

American University Washington College of Law

Digital Commons @ American University Washington College of Law

Upper Level Writing Requirement Research
Papers

Student Works

5-31-2021

Regulating the Digital Resonance

Hassan Salman

Follow this and additional works at: https://digitalcommons.wcl.american.edu/stu_upperlevel_papers



Part of the [Internet Law Commons](#)

Regulating The Digital Resonance

Hassan Salman¹

¹ JD Candidate, Washington College of Law

Table of contents

I.	Introduction	1
II.	Section 1: How Social Media Platforms Work, and Why That's a Problem	5
a.	Part I(A): The Root of the Problem	5
b.	Part I(B): The Problem	8
i.	Quantity of Users & Mistrust	8
ii.	Effect of Excessive Moderation	9
c.	Part I(C): Legal Background	12
i.	Overview of the US & EU Regulatory History	12
ii.	US Approach to Regulating Social Media Platforms	13
iii.	EU Approach to Regulating Social Media Platforms	18
iv.	Gaps in the Frameworks	27
v.	The Facebook Oversight Board	30
III.	Section 2: Synthesizing a Solution	37
IV.	Conclusion	39
V.	Annex	40
VI.	Bibliography	41

Introduction

25 years ago Justice Kennedy postulated that, “minds are not changed in streets and parks as they once were. To an increasing degree, the more significant interchanges of ideas and shaping of public consciousness occur in mass and electronic media.”² Hundreds of millions even billions of users worldwide use platforms like Facebook and YouTube to communicate, share content, or even sell items for whatever purpose they see fit [albeit within limits determined by the platform].³ User content varies in complexity and can be as straightforward as, “I saw Tom Cruise at Trader Joe’s today” to something more complex e.g. critiquing the US immigration system by overlaying an image of caged immigrants with quotes from ‘The Irony of American History’ by Reinhold Niebuhr.⁴ User content can be nuanced and thereby makes content moderation difficult. Human content moderators spend an average of 10-15 seconds per image; however, it has been found that the time frame could be as short as 2 seconds per image.⁵ That said, due to incidents like the livestreaming of the Christchurch shootings on Facebook, lawmakers around the world began closely scrutinizing how these platforms moderate content i.e. the means and methods by which platforms monitor, filter, rank and block user generated content.⁶ This pressure has led to the need for automated moderation since human moderators generally cannot keep up with the content and usually have little training or support; nevertheless, machine assisted moderation has not reached the point where it can sufficiently

² Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech*, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1902, (January 2019)(citing *Denver Area Educ. Telecommunications Consortium, Inc. v. FCC*, 518 US 727, 802-3 (1996)(Kennedy, J., concurring)) [hereinafter Keller, (2019)]

³ Valerie C. Brannon, Cong. Research Serv., R45650, *Free Speech and the Regulation of Social Media Content*, (2019) [hereinafter Brannon (2019)]

⁴ John C. Bennet, *The Irony of American History* work by Niebuhr, Encyclopedia Britannica (2020) <https://www.britannica.com/topic/The-Irony-of-American-History>

⁵ Frederik Stjernfelt & Anne M. Lauritzen, *Facebook’s Handbook of Content Removal*. In: *Your Post has been Removed*, 134 (Springer, Cham., 2020) https://doi.org/10.1007/978-3-030-25968-6_11 (noting that one Facebook moderator working 8 hours a day from 2008 to 2013, processed 15,000 images a day) [hereinafter Stjernfelt & Lauritzen (2020)]

⁶ Hannah Bloch-Wehba, *Automation in Moderation*, 53 Cornell Int’l L.J. 41, 42-3 (2020), [hereinafter Bloch-Wehba (2020)].

understand the contextual nuances of human languages.⁷ Throughout this article, the terms ‘content moderation’ and [mandatory] takedowns will be mentioned but these are distinct notions. Content legality varies between countries and when a company is ordered to take content down, that is a mandatory takedown; however, removing content that is legal albeit undesirable (e.g. certain kinds of hate speech) to the internet service is an example of content moderation.⁸ Moderation standards vary between platforms but they generally rely on several of the same sources to identify content for removal: (1) users [to flag violating content]; (2) employed content moderators; and (3) AI (using machine learning to develop automated content filters).⁹ Few platforms rely on machine learning models to identify new content that could violate their terms and conditions, and most rely on models that have been trained to identify specific phrases and images.¹⁰ Facebook successfully used these models to tag 99% of violent and graphic content before users reported on it; however, with bullying and harassment content that success rate dropped to 16%, indicating that moderators are [for now] better suited to pick up on these types of violation.¹¹ As discussed throughout this article, platforms have an economic incentive to remove offensive and illegal content, thus increasing their reliance on automated content filters.¹² Due to this combination of economic and political pressure, content deemed as objectionable or harmful but not illegal under the platforms’ policies have been falsely removed.¹³ User generated content is also colored by societal norms, personal

⁷ Access Now, *Protecting Free Expression in the Era of Online Content Moderation: Access Now’s preliminary recommendations on content moderation and Facebook’s planned oversight board*, 5 (May 2019)

<https://www.accessnow.org/cms/assets/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf> (citing Casey Newton, *The Trauma Floor*, The Verge (Feb. 25, 2019, 8:00AM) <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>)[hereinafter Access Now (2019)]

⁸ *Id.* at 2

⁹ Clare Y. Cho & Jason A. Gallo, Cong. Research Serv., R46662, *Social media: Misinformation and Content Moderation Issues for Congress* 6 (2021)[hereinafter Cho & Gallo (2021)]

¹⁰ *Id.*

¹¹ *Id.*

¹² *Id.* at 12; Bloch-Wehba (2020) at 61

¹³ Steinfeld & Lauritzen (2020) at 129-30 (noting that despite Facebook’s claim that its approach to adult nudity and sexuality has become more nuanced, Facebook has repeatedly deleted and reinstated photos of ‘Venus from Willendorf’)

experiences, values etc., thus attempting to moderate “objectionable” content without accounting for these underlying issues has, and will continue to silence user speech.

While social media platforms have existed for some time, legal and technical developments regarding the aforementioned issues are ongoing, thus the existing literature will need to be updated in light of those developments and their effects. Existing literature¹⁴ on content moderation practices mainly focused on region centric laws, factors affecting moderation practices, and the ramifications of those practices with respect to various categories of content. However, the literature does not effectively assess how content moderation developments have led to the excessive removal of ‘objectionable content,’ and how this problem might be reined in. That said, these issues are relatively new and have incurred major developments like Europe’s Digital Services Acts Package (which has yet to be implemented); additionally, there is the Facebook Oversight Board which decided to uphold President Trump’s ban from Facebook and Instagram three weeks prior to the writing of this article.¹⁵ The ramifications of these developments have yet to be fully determined, and it is possible that the conclusions drawn in this article will lose some of their relevance.

That said, this article can serve to bridge some of the gaps left by previous studies. The focus here will be on the impact that current moderation practices have on harmful, but not necessarily illegal content, (i.e. objectionable content) and the value in developing alternative practices. This analysis will be grounded in interrelated studies on social media platforms and content moderation practices by the most influential and socio-politically relevant platforms, especially Facebook. Facebook’s broad reach and controversial reputation throughout the US

¹⁴ Statistical data, Case law, Law review articles, Tech blogs, Government research groups, think tanks as well as Academic texts and articles.

¹⁵ See generally, Annabelle Gawer, Nick Srnicek, *Online Platforms: Economic and Societal Effects*, Study Commissioned by the European Parliamentary Research Service, PE 656.336 (March 2021) [hereafter Gawer & Srnicek (2021)]; Trump Decision – FB Oversight Board. Case Decision 2021-001-FB-FBR [concerns ‘Violence and Criminal Behavior’ and ‘Safety’ but not ‘Objectionable Content’]

and EU have, and will continue to affect extensive legal and political ramifications.¹⁶ This article will specifically focus on the US and EU's approaches to regulating social media platforms like Facebook. This is partially due to their similar perspectives on free speech, as well as their readily accessible, and extensive amount of data and studies on content moderation. Further, though the US and EU have similar perspectives on the right to free speech, the EU has demonstrated a greater willingness to relegate free speech in favor of safety and economic concerns. This difference, and the subsequent results of that regulation (the Digital Services Act [DSA]) sharply contrast with the US' relatively static approach to content moderation. The DSA presented an example of how contemporary, stringent and systematically applied regulatory practices are likely to affect major platforms. These developments can also be compared against Facebook's attempt at self-regulation: The Oversight Board. The Board is an independent body developed by Facebook to help the company regulate its content moderation decisions, especially insofar as objectionable content is concerned. By synthesizing these various approaches to content moderation, and assessing their strengths and weaknesses, this article will attempt to present alternative solutions to the objectionable content problem.

In Part I(A) Facebook's [and other platforms'] content moderation methodologies, and the economic rationale behind these methods will be explained. This will help explain some of the main problems with contemporary content moderation practices. In Part I(B)(a), statistical data will be used to highlight Facebook's reach and usage throughout the US and EU, as well as the extent to which users trust Facebook's moderation practices. Following this, Part I (B)(b) will outline how, despite public mistrust, social media companies are nevertheless relied upon to keep their platforms safe and clean; however, this reliance, and sometimes over reliance will be shown to have its own drawbacks. In Part I(C) the differentiated impacts that the US & EU

¹⁶ See *infra* Part I(A)

legal approaches have had on content moderation practices will be analyzed using the existing literature. Further, Facebook's attempt at regulating its moderation practice by establishing an independent Board, a Facebook 'Supreme Court,' will add to this analysis. However, because the Board is so new its true impact on Facebook's practices is unknown, thus this part of the analysis will be comparatively short. That said, the Board has already rendered several decisions, and grounded its holdings in a myriad of established laws and rules. Moreover, because of the Board's structure and alleged influence over Facebook, its impact may likely be similar to an actual court of law. As the current literature largely precedes the Board's decisions on objectionable content, Part II will contain a synthesized analysis of the previous sections to: draw on the implications and conclusions of those sections to add to the existing literature; develop a new approach to moderating objectionable content; and highlight how future research may advance or even diminish the value of this methodology.

I. How Social Media Platforms Work, and Why That's a Problem

A. The Root of the Problem

- a. Social media platforms utilize a business model that is both a source of success and widespread criticism. The core of a social networks' business model, including and perhaps especially Facebook, is the 'attention economy.'¹⁷ This is the tactic of matching users to the most relevant information for them, and monetizing that attention through targeted ads or transactions.¹⁸ This tactic compliments a concept known as 'the network effect,' which refers to the understanding that the more users a platform has (i.e. the bigger the user network), the more useful it becomes to people who

¹⁷ Tambiama Madiega, *Digital Services Act*, Briefing Commissioned by the Directorate-General for Parliamentary Research Service, PE 689.357, 4 (March 2021) [hereinafter Madiega DSA (2021)]

¹⁸ *Id.*

have not already joined the social network.¹⁹ Thus, the better the network can capture its users' attention, the more users it can acquire and the greater its revenue becomes.²⁰ Revenue streams may vary but are mostly derived from online, targeted advertisements.²¹ Maximizing user engagement, and deriving revenue from online ads is the foundation of these platforms' business models, with US social media ad revenues totaling at \$35.6 billion in 2019.²² Additionally, though time spent on Facebook was expected to drop in 2020 to 33 minutes a day among [adult] US users, US advertisers were expected to increase their social network spending by 20.4% in 2020, and 16.9% in 2021.²³ This means that ad revenues for social networks were expected to increase from approximately \$36 billion in 2019, to \$45.53 billion in 2020 to \$50.86 billion in 2021.²⁴ That said, it is the means by which social media platforms generate this revenue that the controversy becomes apparent – the 'attention economy.' Platforms curate, categorize and rank content based on the users' interactions with the site and other factors; however, to do this effectively, platforms use non-human, code based processes to tailor and predict which content is likely to be relevant (algorithms).²⁵ These algorithms allow platforms to sort through massive amounts of user generated posts and behaviors (data) to estimate which

¹⁹ Arjun Sundararajan, *Network Effects*, New York University Stern School of Business (Dec. 23, 2020), <http://oz.stern.nyu.edu/io/network.html>

²⁰ Cho & Gallo (2021) at 3

²¹ *Id.* at 12 (finding that global online advertising accounted for 98% of Facebook's annual revenue in 2019)

²² Interactive Advert. Bureau, Internet Advertising Revenue Report: Full Year 2019 Results & Q1 2020 Revenues, https://www.iab.com/wp-content/uploads/2020/05/FY19-IAB-Internet-Ad-Revenue-Report_Final.pdf (May 2020) (defining social media as ads delivered on platforms including social networking and social gaming websites and apps, across all device types, including desktop, laptop, smartphone and tablet)

²³ Debra A. Williamson, *US Social Trends for 2020: eMarketer's Predictions for the Year Ahead*, eMarketer, <https://www.emarketer.com/content/us-social-trends-for-2020> (Jan. 15, 2020)

²⁴ *Id.*

²⁵ Brannon (2019) at 1 (citing Stuart Minor Benjamin, *Algorithms and Speech*, 161 Univ. Penn. Law Review 1445, 1448 (2013))

content (e.g. ads) would be most relevant to each user, and disseminate that content accordingly.²⁶ Therefore, more data means more refined algorithms which in turn lead to more accurate ads.²⁷ Facebook's algorithm for example can predict a user's personality with greater accuracy than their own spouse by analyzing only 300 'likes'.²⁸ One side effect of this is that echo chambers tend to develop wherein users become exposed to one kind of content instead of a range of voices and opinions.²⁹ Further, since algorithms tend to reward visceral and emotive content, this process has exacerbated the spread of misinformation, especially since the 2016 US elections [see Part I(B)(a) below].³⁰ Although, to keep users engaged and government regulators appeased platforms have been using this technology to try to prevent the dissemination of such harmful or unlawful online content before it is ever seen or distributed.³¹ The problem however, is that the incentive to remove illegal content like the Christchurch shootings has also led to the removal of unpopular but not necessarily illegal content (objectionable content).³² This incentive has broad ramifications for the entire internet speech ecosystem and, as Justice Kennedy noted, will impact the interchange of ideas, and the shaping of public consciousness.³³

²⁶ Cho & Gallo (2021) at 25

²⁷ *Id.*

²⁸ *Social Media Influences our Political Behavior and puts Pressure on our Democracies, New Report Finds*, EU Science Hub (Oct. 27, 2020) <https://ec.europa.eu/jrc/en/news/social-media-influences-our-political-behaviour-and-puts-pressure-our-democracies-new-report-finds>

²⁹ Gawer & Srnicek (2021) at 58

³⁰ *Id.*

³¹ Bloch-Wehba (2020) at 42-3

³² Steinfeld & Lauritzen (2020) at 122-23 (noting that examples of objectionable content include hate speech, graphic violence, and nudity or sexual acts)

³³ Keller (2019) at 23

B. The Problem

- a. Due to its popularity and business model, Facebook has engendered broad albeit divided public scrutiny, especially with regards to how it handles user data and moderates content. A majority of US adults use Facebook (69%), and the majority of those users (74%) visit the site on a daily basis.³⁴ In a 2018 survey, seven of eight Western European countries surveyed responded that a third or more of their adult population use social media to get their news on a daily basis.³⁵ Further, over 60% of social media consumers in all eight countries cited Facebook as the most frequently used social media source for news.³⁶ Though only 43% of US adults use Facebook as a news source, news results may vary for both EU & US users depending on how those users interact with the site.³⁷ These variations occur because Facebook's algorithms curate and present content that they predict will have a more likely chance of engaging the user i.e. divisive or provocative content.³⁸ Slides presented by an internal Facebook team to company executives in 2018 stated, "our algorithms exploit the human brain's attraction to divisiveness," and warned that the algorithms would promote "more and more divisive content in an effort to gain user attention and increase time on the platform."³⁹ Accordingly, US users

³⁴ John Gramlich, *10 Facts About Americans and Facebook*, Pew Research Center (June 1, 2021) <https://www.pewresearch.org/fact-tank/2021/06/01/facts-about-americans-and-facebook/>

³⁵ Amy Mitchell et. al., *In Western Europe, Public Attitudes Toward News Media More Divided by Populist Views Than Left-Right Ideology*, Pew Research Center (May 14, 2018)(polling users in Sweden, Germany, France, Spain, Netherlands, Denmark, UK & Italy) <https://www.pewresearch.org/journalism/2018/05/14/in-western-europe-public-attitudes-toward-news-media-more-divided-by-populist-views-than-left-right-ideology/>

³⁶ *Id.*

³⁷ Cho & Gallo (2021) at 25

³⁸ *See supra* Part I(a)

³⁹ Cho & Gallo (2021) at 10-1 (citing Jeff Horowitz & Deepa Seetharaman, *Facebook Executives Shut Down Efforts to Make the Site Less Divisive*, Wall Street Journal, (May 26, 2020, 11:38AM) <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>)

across the political spectrum largely agree that social media's effects on the US have been mostly negative; however, this belief is particularly widespread among Republicans.⁴⁰ Misgivings among 90% of republicans and right leaning independents partially stem from their belief that social media platforms are biased against conservatives and engage in political censorship.⁴¹ This belief is shared among 59% of Democrats and left leaning independents, and 73% of US adults.⁴² This variance is even greater with regards to how much each party trusts platforms to label content as inaccurate or misleading.⁴³ Similarly, Eurobarometer found that of the respondents from the 28 EU member states, only 26% trusted news and information accessed on social networks; moreover, Eurostat found that only 25% of EU citizens aged between 16 and 74 claimed to have provided personal information to social networks.⁴⁴ Thus, data handling and content moderation seem to be the major areas of concern amongst European and US social media users.

- b. Nevertheless, despite these concerns, users still rely on companies like Facebook and others to keep their platforms clear of objectionable content.

A sizable majority of US users (66%) say companies have a responsibility

⁴⁰ Brooke Auxier, *64% of Americans Say Social Media Have a Mostly Negative Effect on the Way Things are Going in the US Today*, (Oct. 15, 2020), <https://www.pewresearch.org/fact-tank/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/> (finding that 53% of Democrats and left leaning independents, and 78% of Republicans and right leaning independents believe that social media has a negative impact on the US)

⁴¹ Emily A. Vogels, Andrew Perrin and Monica Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints*, Pew Research Center (Aug. 2020) <https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/#fn-26445-1>

⁴² *Id.*

⁴³ *Id.* (finding that 73% of Democrats strongly or somewhat approved of this process, whereas 71% of Republicans at least somewhat disapproved of this process)

⁴⁴ *Flash Barometer 464: Fake News and Disinformation Online*, Survey Requested by the European Commission Directorate-General for Communications Networks, Content & Technology, Project Number 2018.2391, at 2 (2018); *Social Media: Security Concerns of Sharing Information*, Eurostat, <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20201013-1> (last visited May 27, 2021)

to remove objectionable content from their platforms.⁴⁵ However, nearly half (48%) of these users admitted to not being sure about what constitutes objectionable content, and even fewer (31%) have a great deal or fair amount of confidence in these companies to determine what objectionable content should be removed.⁴⁶ Correspondingly, the European Parliament's Committee on Citizens' Rights and Constitutional Affairs concluded that drawing the line between objectionable and unobjectionable, in good faith, is something that goes beyond the law; moreover, it is rooted in the plurality and diversity of the users, and providers of the digital environment within which the objectionable content manifests.⁴⁷ Certain content is universally objectionable (e.g. videos of animal torture) while others may be objectionable to some (e.g. blasphemy), and everything else falls somewhere within the spectrum.⁴⁸ Therefore establishing content moderation rules that work with all types of content is problematic. Despite this problem, Facebook responded to these concerns by developing a continuously developing set of Community Standards wherein it divided objectionable content into four parts: (1) Hate Speech; (2) Violent and Graphic content; (3) Adult Nudity and Sexual Activity; and (4) Sexual Solicitation (the fourth is a new category while 'Cruel & Insensitive' was subsumed into Hate Speech).⁴⁹ Though laudable, these

⁴⁵ John Laloggia, *U.S. Public Has Little Confidence in Social Media Companies to Determine Offensive Content*, Pew Research Center (July 2019) <https://www.pewresearch.org/fact-tank/2019/07/11/u-s-public-has-little-confidence-in-social-media-companies-to-determine-offensive-content/>

⁴⁶ *Id.*

⁴⁷ Giovanni Sartor & Andrea Lorregia, *The Impact of Algorithms for Online Content Filtering of Moderation: Upload Filters*, Study Commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, PE 657.101, at 55 (2020) [hereinafter Sartor & Lorregia (2020)]

⁴⁸ See Part I(C) *infra*

⁴⁹ *Writing Facebook's Rulebook*, Facebook <https://about.fb.com/news/2019/04/insidefeed-community-standards-development-process/> (last visited May 17, 2021); *Objectionable Content in Community Standards*,

efforts were, and continue to be, hampered by two major issues: (a) the nuance of human languages, and (b) over-reliance on content filtering technology.⁵⁰ One early problem that came up were statements which “merely cite or parody” the hateful statements of others.⁵¹ In response, Facebook held that users had to be clear as to their intent otherwise their posts would be removed.⁵² This is a problem when the message is, for example, meant to be ironic, and this is especially an issue when taken at a global scale.⁵³ Facebook has 3 billion active users as of August 2020 with the average user spending 50 minutes a day on FB and Instagram; additionally, in 2019 it received 76 million appeals to restore posts that were taken down, only 23% of which (1.748 million) were restored.⁵⁴ A further 284 million pieces of content were restored without appeal.⁵⁵ Facebook has only 15,000 human moderators, thus the sheer volume of data it deals with on a daily basis requires some amount of automated processes.⁵⁶ Aside from a lack of human moderators, the increased spread of online misinformation, and illegal content galvanized lawmakers around the world to closely scrutinize the content moderation process.⁵⁷ Regulators thus imposed certain regulatory strategies e.g. encouraging platforms to use content moderation technology to prevent the

Facebook, https://www.facebook.com/communitystandards/objectionable_content (last visited May 17, 2021); Steinfeld & Lauritzen (2020) at 122-23

⁵⁰ Daphne Keller, *Facebook Filters, Fundamental Rights, and the CJEU’s Glawischnig-Piesczek Ruling*, 69 GRUR International 616, 619 (2020) <https://doi.org/10.1093/grurint/ikaa047> [hereinafter Keller (2020)]

⁵¹ Steinfeld & Lauritzen (2020) at 123-24

⁵² *Id.* at 124-25

⁵³ *Id.*

⁵⁴ Cho & Gallo (2021) at 7; Gawer & Srnicek (2021) at 18

⁵⁵ Cho & Gallo (2021) at 7

⁵⁶ John Koetsier, *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*, Forbes (June 2020, 8:08PM EDT), <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=30c5a70854d0>

⁵⁷ Bloch-Wehba, (2020) at 42-3

dissemination of unlawful online content before it is ever published.⁵⁸

Through court rulings and legislation, the EU and US implemented different regulatory strategies; however, increased regulatory pressure, along with some misguided assumptions about how content filters work have led to mixed and, sometimes detrimental ramifications in the internet speech ecosystem.⁵⁹ The tools used to exploit the ‘attention economy’ has been a source of success and strife for platforms and their users; however, as seen below, government attempts to [partially] co-opt these tools have served as a double-edged sword.

C. Legal background

- a. Despite facing a similar diffusion of illegal and objectionable forms of content, the United States and Europe have almost completely diverged in their respective regulatory approaches. While the US and EU value similarly the consequence of restricting free speech, the EU nevertheless imposes specific limitations e.g. on hate speech; moreover, EU member states are permitted to restrict certain kinds of speech, including across social media.⁶⁰ The impact of this fragmented form of legislation on the European Single Market, along with the aforementioned spread of illegal and offensive content, has led to major regulatory efforts against platforms like Facebook, the full ramifications of which remain to be seen.⁶¹ Conversely, the United

⁵⁸ *Id.*

⁵⁹ Bloch-Wehba (2020) at 72-4

⁶⁰ Sartor & Lorregia (2020)(establishing that “public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin” constitutes as hate speech with no definition of hate speech in the US); Ruth Levush, *Comparative Summary, in Limits on Freedom of Expression 1, 3*(Law Libr. of Congress, 2019)(noting that Germany and the Netherlands specifically recognize limitations on speech that constitutes a denial or praise of atrocities committed during the holocaust)[hereinafter Levush (2019)]

⁶¹ Madiaga DSA (2021) at 5

States' largely restrained approach to free speech has had relatively less impact on platforms;⁶² however, as seen below, even location specific regulations can have a global impact on the way a platform functions.

- b. Under the US legal system, content moderation by private companies remains largely in the latter's hands except with regards to specific kinds of illegal content. Attempts at directly regulating social media platforms have generally been approached in three ways: (1) The First amendment; (2) Section 230 of the Communications Decency Act [CDA] (47 U.S.C. § 230); and (3) the Digital Millennium Copyright Act [DMCA].⁶³ However, due to the history, particularities and context behind these laws, each law will be addressed separately.
- c. First, the majority of internet related First Amendment cases have focused on the actions of internet companies rather than their character, thus whether or not those companies are social media platforms is irrelevant to the analysis.⁶⁴ With that, attempts at regulating platforms under the First Amendment have focused on, and failed to succeed in treating those platforms as state actors.⁶⁵ Nonetheless, insofar as a private actor exercises functions that were traditionally and exclusively held by the state, or that actor's actions were so closely regulated by the state such that the former's actions can be fairly treated as that of the state itself, that actor can be subject to the First Amendment.⁶⁶ In *Lloyd Corp. v. Tanner* [exclusive powers test],

⁶² Cho & Gallo (2021) at 22

⁶³ Keller (2019) at 4; Brannon (2019) Summary

⁶⁴ Brannon (2019) at 1

⁶⁵ See *Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass'n*, 531 U.S. 288, 298 (2001)

⁶⁶ See *id.* at 295; *Lloyd Corp. v. Tanner*, 407 U.S. 551, 569 (1972); *Jackson v. Metro. Edison Co.*, 419 US 345, 350-52 (1952)

the Supreme Court clarified that claimants arguing that the private actor exercised the traditional and exclusive functions of the state must satisfy a high threshold i.e. that the actor's exercise of these powers was such that the actor essentially "stood in the shoes of the state."⁶⁷ In *Packingham v. North Carolina* [extensive supervision test] the Court held that social media platforms can serve as public forums, insofar as they are important places for people to speak and listen, and that social media users in particular engage in a wide array of protected First amendment activities.⁶⁸ Nevertheless, the Court previously held that simply opening a private place to the public is, without more, insufficient to entitle the public to First Amendment protections.⁶⁹ Alternatively, the Court noted that the public may be granted First Amendment protections if a private company is sufficiently under the state's control; however, this requires extensive state regulation, and lower courts have interpreted this as requiring state operation or management of the company's website.⁷⁰ Thus, whether under the exclusive powers, or the extensive supervision test, social media platforms have been highly resistant, if not immune to First Amendment arguments. Specifically this means that private companies are not required to carry user content, and may remove or limit access to that content as they see fit. The concern here is that, insofar as it is economically beneficial, private intermediaries may not only remove illegal content but unpopular or offensive content as well; moreover, this is especially likely when such

⁶⁷ 407 U.S. at 569

⁶⁸ *Packingham v. North Carolina*, 137 S. Ct. 1730, 1735–36 (2017)

⁶⁹ Brannon (2019) at 24 (citing *Lloyd Corp. v. Tanner*, 407 U.S. at 569)

⁷⁰ See *Jackson v. Metro. Edison Co.* 419 U.S. at 350; see also *Quigley v. Yelp Inc.*, No. 17-cv-03771-RS, 2017 U.S. Dist. LEXIS 103771 at 5-6 (N.D. Cal. July 5, 2017)

content is expected to drive other users away.⁷¹

- d. Additionally, even if intermediaries were sued on non-constitutional grounds, §230 of the CDA provides intermediaries with broad immunity. Social media platforms (i.e. intermediaries) like Facebook constitute as interactive computer services [ICS], and as such are entitled to: (1) not be treated as the publishers or speakers of any information, regardless of the nature of that information; and (2) such services may not be held liable for voluntarily acting in good faith to remove or restrict access to objectionable content on their platforms.⁷² These protections indicate that intermediaries will not only avoid liability for not removing illegal or offensive content uploaded by their users, but can freely restrict access to that content if it is for non-fraudulent reasons.⁷³ Furthermore, several courts have held that an ICS will lose its immunity if it *materially contributed* to the alleged illegality of the content, by being responsible for that illegality through e.g. willfully publishing content that the service knows is unlawful.⁷⁴ Nevertheless, §230 does not provide complete immunity. For example §230(e)(2) allows for suits alleging a violation of intellectual property [IP], otherwise the laws pertaining to IP would be limited.⁷⁵ This leads to the third direct approach to regulating content moderation by social media platforms in the US; the DMCA.

⁷¹ Keller (2019) at 23

⁷² 47 USC §230(f)(2)(defining interactive computer service as “any information service, system, or software provider that provides or enables computer access by multiple users to a computer server.”); 47 USC §230(c)(1), (2); *see Klayman v. Zuckerberg*, 753 F.3d 1354, 1357 (D.C. Cir. 2014)(holding that companies like Facebook are interactive computer services)

⁷³ *See Cho & Gallo* (2021) at 14, 27

⁷⁴ *See Fair House Council v. Roommates.com, LLC*, 521 F.3d 1157, 1162, 1170 (9th Cir. 2008)(quoting 47 U.S.C. § 230(f)(3)); *see also Jones v. Dirty World Entm’t Recordings, LLC*, 755 F.3d 398, 413 (6th Cir. 2014)

⁷⁵ *See, e.g., Gucci Am., Inc. v. Hall & Assocs.*, 135 F. Supp. 2d 409, 413 (S.D.N.Y. 2001)

- e. In contrast with the other approaches, the DMCA imposes strict obligations upon platforms; however, these obligations have negatively impacted the users. Assuming that a particular platform has fulfilled the safe harbor requirements under the DMCA e.g. designating an agent to whom copyright owners may send infringement notices, this platform may qualify as a hosting service under §512(c).⁷⁶ Host services are generally immune from secondary liability if they: (1) do not know or have reason to know that the material posted to websites they host is infringing; (2) are either unable to control what their customers post to their sites or gain no direct financial benefit from those postings (such as a fee for each item posted); (3) adopt and implement a policy for terminating service to repeat infringers, and not undercut the effectiveness of standard technological protection measures (such as encryption); and (4) crucially, they must comply with the “notice-and-take-down” provisions under §512(i).⁷⁷ While this is somewhat similar to §230, a major divergence is the 'notice and takedown' provisions. Therein lies the issue. Platforms are obligated to take down genuinely unlawful material that they are notified of or discover under this provision, but in doing so have been criticized for excessively taking down objectionable content.⁷⁸ Particularly due to the DMCA, these takedowns arguably resulted from either: (a) an overabundance of caution, as well as to avoid the costs of having lawyers assess the legitimacy of the content; or (b) the platforms failed to verify users' claims regarding the content's illegality.⁷⁹ Though other platforms have tried to sift through such user claims, these efforts are

⁷⁶ See 17 U.S.C. §512(c), (c)(2) [hereinafter §512(c)]

⁷⁷ §512(c)(1)(A)-(C); §512(i); see Bloch-Wehba (2020) at 62 (citing §512(c)(1)(A))

⁷⁸ Keller (2019) at 3; Bloch-Wehba (2020) at 75

⁷⁹ Keller (2019) at 3

rare.⁸⁰ With content moderation in general, these takedowns have been exacerbated by automated content filters, and in some cases have even left unlawful content untouched.⁸¹ Adding to this problem are calls by policy makers to apply ex ante automation to specific forms of speech; however, as previously noted, automated filters are not yet capable of avoiding collateral takedowns.⁸² Furthermore, while the DMCA neither explicitly requires platforms to proactively moderate content, nor apply the DMCA's rules extraterritorially, platforms nevertheless tend to opt for global takedowns.⁸³ This tendency is partially because it is logistically and technically easier for companies like Facebook and Google to have a single set of Community Guidelines that are adjusted/expanded in response to governmental pressure.⁸⁴ Moreover, it is easier for their engineers to develop content filters that work globally than locally.⁸⁵ This example of local laws with global effects is even more prevalent in Europe where policy makers both regionally, and at the state levels have been more willing to directly influence moderation practices.

- f. Unlike the US' relatively stagnant approach to content moderation, the EU is currently developing how social media platforms, and other digital service providers behave within the EU's borders.⁸⁶ The EU's regard for free speech

⁸⁰ *Id.*

⁸¹ Cho & Gallo (2021) at 17; Bloch-Wehba, (2020) at 75 (noting that in a random sample of over 1800 DMCA takedown requests, a significant number requests either incorrectly identified or insufficiently specified the allegedly infringing work)

⁸² Bloch-Wehba (2020) at 75, 82-5; *see supra* note 7

⁸³ *Id.* at 86

⁸⁴ *Id.*

⁸⁵ Keller (2019) at 8

⁸⁶ Alice Tidey, Ana Lazaro & Jack Parrock, *Digital Services Act: Brussels vows to put order into chaos of digital world with new tech laws*, euronews (Dec. 15, 2020) <https://www.euronews.com/2020/12/15/digital-services-act-brussels-unveils-landmark-plans-to-regulate-tech-companies>

is highly similar to the US insofar as e.g. shocking or offensive content is concerned, but that regard is not universally applied among EU member states.⁸⁷ Specifically, free speech concerns have sometimes been superseded by safety and economic concerns.⁸⁸ Thus the EU's approach to regulating digital services in general, and content moderation in particular explicitly implicates multiple stakeholders and perspectives. Knowing how this approach applies to platforms like Facebook, and why a new approach is under development requires an understanding of the eCommerce Directive [the Directive].

- g. Since 2000, the EU's governance of digital services has been approached through the Directive with the overall goal of fostering e-commerce throughout Europe; however, the Directive's fragmented application, and the lack of adequate content moderation policies triggered calls for a supplementary approach – The Digital Services Act Package [the DSA Package].⁸⁹ Despite its shortcomings, several of the Directive's principles remain significant, specifically Articles 14 (Hosting providers) and 15 (no general obligation to monitor).⁹⁰ Article 14 provides that if a company storing user data lacks actual or constructive knowledge (i.e. is not aware of facts or circumstances from which the illegal activity or information is apparent), the company is not liable for that data's content.⁹¹ Hosting

⁸⁷ Levush (2019) at 4 (noting that the European Court of Human Rights has declared that freedom of speech applies to ideas that “offend, shock or disturb the State or any sector of the population.”)

⁸⁸ See Article 19, *At a glance: Does the EU Digital Services Act protect freedom of expression?*, EDRI (March 10, 2021) <https://edri.org/our-work/does-the-eu-digital-services-act-protect-freedom-of-expression/> [hereinafter Article 19 (2021)]

⁸⁹ Madiega DSA (2021) at 2-5

⁹⁰ Tambiama Madiega, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act*, Study Commissioned by the Directorate-General for Parliamentary Research Service, PE 649.404, 2-3 (May 2020) [hereinafter Madiega (2020)]

⁹¹ 2000 O.J. (L 178) 13

providers can retain their immunity if they expeditiously remove or disable access to illegal activities or information of which they have actual knowledge.⁹² Although member states cannot mandate that service providers actively monitor for illicit content, the Court of Justice of the European Union [CJEU] established that states can direct platforms to detect and prevent specific types of illegal content under those states' laws.⁹³

- h. In *Eva Glawischnig-Piesczek v. Facebook Ireland Limited*, the plaintiff was the former head of the Austrian Green party, and was called a corrupt oaf and member of a fascist party on Facebook.⁹⁴ Facebook disabled access to that content in Austria, and a criminal court held that Facebook had to cease disseminating equivalent content only if Facebook had knowledge of that content; however, an Austrian civil court held that the content was 'excessively harmful' to the plaintiff's reputation, and held Facebook liable.⁹⁵ Following a referral to the CJEU the Court broadly held that injunctions requiring platforms to proactively remove both identical and equivalent content are permitted by the Directive.⁹⁶ Additionally, injunctions to block particular content identified by a court are also permitted.⁹⁷ That said, such injunctions could not require the platform to *independently assess* whether specific content violates the law.⁹⁸ Moreover, the CJEU previously differentiated between impermissible general monitoring and permissible specific injunctions in *L'Oreal SA and Others*

⁹² Madiega (2020) at 3

⁹³ 2000 O.J. (L 178) 13; Keller (2020) at 616; Madiega (2020) at 3

⁹⁴ Keller (2020) at 617

⁹⁵ *Id.*; Judgment of the Court (Third Chamber) of 3 October 2019 ECLI:EU:C:2019:458, ¶¶ 14, 17-8

⁹⁶ Keller (2020) at 617

⁹⁷ *Id.*

⁹⁸ *Id.*

v. eBay International AG and Others.⁹⁹ The Court held that online service providers cannot be ordered to actively monitor all user data but they can be ordered to terminate a particular user's account or make that user easier to identify.¹⁰⁰ These rulings and provisions highlight several major, and widely supported principles: (1) **country of origin principles** i.e. providers must comply with the laws of a member state to access the EU Single Market; (2) the **Limited Liability Regime** provides that online intermediaries are exempt from the content they convey/host if they fulfill certain conditions [safe harbor principle] e.g. hosts must expeditiously remove illegal content once they know of it; and (3) member states cannot impose a **general obligation to monitor** information that providers e.g. store for their users.¹⁰¹ Still, despite these virtues, the European Commission found large variances in the way the Directive was implemented throughout the EU.¹⁰² Specifically, national case law on intermediary liability remains highly fragmented due to conflicts between court rulings, and uncertainty regarding the application of national norms.¹⁰³ Furthermore, prior to 2021 the European Commission generally relied on online platforms to voluntarily commit to codes of conduct or practices directly related to content moderation practices.¹⁰⁴ These practices included e.g. providing monthly reports to the European Commission on actions undertaken to tackle fake accounts.¹⁰⁵ While several major platforms have agreed to these

⁹⁹ See Judgment of the Court (Grand Chamber) of 12 July 2011 ECLI:EU:C:2011:474

¹⁰⁰ Judgment of the Court (Grand Chamber) of 12 July 2011 ECLI:EU:C:2011:474, ¶¶139, 141-42

¹⁰¹ Madiaga (2020) at 2-3; Madiaga DSA (2021) at 2-3

¹⁰² Madiaga DSA (2021) at 2-3

¹⁰³ *Id.*

¹⁰⁴ Gawer & Srnicek (2021) at 75

¹⁰⁵ *Id.*

codes of conduct, it has been ineffective with respect to moderating some types of objectionable content e.g. misinformation.¹⁰⁶ This fragmentation, coupled with a lack of clear guidance on how to supervise digital services resulted in several issues, namely: users' increased exposure to illegal and harmful content; market dominance by certain platforms; and a divided EU Single market.¹⁰⁷ In response, the Commission [on the basis of Art. 114 of the Treaty on the Functioning of the European Union] put forward the Package to "prevent divergences from hampering the free provision of cross-border digital services and to guarantee the uniform protection of rights and uniform obligations for businesses and consumers across the internal market." Though it is not yet in force the Package's provisions have elicited lively responses from multiple stakeholders, including social media platforms.¹⁰⁸

- i. In light of these issues the DSA package presents promising albeit underdeveloped solutions to the content moderation problem. The package is comprised of two pieces of legislation: (a) the eponymous DSA which focuses on making a safer digital space in which user rights are protected and (b) the Digital Markets Act [DMA] which focuses on curbing the market dominance of platforms like Facebook to increase competitiveness, growth and development both regionally and abroad.¹⁰⁹ The DSA focuses on

¹⁰⁶ *Id.* at 78

¹⁰⁷ *Id.* at 1; Madiaga DSA (2021) at 64

¹⁰⁸ See generally Gawer & Srnicek (2021); Article 19 (2021); Jan Penfrat, *The EU's attempt to regulate Big Tech: What it brings and what is missing*, EDRi (Dec. 18, 2020) <https://edri.org/our-work/eu-attempt-to-regulate-big-tech/> [hereinafter Penfrat (2020)]; DOT Europe, *DOT Europe preliminary remarks on the DSA: Consider the focus, scope and coherence of the proposal*, (Feb. 2021) <https://doteurope.eu/wp-content/uploads/2021/02/DOT-Europe-DSA-high-level-remarks-February-2021-.pdf> [hereinafter DOT Europe (2021)]

¹⁰⁹ Gawer & Srnicek (2021) at 64-5

ensuring platform transparency, and protecting users' fundamental rights through a tiered set of provisions.¹¹⁰ Under the DSA every digital platform or service that connects users to goods is obligated to undertake certain duties with respect to how they handle user data and illegal content.¹¹¹ All platforms are required to clearly and unambiguously keep users informed on how user information is collected, what it is used for, and metadata on user targeted ads.¹¹² Platforms must provide clear-cut notice & takedown mechanisms, along with detailed reports on how the user's content was illegal, or how that content violated the platform's terms and conditions.¹¹³ Additional obligations apply to very large operating platforms (VLOPs) with over 45 million users a month, including: clarifying the key determinants used by their algorithms to curate and rank content; analyzing the systemic risks posed by using the platform, as well as implementing effective content moderation mechanisms to mitigate those risks; and undergoing annual independent audits, along with employing a dedicated compliance officer to ensure the platforms' compliance under the DSA.¹¹⁴ Also, unlike similar ventures by the European Commission in the past, the Package emphasizes fundamental rights like free speech by retaining conditional immunity from liability for hosting providers, and the prohibition on general monitoring.¹¹⁵ By preserving the protections provided under the Directive, the DSA package can reduce the risk of

¹¹⁰ Madiega DSA (2021) at 6

¹¹¹ Gawer & Srnicek (2021) at 65

¹¹² *Id.* at 65-6

¹¹³ *Id.*

¹¹⁴ *Id.* at 66

¹¹⁵ *See generally*, Eur. Conv. On H.R., Art. 10 (freedom of expression); Article 19 (2021)(noting that the Package contains 11 mentions of fundamental rights)

unnecessary takedowns.¹¹⁶ Users' speech rights are further protected by requiring platforms to participate in, and subsequently report on, out of court dispute settlements with users regarding illegal content takedowns.¹¹⁷ That said, the DSA and the DMA are not, and should not be understood as distinct acts.

- j. Though seemingly separate, the DMA & DSA are complimentary, and are rooted in the concept of user data.¹¹⁸ The DMA focuses on so called 'gatekeepers' such as Facebook or Google or other platforms which satisfy a 'three-limbed test': (1) they have a significant impact on the European internal market; (2) they provide a core platform service which serves as an important gateway for business users to reach end users; and (3) they enjoy an entrenched position in their operations, or it is foreseeable that they will enjoy such a position in the near future.¹¹⁹ Here, 53% of all EU enterprises use Facebook, and in the first quarter of 2021 Facebook recorded 423 million European users per month.¹²⁰ Adding to this is that Facebook's worldwide social media market share increased from 64% in 2019 to roughly 70% in 2020, conversely platforms like Twitter and Tumblr remained below 16%.¹²¹ Facebook dominates the social media market. As previously discussed, platforms like Facebook grew and developed by collecting user data to maintain user attention, and thereby implementing more narrowly tailored ads; further, their users' mistrust notwithstanding,

¹¹⁶ Supra Part I(C)(g)

¹¹⁷ Article 19 (2021)

¹¹⁸ *Id.*; See supra Part I(A)(a)

¹¹⁹ Gawer & Srnicek (2021) at 64, 66

¹²⁰ *Id.* at 18; *Facebook: quarterly MAU in Europe Q4 2012-Q1 2021*, Statista (July 29, 2021)

<https://www.statista.com/statistics/745400/facebook-europe-mau-by-quarter/>

¹²¹ Gawer & Srnicek (2021) at 25-6

these platforms continued to exist and even thrive.¹²² It is also because of this business model, and a lack of regulation (e.g. antitrust laws) that major platforms like Facebook have become gatekeepers to their respective markets.¹²³ Gatekeepers can restrict competition through e.g. predatory pricing or purchasing potential competitors thereby limiting consumers' platform choices.¹²⁴ This in turn allows firms to impose oppressive contract terms against advertisers, and extract greater amounts of user data.¹²⁵ Limiting competitor access to user data prevents those competitors from delivering that data to advertisers as advertisers are less likely to utilize those competitors' services.¹²⁶ To mitigate these effects, the DMA regulates gatekeeping activities by prohibiting acts like self-referencing; imposing an obligation to share collected data with both business users and regular users; and even forbidding the reuse of personal data by copying it onto other products (e.g. Facebook copying your WhatsApp address book onto Facebook's main platform).¹²⁷ By understanding the connection between user data and market dominance the complimentary nature of the DSA & DMA can be established.¹²⁸ One problem though, is that a holistic examination of the DSA package has not been sufficiently emphasized, and that this among other shortcomings are likely to undermine the package's

¹²² Supra Parts I(A)(a)-(b)

¹²³ Gawer & Srnicek (2021) at 27; Subcomm. On Antitrust, Com. And Admin. L., 116th Cong., Investigation of Competition in Digital Markets 39 (Comm. Print 2020)

https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf?utm_campaign=4493-519
[hereinafter Subcommittee Print (2020)]

¹²⁴ Mike Isaac, *Facebook posts a 33 percent increase in revenue and a 53 percent jump in profit*, The New York Times (Jan. 27, 2021) <https://www.nytimes.com/2021/01/27/business/facebook-earnings.html>;
Subcommittee Print (2020) at 390

¹²⁵ *Id.*

¹²⁶ Cho & Gallo (2021) at 14

¹²⁷ Penfrat (2020)(exemplifying self-referencing as Google listing Gmail as the first option when users do a google search for email providers)

¹²⁸ See Madiaga DSA (2021) at 2

efficacy.¹²⁹

- k. Though the DSA Package emphasizes user rights by improving platform transparency and accountability, it is unlikely to sufficiently deter excessive takedowns nor rein in the gatekeepers.¹³⁰ Despite safeguards against excessive takedowns such as: Articles 15 (platforms must explain the reasoning behind the removal); 14.2(a) (users have to explain why they believe the specific content is illegal); and 20.2 (online platforms must adopt measures against misuse e.g. users submitting a manifestly ill-founded notice regarding some content's alleged illegality), the DSA Package still falls short of the mark.¹³¹ Platforms generally have a 'delete first, ask questions later' mentality, and though some DSA critics applauded these new safeguards, other provisions arguably aggravate this mentality. For example, once platforms receive "substantiated notice" of some content's illegality, that constitutes actual knowledge for the purpose of host immunity under Art. 5.¹³² Therefore hosting providers gain a strong incentive to remove content upon notice. Connected to this is that VLOPs [which are not always gatekeepers] are required to annually assess, among other things, their content moderation systems for weaknesses or shortcomings; however, a logistical and practical flaw here is that a VLOPs' compliance with the DSA Package must be assessed by the European Commission [rather than a dedicated independent regulator].¹³³ Efforts to

¹²⁹ Article 19 (2021)

¹³⁰ *Id.*; DOT Europe (2021); *The Digital Markets Act must do more to protect end users' rights*, EDRI (Feb. 11, 2021) <https://edri.org/our-work/eu-the-digital-markets-act-must-do-more-to-protect-end-users-rights/> [hereinafter EDRI (2021)]; Penfrat (2020); see Keller (2020) at 620

¹³¹ Article 19 (2021)

¹³² *Id.*

¹³³ *Id.* (verifying that smaller platforms will subject to independent regulation by 'digital service coordinators')

improve content moderation practices are also undercut by the DSA [act]'s language which, though only focused on illegal content is still overbroad, and despite certain safeguards¹³⁴ raises the risk of objectionable content being unnecessarily removed.¹³⁵ Finally, despite the DMA's virtues it does not adequately focus on easing barriers to enter the social media market, and thereby fails to give [both business and normal] users more choice between platforms.¹³⁶ The provisions and potential outcomes of the DSA Package notwithstanding, the actual long term effects and ramifications are currently unknown. However, now that platforms face mandatory regulations, they have begun to take a more proactive rather than reactive approach to content moderation.¹³⁷ For example, in 2020 the European Commission Vice-President for Values and Transparency, and Twitter's CEO sought to develop rules which, rather than compel platforms to remove objectionable content, instead affect how objectionable content manifests and propagates on those platforms.¹³⁸ At the risk of speculating, this example illustrates how aggressive government regulatory practices in conjunction with meaningful government-platform collaborations can effectively balance safety concerns with users' right to speak. This argument will be studied further in Part II, but as examined below the DSA Package alone is insufficient to mollify excessive content moderation.

1. As social media platforms generally apply their community standards

¹³⁴ See e.g. Penfrat (2020) (noting that the Commission requires platforms to use independent, certified dispute settlement bodies with whom complaints of wrongful content removal can be submitted)

¹³⁵ DOT Europe (2021)(arguing that Art 2(g) partially defines illegal content as including information that references activities that are illegal either under EU or member states' laws)

¹³⁶ EDRi (2021)

¹³⁷ Gawer & Srnicek (2021) at 78

¹³⁸ *Id.*

worldwide, regional or even country specific regulations can have a global impact. To avoid sudden structural changes, platforms have attempted to predict potential government regulations, and preemptively adapt through e.g. content filtering; however, this technology is imperfect, and pressure from both well-informed and misinformed policy makers has sometimes caused unnecessary takedowns or excessive moderation on a global scale.¹³⁹ The US & EU have, until the DSA Package, emphasized a ‘notice and takedown’ system, and were thus more reactive rather than proactive.¹⁴⁰ Platforms on the other hand have mostly attempted to stay ahead of the regulatory curb through a kind of ‘anticipatory obedience,’ by predicting new laws, and adjusting their policies accordingly.¹⁴¹ Content filtering technology like Facebook’s hash database for violent extremists or YouTube’s Content ID system are manifestations of this approach.¹⁴² Unfortunately this technology is still unrefined.¹⁴³ For example, hashes are essentially unique digital fingerprints of specific kinds of content (e.g. beheadings conducted by terrorists) which are then collected in a database that the platform’s algorithm ‘learns’ from and uses to compare to user content that might match the contents of the database.¹⁴⁴ However, while this process has seen great success in, for example, deleting 99% of ISIS

¹³⁹ Article 19 (2021); Elizabeth Dwoskin and Gerrit de Vynck, Facebook’s AI treats Palestinian Activists like it treats American Black Activists. It blocks them, The Washington Post (May 29, 2021, 8:09PM) <https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/> [hereinafter Dwoskin & de Vynck (2021)]; Bloch-Wehba (2020) at 63, 82-5; Keller (2020) at 619; Keller (2019) at 5-9; Judit Bayer & Petra Bárd, *Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches*, Study Commissioned by the European Parliament’s Committee on Civil Liberties, Justice and Home Affairs, PE 655.135, 36-47 (2020) [hereinafter Bayer & Bárd (2020)]

¹⁴⁰ Bloch-Wehba (2020) at 63

¹⁴¹ Keller (2019) at 2

¹⁴² Bloch-Wehba (2020) at 63, 65-6

¹⁴³ Keller (2020) at 619

¹⁴⁴ Bloch-Wehba (2020) at 65-6

content before it was flagged by users, in late May 2021 Facebook and Twitter mistakenly blocked and later restored millions of Palestinian accounts and posts related to the recent strife.¹⁴⁵ Facebook's explanation was that its hate speech detection software mistakenly classified a key hashtag as belonging to a terrorist group.¹⁴⁶ A father's happy birthday wish to his son 'Qassam' was also likely blocked because Facebook blocks many posts about Hamas' military branch: the al-Qassam Brigades.¹⁴⁷ The exact capabilities of Facebook's filtering tools are not known, and that is part of the problem.¹⁴⁸ Though the DSA Package is an encouraging step, some critics underscored the lack of transparency on content filters, specifically the need for detailed reports on false positives and negatives.¹⁴⁹ Additionally, despite these problems outside actors such as courts have demonstrated some undue optimism here: specifically by assuming that platforms are capable of certain kinds of filtering that are actually beyond those platforms' capabilities.¹⁵⁰ In the EU, several platforms have even warned against including the language 'not-illegal-but-harmful' content in potential regulations since it forces the platforms to draw the line between user safety, and freedom of speech and information.¹⁵¹ Thus to avoid a conflict of laws, platforms have urged that they should only be responsible for removing illegal content.¹⁵² When these companies draw lines, they

¹⁴⁵ Dwoskin & de Vynck (2021); Bloch-Wehba (2020) at 57-62

¹⁴⁶ Dwoskin & de Vynck (2021)

¹⁴⁷ *Id.*

¹⁴⁸ *See generally* Keller (2020)

¹⁴⁹ Article 19 (2021)

¹⁵⁰ Bloch-Wehba (2020) at 72-4 (noting how in the Glawischnig-Piesczek case, the CJEU made inferences that Facebook had the 'software magic' to prevent the republication of defamatory content)

¹⁵¹ Gawer & Srnicek (2021) at 77 (giving an example where public insult of religion or blasphemy, is considered legal in Denmark or France, but is illegal in Germany, Poland, Spain, or Italy)

¹⁵² *Id.* at 77

often do so throughout their platform, and in every country wherein that platform is used.¹⁵³ EU regulations can therefore impact Facebook's content moderation throughout the world. As discussed in Part I(C)(e) above, it is often technically, logistically, and legally easier for companies to filter out certain kinds of content everywhere rather than filtering them in a particular location. On that note, it is precisely the desire to avoid drawing lines that Facebook developed an alternative, and allegedly independent oversight solution; The Facebook Oversight Board.¹⁵⁴

- m. The Facebook Oversight Board [the Board] is a private appeals system focusing solely on Facebook and Instagram's content moderation decisions, specifically to help Facebook understand what content should be kept up, or taken down and why.¹⁵⁵ The Board is currently composed of 20 (with a minimum of 11 and a maximum of 40) multinational, multidisciplinary individuals whose authority is provided by a trust agreement which explicitly separates the Board from Facebook, and places the former under the authority of an Independent trust.¹⁵⁶ The Board's independence was enhanced by Facebook's initial gift of \$130 million to the Trust, which used it to fund, manage, and develop an LLC that would provide the Board's future source of funding.¹⁵⁷ This independence is vital to the Board's

¹⁵³ Keller (2019) at 6-8

¹⁵⁴ Kate Cox, *Facebook plans launch of its own "Supreme Court" for handling takedown appeals*, Ars Technica (Sept. 18, 2019, 3:17PM) <https://arstechnica.com/tech-policy/2019/09/facebook-plans-launch-of-its-own-supreme-court-for-handling-takedown-appeals/> [hereinafter Cox (2019)]

¹⁵⁵ *Oversight Board Charter* Article 1 §4 (2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [hereinafter Oversight Board Charter]

¹⁵⁶ *Id.* at Article 1 §1; Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L. J. 2418, 2481 (2020) [hereafter Klonick (2020)]; *Expertise from Around the World*, Oversight Board (accessed May 30, 2021) <https://oversightboard.com/meet-the-board/> (highlighting the diversity of the members, as well as their various areas of expertise e.g. journalism, constitutional law, technology regulation and policy design)

¹⁵⁷ Oversight Board Charter, Article 1 §5; Klonick (2020) at 2486;

mandate of promoting free expression through principled, independent decisions regarding Facebook and Instagram's (owned by Facebook) content moderation decisions.¹⁵⁸ Although, this independence is perhaps undermined by both the rules that govern the appeals process, and those governing the Board's decision making process.¹⁵⁹ First, appealing content decisions on Facebook is a five step process requiring users to satisfy several pre-requisites, namely that Facebook has to review a user's case and render a final decision; in addition, that decision must include a **reference ID** that can be used to submit an appeal.¹⁶⁰ Though this might suggest that Facebook controls which cases are subject to the Board's review, this is undercut by the manner in which the Board selects cases. The Board uses its discretion to choose cases that are emblematic of major issues in content moderation such as censorship of hate speech, female nudity, and covid-19 misinformation.¹⁶¹ However, while the rules governing the appeals process are arguably adequate, the rules governing the decision making process present a different issue. The Board's decisions are derived from several sources, and with objectionable content in particular they are: (a) Facebook's Community Standards [specifically under 'Objectionable Content']; (b) Facebook's Values (e.g. users' rights to voice their views); and (c) Human Rights Standards (e.g. the International Covenant on

¹⁵⁸ Cox (2019)

¹⁵⁹ See Klonick (2020) at 2478, 2488

¹⁶⁰ *Appealing Content Decisions on Facebook or Instagram*, Oversight Board (accessed May 30, 2021) <https://oversightboard.com/appeals-process/>

¹⁶¹ Elena DeBré, *The Independent Facebook Oversight Board has made its First Rulings*, SLATE (Jan. 2021, 7:23PM) <https://slate.com/technology/2021/01/facebook-oversight-boards-content-moderation-rulings.html> (finding that out of 150,000 cases submitted in December 2020, the Board chose 6); Oversight Board Charter, Article 2 §1 (establishing that the Board will select cases "that have the greatest potential to guide future decisions and policies")

Economic, Social and Cultural Rights [ICESCR]).¹⁶² Though the Board began accepting cases around October 2020, it has rendered twelve decisions as of May 26th 2021; nevertheless, there are few cases pertaining to Facebook's Objectionable Content standards in the US and EU, and most of those cases focus on hate speech.¹⁶³ Facebook defines hate speech as a direct attack on people based on protected characteristics like race.¹⁶⁴ Facebook prohibits users from posting content that targets a person or group on the basis of such characteristics using "designated dehumanizing" generalizations or behavioral statements such as blackface.¹⁶⁵ However, content shared to condemn or raise awareness about such types of hate speech is an exception.¹⁶⁶ These exceptions to the Community Standard exist to uphold one of Facebook's core values: Voice i.e. creating a place for users to express their diverse views, ideas and information.¹⁶⁷ However, given the potential for abuse on the internet, these values are balanced against considerations like safety (making Facebook a safe, user friendly environment) and dignity (mitigating the harassment and degradation of others to ensure their dignity).¹⁶⁸ Furthermore, given Facebook's impact on human rights, user content is voluntarily assessed under the UN Guiding Principles on Human Rights, under which several human rights instruments are available.¹⁶⁹ Despite the dearth of cases, content moderation is an ever

¹⁶² See generally Oversight Board Charter, Article 2 §2

¹⁶³ Brian Fung, *Facebook's Oversight Board is Finally Hearing Cases, Two Years after it was First Announced*, CNN (Oct. 2020, 4:45PM GMT) <https://edition.cnn.com/2020/10/22/tech/facebook-oversight-board/index.html>

¹⁶⁴ See Facebook Community Standards, §3(12)

¹⁶⁵ *Id.*

Case Decision 2021-002-FB-UA, Reference ID: FB-S6NRTDAJ, at 8 (Oversight Board, April 13, 2021) <https://oversightboard.com/decision/FB-S6NRTDAJ/> (hereinafter Zwarte Piet Decision)

¹⁶⁷ *Updating the Values that Inform our Community Standards*, Facebook (Sept. 12, 2019), <https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/>

¹⁶⁸ *Id.*; Zwarte Piet Decision at 5-6

¹⁶⁹ Zwarte Piet Decision at 6 (highlighting the ICESCR and its general comments)

developing process, and by assessing how the Oversight Board has led this issue, insight into a platform-perspective approach can be gained, and existing gaps to this process can be filled.

- n. Facebook cannot effectively evaluate whether or not a user's contribution constitutes hate speech without first understanding how its users think, and holistically examining the context of that speech. User content constitutes hate speech when the factual circumstances, including user intent, are assessed through three lenses: (1) Facebook's Community Standards; (2) Facebook's Values; and (3) International Human Rights law, and is subsequently found to be an unprotected form of speech.¹⁷⁰ However, unless used to condemn or raise awareness about some hate speech, certain forms of speech are sufficiently egregious as to raise the risk to people's safety and dignity, and may be removed regardless of user intent or cultural values.¹⁷¹ In April 2021, the Board rendered its decision on whether to uphold Facebook's removal of a video of 'Sinterklaas' and 'Zwarte Piet' that were posted for the user's friends and family.¹⁷² In Dutch Christmas tradition, 'Zwarte Piet' or 'Black Pete' is 'Sinterklaas' (St. Nicholas) helper, and is often portrayed in blackface with exaggerated lips, and gold earrings.¹⁷³ Although the user and many Dutch people view Zwarte Piet as lacking any racial intent, Facebook has, since August 2020, explicitly prohibited caricatures of Black people under its Hate Speech Community

¹⁷⁰ Zwarte Piet Decision at 5-8; Case Decision 2021-005-FB-UA, Reference ID: FB-RZL57, at 2, 5-8 (Oversight Board, May 20, 2021) <https://oversightboard.com/decision/FB-RZL57QHJ/> (hereafter Two Buttons Meme Decision)

¹⁷¹ Zwarte Piet Decision at 11

¹⁷² *Id.* at 1, 5

¹⁷³ Becky Little, *This Notorious Christmas Character is Dividing a Country*, (Dec. 2018) <https://www.nationalgeographic.com/history/article/black-pete-christmas-zwarte-piet-dutch>

Standards, specifically blackface.¹⁷⁴ The Board held that despite the user's innocent intent, blackface is inherently discriminatory under Facebook's Community Standards, regardless of user intent.¹⁷⁵ Moreover, even though the video was intended for a small number of people, portrayals of Zwarte Piet have been inextricably linked to negative and racist stereotypes that can harm Black People's dignity and safety if left unchecked.¹⁷⁶ Finally, though the right to participate in cultural life, and the freedom of expression (including 'deeply offensive' expression) are enshrined in international law, those rights are not absolute.¹⁷⁷ Nevertheless, Facebook properly restricted those rights by: (1) clearly and precisely notifying users through information videos and newsfeeds, that barring an exception, content featuring blackface will be removed (legality); (2) Facebook's restrictions were legitimate because they were aimed at protecting the right to equality and non-discrimination; and (3) the restrictions were necessary and proportionate to those interests because of the harms posed (both physically and emotionally) in allowing this type of content to accumulate.¹⁷⁸ Facebook's restrictions were thus deemed valid under the Community Standards, Facebook's values, and Facebook's human rights obligations, and as such the Board upheld the removal.¹⁷⁹

- o. Conversely, in May, 2021, the Board overturned Facebook's removal of a meme relating to Turkey's views of the Armenian genocide because that

¹⁷⁴ Zwarte Piet Decision at 2

¹⁷⁵ *Id.* at 10

¹⁷⁶ *Id.* 11

¹⁷⁷ *Id.* 12 (citing United Nations, International Covenant on Civil and Political Rights, 999 UNTS 171, Arts. 2, 19(3) (23 March 1976) and Human Rights Comm., General Comment No. 34, Article 19: Freedoms of Opinion and Expression, CCPR/C/CG/34, ¶¶11-12 (Sept. 12, 2011))

¹⁷⁸ Zwarte Piet Decision at 13-4

¹⁷⁹ *Id.* at 15-8

meme fell under Facebook’s exception for content that condemns or raises awareness of hatred.¹⁸⁰ The user used the ‘Two Buttons’ meme to point out the irony of Turkey’s denial of Armenian genocide, while also claiming that the genocide was justified.¹⁸¹ The ‘Two Buttons’ or ‘Daily Struggle’ meme is an image of a sweating character attempting to push one of two red buttons with contradictory statements.¹⁸² Here, the Board applied the same standards as the ‘Zwarte Piet’ case; however, to assess the content under the Community Standards, the Board addressed the comments on the buttons individually before juxtaposing them in the context of the meme.¹⁸³ The first statement, ‘The Armenian Genocide is a lie’ was viewed by Facebook as a direct attack on a protected characteristic (ethnicity or national origin), but, upon considering the user’s intent, and type of meme used, the Board held that the statement was intended to satirize Turkey’s denial.¹⁸⁴ The second statement: ‘The Armenians were terrorists that deserved it,’ though seemingly dehumanizing was quite the opposite.¹⁸⁵ When assessing the phrases as part of the meme, the majority held that the user clearly intended to use satire to raise awareness of, and condemn Turkey’s efforts to deny the Armenian Genocide whilst simultaneously vilifying the victims as terrorists.¹⁸⁶ With regards to Facebook’s values, the Board held that despite the Armenians’ struggle during those events, and their efforts to gain recognition and justice for those harms, the meme was unlikely to

¹⁸⁰ Two Buttons Meme Decision at 1

¹⁸¹ See Annex 1; Two Buttons Meme Decision at 1, 7

¹⁸² Know Your Meme, Daily Struggle, (updated May 27, 2021) <https://knowyourmeme.com/memes/daily-struggle>

¹⁸³ Two Buttons Meme Decision at 10

¹⁸⁴ *Id.*

¹⁸⁵ *Id.* at 11-2

¹⁸⁶ *Id.* at 12

undermine the peoples' safety and dignity.¹⁸⁷ Finally, the Board held that given the particularly high value of artistic expressions concerning public figures under international law, and Facebook's misunderstanding of the meme, Facebook's restrictions were invalid under international law.¹⁸⁸ Specifically: (1) Facebook both wrongfully applied an inappropriate standard to the meme (the Cruel and Insensitive Community Standard), and failed to properly notify the users of the reason for the enforcement; (2) as there were no legitimate safety or dignity concerns, Facebook's restrictions lacked a legitimate aim; and (3) the restriction was unnecessary because rather than undermine the right of Armenians to equality and non-discrimination, the meme was intended to do the opposite by condemning the Turkish government's "contradictory and self-serving position."¹⁸⁹ While not necessarily indicative of how the Board might deal with other types of objectionable content, the Board's method of dealing with Hate speech reflects a globalized approach to content moderation. Though the Board is obviously required to assess cases using Facebook's constantly updating community standards, that is inevitable given the ever-developing definition of what constitutes objectionable content. Further, by assessing those standards and values under the rubric of International Human Rights law, the Board can avoid a country/region specific analysis, and instead use a holistic assessment that can apply to people and cultures worldwide. The Board also demonstrated an additional, and vital approach to content assessment in the Two Buttons Case - it understood how the meme worked,

¹⁸⁷ *Id.* at 13

¹⁸⁸ *See* Two Buttons Meme Decision at 16

¹⁸⁹ Two Buttons Meme Decision at 16

and more generally how users think. Different memes have different uses, but the way they come about is chaotic and highly context dependent. Here, Facebook's moderators partially missed the intent behind the user's submission, by failing to understand what the Two Buttons meme is used for. That said, Facebook only has 15,000 moderators for 3 billion users, and attempting to apply the nuanced, contextual approach recommended by the board requires a different approach.

II. Synthesizing a Solution

- A. First, attempting to suggest solutions to a problem before other solutions like the DSA Package, and the Oversight Board have properly taken root is premature. That said, based on the developments thus far there are several key issues regulators should keep in mind going forward. First, the core of the content moderation problem is the 'attention economy' and the complimentary 'network effect.' Social media companies develop their content filters, and other algorithms in order to acquire as many users as possible and retain them. These developments are not simply based on the desire for a larger market share, but can also be caused by overly optimistic or even short sighted regulations. Additionally, though easing the barriers to entering the social media market might improve platform responsibility by incentivizing those platforms to develop safe, user friendly platforms, it may exacerbate some issues as well. For example, though more platforms means more choices for consumers, platforms may attempt to more aggressively retain their consumers by delivering more divisive i.e. more attention catching, content. Thus understanding and either mitigating or disincentivizing the triggers that push platforms like Facebook to develop these practices is vital to future regulatory practices. Second, the EU has, at least for now, demonstrated that aggressive regulation gets the platforms to take notice and start suggesting solutions that help the

governments while minimizing damage to the platforms' bottom line. Although there is a risk of overcorrection, and even greater takedowns of objectionable content, this risk can be reduced in at least four ways: (a) detailed, transparent reports on all aspects of a platform's content moderation practices, especially how often its algorithms mis-report or fail to report illicit content; (b) annual *independent* audits by government provided auditors to assess how, and to what extent, a platform is developing its content moderation practices; (c) focusing on regulating illegal content specifically, and requiring platforms to hire more human moderators and independent auditors for the sake of objectionable content; (d) clarifying internal laws so that the platforms have an easier time knowing what to look for and thereby help mitigate the 'delete first, ask questions later' mentality. Third, local laws often have global effects and part of the problem here is that the laws of more restrictive countries end up applying to less restrictive countries. Though indirect, this can affect a large portion of the public discourse. The upside though is that the reverse is technically also true; you do not need to change the rules everywhere for them to apply everywhere – aggressive but unified regulatory measures in a few places can result in widespread changes to content moderation practices. This however is a double-edged sword, and risks imposing the values of states who can impose regulatory obligations over those that cannot. Nevertheless, by understanding how and why platforms work the way they do, Governments may be able to impose more nuanced regulations that effectively impact the platforms' bottom line. Moreover, by collaborating with social media platforms, Government regulations can be tailored to suit those platforms' capabilities. This in turn could mitigate the 'delete first, ask questions later' mentality, and result in more careful moderation practices.

Conclusion

In the last several decades, social media platforms have expanded into multinational entities affecting every aspect of human life from science and medicine to art and religion, whether for good, bad or neither. A core part of this expansion has been people, normal users whose reliance on platforms like Facebook or others continues to grow. This reliance partially stems from the need to communicate ideas, creations, and developments that range from wholesome to vile. On either end of that spectrum the users' reliance could simply be based on the need for attention. User mentality however is beyond the scope of this paper. The recent conflagration of misinformation, bullying, harassment, hateful messages and overall negative albeit not necessarily illegal content, has accompanied the growth of social media. However, as most users and as such, their governments, prefer reasonably safe and comfortable forums within which to express themselves, platforms have either on their own initiative or from government pressure, attempted to establish those forums. Algorithmic efforts like Facebook's terrorism hash databases have both succeeded and failed to moderate negative content, whether illicit or objectionable, and has resulted in widespread harm to those users' abilities to freely express themselves. This failure is partially due to the scale at which such platforms operate, overzealous yet shortsighted regulatory efforts, and in cases where regulation is relatively lacking (e.g. the US) the failure can be caused by a platform's excessive desire to retain users. Though some of these efforts are economically efficient, they have resulted in an overdependence on computer based moderation that is not yet capable of dissecting the nuances of human language and intent. That said, and to their credit, some platforms like Facebook have recognized this problem, and the broader issue of differentiating between illegal content and offensive content which [though uncomfortable] should nevertheless be allowed rather than shut out. This is especially so when the content itself has been mischaracterized by those algorithms. This

realization however, and the subsequently developed Facebook Oversight Board, as well as the EU's DSA Package are encouraging for the future of content moderation. Nevertheless, both developments are either too new or not yet in force, thus long term studies of where these developments lead will likely be necessary. Further, even though the Oversight Board has delivered some promising holdings, it is still too soon to say how that will affect Facebook's moderation practices in the long run. For example, future researchers can analyze multiple Oversight Board cases on specific topics over several years to see how much, and the way in which the Board attaches weight to different sources like Facebook's community standards, human rights standards, Facebook values etc. Adding to this, I would argue that a language analysis of different community standards/terms and conditions across different platforms should be analyzed for similarities, gaps, whether the gaps are similar, and whether the differences are all that different. These recommendations, along with those previously discussed must be considered alongside developments in the DSA Package and the Oversight Board in order to protect not just user speech but user safety as well.

Annex



Two Buttons/Daily
Struggle Meme

Bibliography:

2000 O.J. (L 178)

17 U.S.C. §512

47 U.S.C. §230

Access Now, *Protecting Free Expression in the Era of Online Content Moderation: Access Now's preliminary recommendations on content moderation and Facebook's planned oversight board*, (May 2019)

<https://www.accessnow.org/cms/assets/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf>

Alice Tidey, Ana Lazaro & Jack Parrock, *Digital Services Act: Brussels vows to put order into chaos of digital world with new tech laws*, euronews (Dec. 15, 2020)

<https://www.euronews.com/2020/12/15/digital-services-act-brussels-unveils-landmark-plans-to-regulate-tech-companies>

Amy Mitchell et. al., *In Western Europe, Public Attitudes Toward News Media More Divided by Populist Views Than Left-Right Ideology*, Pew Research Center (May 14, 2018)

<https://www.pewresearch.org/journalism/2018/05/14/in-western-europe-public-attitudes-toward-news-media-more-divided-by-populist-views-than-left-right-ideology/>

Annabelle Gawer & Nick Srnicek, *Online Platforms: Economic and Societal Effects*, Study Commissioned by the European Parliamentary Research Service, PE 656.336 (March 2021)

Appealing Content Decisions on Facebook or Instagram, Oversight Board (accessed May 30, 2021) <https://oversightboard.com/appeals-process/>

Article 19, *At a glance: Does the EU Digital Services Act protect freedom of expression?*, EDRI (March 10, 2021) <https://edri.org/our-work/does-the-eu-digital-services-act-protect-freedom-of-expression/>

Arjun Sundararajan, *Network Effects*, New York University Stern School of Business (Dec. 23, 2020), <http://oz.stern.nyu.edu/io/network.html>

Becky Little, *This Notorious Christmas Character is Dividing a Country*, (Dec. 7th, 2018) <https://www.nationalgeographic.com/history/article/black-pete-christmas-zwarte-piet-dutch>

Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass'n, 531 U.S. 288 (2001)

Brian Fung, *Facebook's Oversight Board is Finally Hearing Cases, Two Years after it was First Announced*, CNN (Oct. 2020, 4:45PM GMT)

<https://edition.cnn.com/2020/10/22/tech/facebook-oversight-board/index.html>

Brooke Auxier, *64% of Americans Say Social Media Have a Mostly Negative Effect on the Way Things are Going in the US Today*, (Oct. 15, 2020), <https://www.pewresearch.org/fact-tank/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/>

Casey Newton, *The Trauma Floor*, The Verge (Feb. 25, 2019, 8:00AM) <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>

Clare Y. Cho & Jason A. Gallo, Cong. Research Serv., R46662, *Social media: Misinformation and Content Moderation Issues for Congress*, (2021)

Daphne Keller, *Facebook Filters, Fundamental Rights, and the CJEU's Glawischning-Piesczek Ruling*, 69 GRUR International 616, (2020) <https://doi.org/10.1093/grurint/ikaa047>

Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech*, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1902, (January 2019) <https://www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech>

Debra A. Williamson, *US Social Trends for 2020: eMarketer's Predictions for the Year Ahead*, eMarketer, <https://www.emarketer.com/content/us-social-trends-for-2020> (Jan. 15, 2020)

Denver Area Educ. Telecommunications Consortium, Inc. v. FCC, 518 US 727, 802-3 (1996)(Kennedy, J., concurring)

DOT Europe, *DOT Europe preliminary remarks on the DSA: Consider the focus, scope and coherence of the proposal*, (Feb. 2021) <https://doteurope.eu/wp-content/uploads/2021/02/DOT-Europe-DSA-high-level-remarks-February-2021-.pdf>

Elena DeBré, *The Independent Facebook Oversight Board has made its First Rulings*, SLATE (Jan. 2021, 7:23PM) <https://slate.com/technology/2021/01/facebook-oversight-boards-content-moderation-rulings.html>

Elizabeth Dwoskin & Gerrit de Vynck, Facebook's AI treats Palestinian Activists like it treats American Black Activists. It blocks them, The Washington Post (May 29, 2021, 8:09PM) <https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/>

Emily A. Vogels, Andrew Perrin and Monica Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints*, Pew Research Center (Aug. 2020) <https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/#fn-26445-1>

Eur. Conv. On H.R., Art. 10 (freedom of expression)

Social Media Influences our Political Behavior and puts Pressure on our Democracies, New Report Finds, EU Science Hub (Oct. 27, 2020) <https://ec.europa.eu/jrc/en/news/social-media-influences-our-political-behaviour-and-puts-pressure-our-democracies-new-report-finds>

Flash Barometer 464: Fake News and Disinformation Online, Survey Requested by the European Commission Directorate-General for Communications Networks, Content & Technology, Project Number 2018.2391 (2018)

The Digital Markets Act must do more to protect end users' rights, EDRI (Feb. 11, 2021) <https://edri.org/our-work/eu-the-digital-markets-act-must-do-more-to-protect-end-users-rights/>

Jan Penfrat, *The EU's attempt to regulate Big Tech: What it brings and what is missing*, EDRI (Dec. 18, 2020) <https://edri.org/our-work/eu-attempt-to-regulate-big-tech/>

Social Media: Security Concerns of Sharing Information, Eurostat, <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20201013-1> (last visited May 27, 2021)

Fair House Council v. Roommates.com, LLC, 521 F.3d 1157(9th Cir. 2008)

Updating the Values that Inform our Community Standards, Facebook (Sept. 12, 2019), <https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/>

Giovanni Sartor & Andrea Lorregia, *The Impact of Algorithms for Online Content Filtering of Moderation: Upload Filters*, Study Commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, PE 657.101 (2020)

Gucci Am., Inc. v. Hall & Assocs., 135 F. Supp. 2d 409 (S.D.N.Y. 2001)

Facebook: quarterly MAU in Europe Q4 2012-Q1 2021, Statista (May, 2021) <https://www.statista.com/statistics/745400/facebook-europe-mau-by-quarter/>

Hannah Bloch-Wehba, *Automation in Moderation*, 53 Cornell Int'l L.J. 41 (2020)

Human Rights Comm., General Comment No. 34, Article 19: Freedoms of Opinion and Expression, CCPR/C/CG/34 (Sept. 12, 2011)

Interactive Advert. Bureau, Internet Advertising Revenue Report: Full Year 2019 Results & Q1 2020 Revenues, https://www.iab.com/wp-content/uploads/2020/05/FY19-IAB-Internet-Ad-Revenue-Report_Final.pdf (May 2020)

Jackson v. Metro. Edison Co., 419 US 345 (1952)

Jeff Horowitz & Deepa Seetharaman, *Facebook Executives Shut Down Efforts to Make the Site Less Divisive*, Wall Street Journal, (May 2020, 11:38AM) <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>

John C. Bennet, *The Irony of American History* work by Niebuhr, Encyclopedia Britannica (2020) <https://www.britannica.com/topic/The-Irony-of-American-History>

John Gramlich, *10 Facts About Americans and Facebook*, Pew Research Center (June 1, 2021) <https://www.pewresearch.org/fact-tank/2021/06/01/facts-about-americans-and-facebook/>

John Koetsier, *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*, Forbes (June 2020, 8:08PM EDT), <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=30c5a70854d0>

John Laloggia, *U.S. public has little confidence in social media companies to determine offensive content*, Pew Research Center (July 2019) <https://www.pewresearch.org/fact-tank/2019/07/11/u-s-public-has-little-confidence-in-social-media-companies-to-determine-offensive-content/>

Jones v. Dirty World Entm't Recordings, LLC, 755 F.3d 398 (6th Cir. 2014)

Judgment of the Court (Grand Chamber) of 12 July 2011 ECLI:EU:C:2011:474

Judgment of the Court (Third Chamber) of 3 October 2019 ECLI:EU:C:2019:458

Judit Bayer & Petra Bárd, *Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches*, Study Commissioned by the European Parliament's Committee on Civil Liberties, Justice and Home Affairs, PE 655.135 (2020)

Kate Cox, *Facebook plans launch of its own "Supreme Court" for handling takedown appeals*, Ars Technica (Sept. 18, 2019, 3:17PM) <https://arstechnica.com/tech-policy/2019/09/facebook-plans-launch-of-its-own-supreme-court-for-handling-takedown-appeals/>

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

Klayman v. Zuckerberg, 753 F.3d 1354 (D.C. Cir. 2014)

Know Your Meme, *Daily Struggle*, (updated May 27, 2021) <https://knowyourmeme.com/memes/daily-struggle>

Lloyd Corp. v. Tanner, 407 U.S. 551 (1972)

Mike Isaac, *Facebook posts a 33 percent increase in revenue and a 53 percent jump in profit*, The New York Times (Jan. 27, 2021) <https://www.nytimes.com/2021/01/27/business/facebook-earnings.html>

Case Decision 2021-002-FB-UA, Reference ID: FB-S6NRTDAJ, at 8 (Oversight Board, April 13, 2021) <https://oversightboard.com/decision/FB-S6NRTDAJ/>

Case Decision 2021-005-FB-UA, Reference ID: FB-RZL57, at 2, 5-8 (Oversight Board, May 20, 2021) <https://oversightboard.com/decision/FB-RZL57QHJ/>

Objectionable Content in Community Standards, Facebook, https://www.facebook.com/communitystandards/objectionable_content (last visited May 17, 2021)

Oversight Board Charter, Facebook (2019) https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf

Packingham v. North Carolina, 137 S. Ct. 1730 (2017)

Quigley v. Yelp Inc., No. 17-cv-03771-RS, 2017 U.S. Dist. LEXIS 103771 (N.D. Cal. July 5, 2017)

Ruth Levush, *Comparative Summary, in Limits on Freedom of Expression 1* (Law Libr. of Congress, 2019)

Frederik Stjernfelt & Anne M. Lauritzen, *Facebook's Handbook of Content Removal. In: Your Post has been Removed* (Springer, Cham., 2020) https://doi.org/10.1007/978-3-030-25968-6_11

Stuart Minor Benjamin, *Algorithms and Speech*, 161 Univ. Penn. Law Review 1445 (2013)

Subcomm. On Antitrust, Com. And Admin. L., 116th Cong., Investigation of Competition in Digital Markets (Comm. Print 2020)
https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf?utm_campaign=4493-519

Tambiana Madiega, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act*, Study Commissioned by the Directorate-General for Parliamentary Research Service, PE 649.404 (May 2020)

Tambiana Madiega, *Digital Services Act*, Briefing Commissioned by the Directorate-General for Parliamentary Research Service, PE 689.357 (March 2021)

United Nations, International Covenant on Civil and Political Rights, 999 UNTS 171, (23 March 1976)

Valerie C. Brannon, Cong. Research Serv., R45650, *Free Speech and the Regulation of Social Media Content*, (2019)

Writing Facebook's Rulebook, Facebook <https://about.fb.com/news/2019/04/insidefeed-community-standards-development-process/> (last visited May 17, 2021)