.[52] One of the concerns that motivated this need for a systemic response is the concern that groups of accounts frequently violate platform rules.[53]

The Act authorizes member states to impose penalties for infringements by providers of intermediary services under their jurisdiction.[54] These penalties should be effective and proportionate, and they should be serious enough to dissuade violations.[55] However, the maximum amount of fines that member states may impose for a failure to comply with the DSA is 6% of the provider's annual worldwide turnover.[56]

IV. CHALLENGES TO BE ADDRESSED

Online communication raises several challenges in the area of hate speech and misinformation. These challenges threaten the very democratic character of online communication itself. Compelling as it is to regulate hate speech and misinformation, the state of the relevant detection technologies raises several concerns. The imperfections of these technologies may lead to limiting more speech than is necessary. This makes it imperative to explore alternative ways for limiting the spread of hate speech online. Further research is required in all these areas in relation to the implementation of the Act.

As mentioned earlier, the Act creates obligations for very large online platforms to manage systemic risks. One of those risk mitigation measures is the Code of Practice on Disinformation, which was strengthened by the EU in 2022.[57] The Code was elaborated by the EU in response to the fact that platforms rely on

---

48 DSA pmbl. ¶ 54.

49 *Id.*

50 *Id.*

51 *Id.* art. 35(1).

52 *Id.* pmbl. ¶ 58.

53 *See* Douek, *supra* note 36, at 540.

54 DSA art. 52.

55 *Id.* art. 52(2).

56 *Id.* art. 52(3).

57 *2022 Strengthened Code of Practice on Disinformation*, *supra* note 2.

third-party fact-checkers' judgments to guide content moderation.[58] The Code provides that the platforms 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.[59] The platforms are obliged to report on the independent fact-checkers they have used.

The Code of Practice on Disinformation also foresees that the signatories will provide details about the basic criteria they use to review information sources and disclose relevant safeguards put in place to ensure that their services are apolitical, unbiased, and independent.[60] The Code requires platforms to: inform users whose content or accounts have been subject to enforcement actions taken on the basis of violation of policies relevant to this section; provide them with the possibility to appeal the enforcement action at issue; handle complaints in a timely, diligent, transparent, and objective manner; and to reverse the action without delay when the complaint is deemed to be unfounded.[61] It provides that the platforms integrate, showcase, and consistently use fact-checkers' work in their services, processes, and content across member states.[62] Platforms commit to creating a repository of fact-checking content that will be governed by the representatives of fact-checkers.[63] The platforms commit to operate on the basis of strict ethical and transparency rules, which must comply with the requirements of instruments such as the International Fact-Checking Network Code of Principles or the proposed Code of Professional Integrity for Independent European fact-checking organizations.[64]

In the area of misinformation, the imperfections of the relevant technology may lead to limiting much more speech than is necessary. Distinguishing truth from falsity is most challenging. Flagging and filtering content may imply serious disempowerment for speakers and users of online information.[65] In the area of "false information," the state of the art technology lies in automated detection systems. These are computer models that can recognize, filter, and flag certain content that contains false information.[66] These models use datasets, which may

---

58 Douek, *supra* note 36, at 544.

59 *2022 Strengthened Code of Practice on Disinformation*, *supra* note 2, measure 21.1.

60 *Id.* measure 22.4, QRE 22.4.1.

61 *Id.* commitment 24.

62 *Id.* commitment 31.

63 *Id.* commitment 32, measure 32.2.

64 *Id.* commitment 33, measure 33.1.

65 *See* Sille Obelitz Søe, *Algorithmic Detection of Misinformation and Disinformation: Gricean Perspectives*, 74 J. DOCUMENTATION 309, 309–31 (2018).

66 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 1 (2022) (unpublished research thesis) (on file with author). I am grateful to Lynn for the references to literature in Linguistic and Communication Sciences in this paper.

include news articles that have been labeled as "false" or "true." Once models are built and their performance is evaluated, they can be implemented for real-world use, where the process for labeling information becomes less clear.[67] One type of model uses natural language processing, which can lead to problematic situations to the extent it picks up cultural biases about gender, race, ethnicity, and religion.[68] This means that it is important to ensure that datasets 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 preexisting biases.[69] Those gender-based and culture-based biases are due to word embedding, which consists of providing a sort of dictionary for computer programs. Words are associated with other words and with semantic meanings, and when these are embedded in arithmetic models, those models can capture a variety of word relationships that reflect sexist and other pernicious attitudes.[70] Computer programming has evolved toward debiasing algorithms, but nevertheless, this debiasing does not work perfectly. In some attempts to debias the original embedding of those algorithms, 6% of the new embedding was still judged as reproducing stereotypes.[71] Such difficulties complicate the task of regulating misinformation.

All these models are created with some underlying assumptions that inform the data collection and labeling.[72] Fact-checking is done by journalists and researchers collecting data from sources that are deemed reliable or unreliable.[73] To evaluate the accuracy of information, the models use mainstream online newspapers and data labeled by journalists. There are several practices for labeling. It is mostly done by experts or researchers making use of journalist-managed sources.[74] The difficulties that arise in this process relate to the fact that there are many cases in between truth and falsity. The example of satire is particularly interesting: some models consider it "false information" while others do not.[75]

---

67 *Id.* at 4.

68 Tolga Bolukbasi et al., *Man is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*, 29 ADVANCES IN NEURAL INFO. PROCESSING SYS. 1, 1–8 (2016); Robyn Speer, *ConceptNet Numberbatch 17.04: Better, Less-Stereotyped Word Vectors*, CONCEPTNET (Apr. 24, 2017), https://perma.cc/FQ9V-QF33.

69 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, 587–604 (2018).

70 Bolukbasi et al., *supra* note 68, at 8.

71 *Id.*

72 De Rijk, *supra* note 66, at 11.

73 *Id.* at 13.

74 *Id.* at 21.

75 *Id.* at 17.

The methods of data labeling are not always clear.[76] The entire process depends on the subjectivity of the evaluator.[77] Truth is often not so clearly distinguished from untruth. 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 in-between cases of misinformation.[78] Given the state of the art in detecting misinformation, strengthening the Code of Practice on Disinformation is a very important step in the right direction. The journalists entrusted with the mission to form these models must continue to receive rigorous training in professional ethics to be able to form models that are reliable to the extent possible.

Researchers suggest that the solution to the problem of developing a misinformation detection model should focus on where the model will be implemented, in order to reduce the risk of false positives.[79] They also suggest involving users in determining what should be considered false information, as well as exploring models that adapt to "the changing nature of truth."[80] The DSA attempts to include users in some respects by giving them the possibility to "flag" speech they consider hateful or misinformation.[81] Any solution in the area of misinformation should involve raising awareness, as some have found this serves the role of "immuniz[ing]" users.[82]

Similar difficulties emerge in algorithms' efforts to identify hate speech. 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.[83] False positives are very frequent, especially when algorithms are entrusted with the mission to impose limits upon speech, as has been the case throughout the pandemic.[84] This is because moderation technology is not completely accurate, and it is not certain whether hate speech detection algorithms are capable of detecting all nuances of speech.[85] European Commission reports note that the average removal rate of suspicious communication is 63.6%.[86] Any user may report a case of hateful content, and a large number of

---

76 *Id.* at 21.

77 *Id.* at 23.

78 De Rijk, *supra* note 66, at 23.

79 *Id.* at 24.

80 *Id.* at 25.

81 DSA art. 16.

82 *See* Nico Grant & Tiffany Hsu, *Google Finds 'Inoculating' People Against Misinformation Helps Blunt Its Power*, N.Y. TIMES (Aug. 24, 2022), https://perma.cc/UQ7D-MWSS.

83 Douek, *supra* note 36, at 548.

84 *Id.* at 549–50.

85 *Id.* at 569.

86 REYNDERS, *supra* note 33, at 1.

communications are removed following notifications to major platforms submitted by the "trusted flaggers," which are organizations all over Europe that already participate in online monitoring exercises.[87] There is a growing difference of treatment between general users' notifications and those sent by trusted flaggers.[88] In several instances, major social media platforms have disagreed with the notifying organizations.[89] Under the DSA, the National Digital Services Coordinators will play an important role in evaluating the platforms' decisions in enforcing national standards of hate speech. This means that local community standards will carry great weight.

Receiving technology has also improved significantly. Those receiving information online can now affect the content of information that is reaching them. Furthermore, they can edit the content that reaches them so that is not felt as hateful or insulting. For instance, it is possible to create a smart filter that reformulates hate speech or replaces it with something that approximates its semantic value.[90] Any user can use this technology during online searches. Social media platforms are not using it yet, though it is worth exploring whether it is preferable for social media platforms to use this technique instead of limiting speech because paraphrasing technology allows for solutions that are not black and white. Users have the possibility to alter what reaches them, while the speaker's expression is not limited entirely. It is extremely important to explore the philosophical and epistemological status of this practice.

The use of paraphrasing technology in the area of hate speech has some significant advantages. It may lead to a situation where platforms no longer need to make decisions about moderating speech. However, this technology, if further used on online platforms, raises several concerns. The primary problem is that the speaker does not become aware of the changes made to her utterances. This technological development raises significant questions in relation to protecting the speaker's autonomy and self-definition. The use of paraphrasing technology may get out of hand and completely distort the text of the author. The author must be protected against further uses of the paraphrased text.

Under international human rights law, speech may be regulated only if the principle of proportionality is respected. For instance, the European Convention on Human Rights foresees the factors that the European Court of Human Rights uses in its proportionality analysis for evaluating government limitations on

---

87 *Id.* at 2 (in 2022, IT companies surveyed removed 63.6% of content flagged, most of which is reported by "trusted flaggers"); DSA pmbl. ¶ 61.

88 REYNDERS, *supra* note 33, at 2.

89 *Id.*

90 *See* 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 5075–86 (Ass'n of Computational Linguistics, 2021).

speech.[91] The question that emerges here is whether, if harm to others is to be averted, the speech limitation is an appropriate and proportionate response. There are several interests and core values that are protected by freedom of expression.[92] The classical defenses for free speech emphasize the importance of exercising this liberty for individuals and societies alike.[93] Free speech has been characterized as "the most human right."[94] Democratic interests are also served by the protection of freedom of speech.[95] All these interests and values are seriously compromised when a person's speech is altered.

We need to think further about whether it is permissible for platforms to use these filters. To limit speech through moderation and inform users accordingly (as the DSA foresees) may be more preferable than to paraphrase a user's speech. Implementing paraphrasing mechanisms may involve a serious infringement of a user's speech, without them knowing about it or having the ability to defend themselves against the practice. Further research is required in this area to explore the ethical issues that arise from the prospect of using paraphrasing technology in the area of online hate speech.

Further research in social psychology is also required to evaluate the effects of exposure to hate speech online and to detect the practices of extremists once blocked from online social media platforms. NGOs in favor of eliminating limits to speech claim that blocking extremists from platforms leads them to further

---

91 European Convention on Human Rights art. 10, Sept. 3, 1953, 213 U.N.T.S. 222. Article 10 of the Convention provides:

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

92 *See generally* Joshua Cohen, *Freedom of Expression*, 22 PHIL. & PUB. AFFS. 207 (1993) (discussing the following interests: an expressive interest, an interest in obtaining information in view of finding out which is the right way to act, and an interest in obtaining information from a secure source on the conditions that are necessary for the pursuit of our goals and aspirations).

93 JOHN STUART MILL, ON LIBERTY 21–61 (John Gray ed., 1998) (offering a classical consequentialist defense for free speech).

94 *See* ERIC HEINZE, THE MOST HUMAN RIGHT: WHY FREE SPEECH IS EVERYTHING (2022).

95 Kent Greenawalt, *Free Speech Justifications*, 89 COLUM. L. REV. 119, 125–29 (1989); *see also* ERIC HEINZE, FREE SPEECH AND DEMOCRATIC CITIZENSHIP (2016).

radicalization on less moderated platforms.[96] Early research in communication and media studies indicates that bans on right-wing extremists imposed by mainstream social media platforms (Facebook and Twitter) very likely led them to "migrate" to other platforms (such as Telegram) which offer enhanced privacy and anonymity along with opportunities to gain publicity, coordinate, and mobilize.[97] The decrease in transparency in these alternative social media platforms may reduce the size of extremists' audiences yet increase the radicalization of the audience members that remain.[98] The same research suggests that outright bans might not be the best way to reduce extremists' influence; gradual bans administered to a few actors might be preferable,[99] since gradual bans cause serious coordination problems for such users.

V. POSSIBLE AREAS OF CONFLICT WITH U.S. LAW

The DSA will likely create a clash in free expression standards between the U.S. and the EU. Some U.S. states, such as Texas and Florida, have already enacted legislation prohibiting the platforms from engaging in viewpoint discrimination. In fact, Texas and Florida have enacted legislation in response to the moderation practices platforms have implemented in conformity with the EU Code of Conduct. Two federal courts of appeals have examined the constitutionality of the relevant legislation under the First Amendment and are split in their rulings.

Florida's SB 7072[100] prevents social media platforms from unfairly censoring, shadow banning, deplatforming, or applying post-prioritization algorithms to Florida candidates, users, or residents. The Eleventh Circuit found that this law violates the First Amendment rights of social media platforms.[101] For the court, the social media platforms express themselves through their content-moderation decisions.[102] The platforms are "curating" speech, and this activity is analogous to the editorial judgments of the press, which the Supreme Court has held are protected under the First Amendment.[103]

Texas' HB 20 prohibits large social media platforms from censoring speech based on speaker viewpoint.[104] The legislation provides that a social media

---

96 JACOB MCHANGAMA ET AL., THOUGHTS FOR THE DSA: CHALLENGES, IDEAS AND THE WAY FORWARD THROUGH INTERNATIONAL HUMAN RIGHTS LAW 5 (2022).

97 Aleksandra Urman & Stefan Katz, *What They Do in the Shadows: Examining the Far-Right Networks on Telegram*, 25 INFO. COMMC'N & SOC'Y 904, 904–23 (2022).

98 *Id.* at 918–19.

99 *Id.* at 919.

100 Fla. SB 7072.

101 NetChoice, LLC v. Attorney General, Florida, 34 F.4th 1196 (11th Cir. 2022).

102 *Id.* at 1212.

103 *See*, *e.g.*, Miami Herald Pub. Co. v. Tornillo, 418 U.S. 241 (1974).

104 H.B. 20.

platform may not censor 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. The Fifth Circuit held the legislation constitutional under the First Amendment.[105] The court found that the law regulates the platforms' conduct, not their speech.[106] It applied the common carrier doctrine, which allows states to impose non-discrimination obligations on communication and transportation providers that serve the public.[107] The court disagreed with the Eleventh Circuit's ruling that platform editorial discretion is protected under the First Amendment.[108] For the Fifth Circuit, the platforms use algorithms to screen out certain content, which does not involve any editorial control. Furthermore, the platforms disclaim any reputational or legal responsibility for the content they host.[109] They merely engage in viewpoint-based censorship with respect to expression they already have disseminated.[110] The court cited in this respect 47 U.S.C. § 230, which provides that the platforms "shall [not] be treated as the publisher or speaker" of content developed by other users.[111]

The Fifth Circuit's ruling, if upheld, will likely clash with the DSA on hate speech. Under HB 20, limiting hate speech is impermissible viewpoint discrimination. While the Fifth Circuit also noted that HB 20 "expressly permits the Platforms to censor any unlawful expression and certain speech that 'incites criminal activity or consists of specific threats,"[112] it may still conflict with the DSA on the definition of incitement. The DSA refers to the Code of Conduct, which platforms have agreed to abide by, to define hate speech. The Code defines it, in part, as "public incitement to violence or hatred."[113] This standard is much broader than the *Brandenburg* criteria in the U.S., according to which only speech that encourages imminent lawless action that will very likely occur may be outlawed.[114] The standard is motivated by the need to preserve social peace. In Europe, incitement is limited independently of whether it encourages imminent lawless action or not. Thus, if the common carrier doctrine is upheld, social media

---

105 NetChoice, LLC v. Paxton, 49 F.4th 439 (5th Cir. 2022).

106 *Id.* at 448.

107 *Id.*

108 *Id.* at 488.

109 *Id.* at 464.

110 *Id.*

111 Protection for Private Blocking and Screening of Offensive Material, 47 U.S.C. § 230(c)(1) (2018).

112 *NetChoice*, 49 F.4th at 452.

113 *EU Code of Conduct*, *supra* note 8, 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.

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

companies will have to modify their moderation standards in the U.S. compared to the rest of the world, a task which can be challenging. The common carrier doctrine enables access to social media platforms and enhances speech rights.[115] In this respect, it mirrors the spirit of the DSA. The application of the doctrine, though, is likely to have negative unintended consequences in relation to the platform's abilities to moderate harmful speech.

VI. CONCLUSION

Increased online speech has raised concerns about hate speech and disinformation. The EU is tackling the problem with enhanced online speech regulation through the DSA. The EU has opted in favor of a sophisticated system of government regulation of online speech. The system enhances the mechanisms of complaints available to citizens who feel excluded from social media. It also foresees an important role for Independent Administrative Authorities and the Digital Services Coordinators to oversee the content moderation platforms engage in. The major online platforms' moderation practices will be supervised by these coordinators. This supervision will contribute to defining the standards of "illegal hate speech" that the platforms will need to limit. Local community standards will carry great weight in the formation of these criteria. Although the DSA offers a sophisticated system of regulating online social media platforms, further research is needed on its implementation. There are important technological challenges that apply to companies implementing the DSA and to companies' self-regulating content. Research is needed on improving the accuracy of algorithms that limit hate speech. Research is also needed in the area of social psychology to investigate the impact that online incitement to hatred and violence may have. Furthermore, research is needed about whether alternative technologies, such as paraphrasing technology, are appropriate in the area of limiting hate speech online. The enhanced procedural requirements that the DSA imposes are balanced out by the opportunities for redress it institutes for users whose rights have been violated.

In any attempt to limit misinformation, we must be conscious of the limits of the state of technology and of the challenges that technological developments raise for democracy. In the area of misinformation, appropriate training of journalists is required as they elaborate the models that are evaluating the truth or falsity of information available online. Given the shortcomings of the state of technology, strengthening the Code of Practice on Disinformation is a very important step. Further research on improving the state of the art of misinformation technology is also necessary. The global impact the Act is likely

---

115 *See* Eugene Volokh, *Treating Social Media Platforms Like Common Carriers?*, 1 J. FREE SPEECH L. 377 (2021).

to have makes all the more compelling the need for further research on several aspects related to its implementation.

Another challenge in the application of the DSA for the platforms relates to potential conflicts with U.S. law. In the U.S., content moderation has been practiced by the platforms themselves thanks to the absence of regulation in this area. Platforms have generalized the standards of content moderation they had to develop to abide by EU requirements. Whether the platforms will be able to engage in content moderation in the U.S. will depend on future court rulings on the constitutionality of legislation against viewpoint discrimination. If the common carrier doctrine is upheld in the U.S., the platforms will need to maintain different standards of moderation in the U.S. compared to Europe. Applying the common carrier doctrine will have the positive intended consequence of protecting users against exclusion from platforms and the negative unintended consequence of limiting the platforms' ability to engage in content moderation. In the area of incitement to hatred and violence, platforms will not be able to apply the same standards of content moderation in the U.S. that they apply in Europe.