WAKE FOREST JOURNAL OF BUSINESS AND INTELLECTUAL PROPERTY LAW

VOLUME 21 WINTER 2021 NUMBER 2

REGULATING PLATFORMS' INVISIBLE HAND: CONTENT MODERATION POLICIES AND PROCESSES

Karanjot Gill[†]

 A daughter of immigrants and the first in my family to attend law school. I would like to thank my family and friends for their continous love and support as I go through law school and life in general. An extra special thank you to my parents and siblings, my rocks: Jagjit Gill, Anchla Rani, Jasmine Gill, and Sameer Gill.

ABSTRACT

Before 2016, the calls for comprehensive regulation of social media platforms were small in the United States. Since then, the push for regulation, spanning the entire political spectrum, has been raised. Yet, despite the introduction of several Congressional bills, extensive regulation has not passed on a national scale in the United States. Instead, Europe is taking the lead on disseminating law governing social media platforms, covering issues ranging from terrorism to political advertisements. As the United States continues to debate and assess its approach, whether to follow in Europe's footsteps or retain laissez-faire policies, it is important to understand that any regulation should be nuanced and tailored to the specific issues at hand.

By addressing how to regulate content moderation, this Article endeavors to demonstrate the refinement required when approaching any issue involving social media platforms. Through critiquing the divergent approaches to content moderation assumed by the United States and Europe, laissez-faire and heavy-handed regulation respectively, this Article argues both that platforms are best suited to create, apply, and enforce content moderation policies, and that the United States should regulate the consequences of such freedom by compelling transparency and procedural due process.

I. OUTLINE

Section II clarifies the definition of a social media platform, discusses which platforms this Article will focus on and why, and addresses the limitations of the Article. Section III explains how content moderation functions today by focusing on the current moderation policies of major social media platforms and their characteristics, along with analyzing how moderation policies are implemented. Section IV describes the current moderation regulations in place in the United States and Europe. Section V argues that platforms are best suited to create, apply, and enforce content moderation policies by presenting the limitations of American law when creating regulations, critiquing the European policies approach, and focusing on the global implications of an individual government's content moderation policies. Section VI analyzes United States' laissez-faire approach by assessing the consequences of platforms' self-regulation. The analysis finds that the United States, in order to fix these issues, should compel transparency in both the decision-making and rule-making processes of these platforms, along with the disclosure of the detailed moderation policies themselves. Further, it argues that the United States should require procedural protections and clear appeals processes for users.

II. FOCUS AND LIMITATIONS OF THIS ARTICLE

With various meanings surrounding the term "platform," it is important to clarify the definition utilized by this Article. As proposed and applied by Tarleton Gillespie, this Article defines "platforms" as:

> online sites and services that (a) host, organize, and circulate users' shared content or social interactions for them, (b) without having produced or commissioned the bulk of the content, (c) built on an infrastructure beneath that circulation of information, for processing data for customer service, advertising, and profit.[1]

Further, when this Article discusses social media platforms, it is referring, unless otherwise specified, to the following: Facebook, Twitter, and YouTube.[2] These platforms are highlighted because they are the largest, most widely used, and most well-known platforms in the United States. Specifically, sixty-nine percent of adults use Facebook, seventy-three percent use YouTube, and twenty-two percent use Twitter.[3] Moreover, seventy-four percent of those adults use Facebook daily, fifty-one percent visit YouTube daily, and forty-two percent visit Twitter daily.[4] Thus, these three social media platforms capture much of the American population and, consequently, their content moderation policies significantly impact Americans', as online users, individual expression and speech.

Now, this Article has several limitations to its discussion. First, it does not consider platforms other than Facebook, YouTube, or Twitter, such as Reddit or Instagram. Second, it does not address platforms that are prominent in other countries and languages, such as Russia's VK.[5] Lastly, it overlooks other platform types, such as messaging services,

---

[1] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 18 (2018).

[2] YouTube is a subsidiary of Alphabet, but I have used YouTube in my Article to remain consistent with my discussion of platforms.

[3] Pew Research Center, *Social Media Fact Sheet*, PEW RESEARCH CENTER (June 12, 2019), https://www.pewresearch.org/internet/fact-sheet/social-media. Although Twitter is not used by most of the country's population, its influence dramatically increased with President Trump's use of his Twitter account to announce policy decisions, promote his legislative agenda, engage with foreign political leaders, and publicize state visits. *Id.* The official nature of his Twitter account (@realDonaldTrump) was firmly established in *Knight First Amendment Institute at Columbia University v. Trump*, which found the President's twitter account was a government actor and that he engaged in unconstitutional viewpoint discrimination by blocking certain Twitter users from his social media account because he disagrees with their speech. *Id.* (citing 928 F.3d 226 (2d Cir. 2019)).

[4] Pew Research Center, *supra* note 3.

[5] *About Us*, VKONTAKTE, https://vk.com/?lang=en (last visited Nov. 24, 2020).

unmoderated social spaces, online gaming, and comment threads. Despite these limitations, this Article endeavors to provide a useful discussion regarding the way in which the United States should approach regulation of social media, specifically content moderation.

III. CONTENT MODERATION EXPLAINED

Content moderation policies are the unspoken rules behind social media platforms. Despite the impression platforms want to create, platforms are not passive.[6] Instead, each retains its own policy and executes that policy to determine what content can be seen and what must be hidden from users. Facilitating this moderation is an established set of laws or rules, representing democratic values, that are modified and updated through external input.[7] Such a structure, it is safe to say, is similar to a legal or governance system.[8] Nevertheless, through this lens, platforms both protect users from each other, or their antagonists, and remove offensive, vile, or illegal items; thereby providing a curated experience to users.[9]

Platforms have had a great deal of incentives to develop these content moderation policies and systems. Specifically, moderation arose from (1) the necessity of meeting users' norms in order to retain and gain users, ensuring the economic viability of the platform; (2) a sense of corporate responsibility to keep users safe; and (3) an underlying belief in free speech norms, by creating a space where free speech may flourish.[10] Simply, platforms could not run successfully if users were bombarded with violence, pornography, bullying, etc. So, in order to ensure platforms continue to function economically, and as embodiments of free speech on the internet, users must remain protected and some speech must be moderated.

This section explains platform's current, publicly available content moderation policies, labeled as community guidelines in this Article, and how they are employed.

A. Current Platform Policies

Platforms provide community guidelines to reflect the nature of their content moderation decisions.[11] Inferable from these guidelines

---

[6] *See* GILLESPIE, *supra* note 1, at 7.

[7] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1663 (2018).

[8] *See id.* at 1602.

[9] *See* GILLESPIE, *supra* note 1, at 5, 13.

[10] *See* Klonick, *supra* note 7, at 1615; *see also* Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1362 (2018).

[11] GILLESPIE, *supra* note 1, at 45.

are blanket similarities between platforms, in terms of what influenced the creation of these guidelines, what they cover, and how they guide users. Such consistency across multiple platforms demonstrates that the industry is establishing universal norms regarding content moderation policies.

Substantively, community guidelines are written in plain language, which simply announce the platforms' principles and list the prohibited content.[12] For each prohibition, platforms provide varying degrees of explanation to justify their decisions.[13]

Facebook's guidelines, known as Community Standards, cover six categories: violence and criminal behavior, safety, objectionable content, integrity and authenticity, respecting intellectual property, and content-related requests.[14] Each section contains several subcategories, such as hate speech under objectionable content, which provides a policy rationale and a list of prohibited content in plain language for each.[15]

YouTube's Community Guidelines include categories similar to those Facebook covered.[16] In each subcategory, YouTube explains the policy, exceptions to the rules, scenarios in which it removes the content itself, and the consequences for users when they violate its policy, along with a list of examples.[17]

Additionally, Twitter's guidelines, The Twitter Rules, cover similar categories and provide analogous information to those policies provided by Facebook and YouTube, but, unlike both, it furnishes a clear link to its appeals process for when a post is flagged.[18]

Together, such guidelines provide general rules and explanations regarding content moderation to inform users of the platform's moderation policies in plain, unambiguous language.

From these guidelines, it is first inferable that United States (US) common law is the natural reference point in developing these content moderation policies.[19] For example, platforms' prohibitions on hate

---

[12] *See id.* at 46.

[13] *See id.*

[14] Facebook, *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/ (last visited Oct. 24, 2020).

[15] *Hate Speech*, FACEBOOK, https://www.facebook.com/communitystandards/hate_speech/ (last visited Oct. 24, 2020).

[16] *See Community Guidelines*, YOUTUBE, https://www.youtube.com/about/policies/#community-guidelines (last visited Oct. 24, 2020).

[17] *Harassment and Cyberbullying Policy*, YOUTUBE: YOUTUBE HELP, https://support.google.com/youtube/answer/2802268?visit_id=1-636215053151010017-1930197662&rd=1&hl=en (last visited Oct. 24, 2020).

[18] *See* Twitter, *The Twitter Rules*, TWITTER: HELP CENTER, https://help.twitter.com/en/rules-and-policies/twitter-rules (last visited Oct. 24, 2020).

[19] *See* GILLESPIE, *supra* note 1, at 51.

speech utilize US anti-discrimination legal diction when providing the platforms' protected categories: race, ethnicity, national origin, religious officiation, and sex.[20] US common law is so prevalent across platforms because these platforms are American-based companies who hire American lawyers or policymakers, trained and educated in First Amendment and American free speech norms, to develop the modern content moderation policies.[21]

As a further consequence of platforms' American origins and influence, not only is US common law the natural reference point for moderation policies, but these policies reflect American cultural norms, which prioritize free speech above all else.[22] For example, YouTube's guidelines are more permissive of a wide range of ideas and how they are expressed.[23] Somewhat similarly, Facebook attempts to strike a balance between allowing content to remain on the platform, echoing free speech principles, and creating rules that satisfy concerned users.[24] More generally, Twitter's policies embodied the ultimate expression of free speech by refusing to moderate content for a long time.[25]

Moreover, another feature of platform policies are that the guidelines across platforms cover similar topics, which are simply distinguished by labels and approaches unique to each platform. Topics covered across all platforms include: (1) violence, (2) terrorism, (3) child sexual content, (4) harassment, (4) hate speech, (5) self-harm, (6) nudity/adult content, (7) illegal goods/services, (8) platform manipulation, (9) spam, (10) impersonation, (11) manipulated media, and (12) copyright and trademark infringement.[26] Along with these categories, platforms address additional topics as well, based on deliberate policy decisions exclusive to each platform. For instance, Twitter provides guidelines for non-consensual nudity and election integrity, and Facebook posts guidelines for sexual solicitation.[27]

---

[20] *See id.*; *Hate Speech*, *supra* note 15; *Hate Speech Policy*, YOUTUBE: YOUTUBE HELP, https://support.google.com/youtube/answer/2801939?hl=en (last visited Oct. 24, 2020); Twitter, *Hateful Conduct Policy*, TWITTER: HELP CENTER, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (last visited Oct. 24, 2020).

[21] *See* Klonick, *supra* note 7, at 1621 (discussing Facebook's tier structure).

[22] *See id.* (noting that YouTube's policies are, of course, reflective of Alphabet's commitment to free speech).

[23] *See id.* at 1626.

[24] *See id.* (finding that it is likely that such focus on balancing is caused by the widespread attention on Facebook's role in the 2016 election in being a tool in proliferating fake news).

[25] *See id.* at 1626–27.

[26] *Community Standards*, *supra* note 14; *Community Guidelines*, *supra* note 16; *The Twitter Rules*, *supra* note 18.

[27] Twitter, *Non-Consensual Nudity Policy*, TWITTER: HELP CENTER, https://help.twitter.com/en/rules-and-policies/intimate-media (last visited Oct. 29,

Along with these topics, each platform similarly provides its rationale for its policies and explains, or lists, prohibited conduct for user reference. Such banned content follows common-sense notions of what should not be available on platforms. For example, each platform prohibits posting content featuring sexual intercourse and adult nudity.[28] Yet, despite this convergence, each platform provides some nuance to its rules. For instance, Facebook exempts nudity in the context of birth giving and after-birth moments,[29] YouTube specifically prohibits bestiality,[30] and Twitter mentions hentai as prohibited sexual content.[31] Such slight differences in content moderation guidelines are common, but the basic rules and topics covered remain the same across platforms, showcasing industry norms.

There are multiple reasons as to how platforms have reached this level of consensus in covering similar topics and providing commonsense-based guidance to users. Foremost, it is likely that some of it occurred naturally, as platforms developed over time and identified problem areas.[32] This rationale is demonstrated by the way in which platforms have instituted a similar, routine internal process to create these intricate rules.[33] Specifically, across platforms, content policy teams may create rules by either deliberately deciding to translate a new legal obligation into an actionable rule, reacting to an emergency of a type of content they do not want on the platform, or responding to a surge of complaints from users.[34] In addition, the development of these similarities occurred due to the fact that smaller platforms tend to use the rules of larger platforms as examples and, in the case of some platforms that depend on other platforms to exist, deliberately choose to align their rules with platforms they are dependent upon, thereby creating similar policies across platforms.[35]

However, it is important to note that these community guidelines do not fully reflect the content moderation policies internally used by

---

2020); Facebook, *Sexual Solicitation*, FACEBOOK, https://www.facebook.com/communitystandards/sexual_solicitation (last visited Nov. 24, 2020).

[28] *See* Twitter, *Sensitive Media Policy*, HELP CENTER, https://help.twitter.com/en/rules-and-policies/media-policy (last visited Oct. 29, 2020).

[29] Facebook, *Adult Nudity and Sexual Activity*, FACEBOOK, https://www.facebook.com/communitystandards/adult_nudity_sexual_activity (last visited Oct. 29, 2020).

[30] YouTube, *Nudity and Sexual Content Policies*, YOUTUBE: YOUTUBE HELP, https://support.google.com/youtube/answer/2802002?hl=en (last visited Oct. 29, 2020).

[31] *Id.*

[32] *See* GILLESPIE, *supra* note 1, at 66.

[33] *See id.* at 67.

[34] *See id.*

[35] *See id.* at 71–72.

platforms to make moderation decisions. In addition to these community guidelines, platforms have their own internal, more detailed, rules used to make moderation decisions on a case-by-case basis.[36] For instance, Facebook provides internal memoranda and training documents to moderators so as to better instruct their decision-making.[37] Unlike community guidelines, platforms do not make these detailed and intricate rules, essential to the content moderation policies, public.

Overall, some content moderation policies are available to the public in the form of community guidelines. Inferable from the content itself is that, among platforms, these guidelines are remarkably similar in their origins, coverage of topics, and commonsense instructions for users. Now, such similarities are further prevalent in platforms in regard to how content is moderated and by whom.

B. Implementing Moderation Policies

Across online platforms, content moderation generally occurs through three different methods: editorial review, automatic detection, and community flagging.[38] Each method utilizes some combination of employees or third-party contractors, artificial intelligence (AI), and platform users to conduct the moderating itself.[39]

Despite being the most pervasive type of moderation historically, online platforms use editorial review for content moderation on a lesser scale today.[40] Editorial review consists of a moderator, AI or human, scrutinizing content by either approving or rejecting it for posting.[41] Today, platforms utilize this method, but instead of using human moderators to comb through millions of posts before publication, they usually apply AI to screen posts for content that the software can reliably identify.[42] Specifically, this consists of identifying material that is illegal or that the platform otherwise prohibits.[43]

---

[36] *See id.* at 72.

[37] *See* Sarah Koslov, Note, *Incitement and the Geopolitical Influence of Facebook Content Moderation*, 4 GEO. L. TECH. REV. 183, 188 (2019).

[38] *See* Klonick, *supra* note 7, at 1638 (stating that platforms actively screen and remove content themselves after content is published. However, this method of moderation is mostly restricted to extremist and terrorist content). Platforms also actively screen and remove content themselves after content is published. However, this method of moderation is mostly restricted to extremist and terrorist content.

[39] *See* GILLESPIE, *supra* note 1, at 77.

[40] *See id.* at 78.

[41] *See id.*

[42] *See* Klonick, *supra* note 7, at 1636.

[43] *See id.* at 1637.

The lack of human moderators within platforms under editorial review is reasonable as the resources required for full-scale, proactive review are immense and grow exponentially with the platform's growth.[44] Moreover, despite these substantial limitations, utilizing editorial review expansively, where each review is presented as judgement in the platform's name, would open the platforms to charges of subjectivity, hypocrisy, and self-interest.[45] Yet, some platforms choose to use this editorial review form outside the use of AI. For example, Apple's App Store requires that any submission for apps must be screened before publication.[46] Nevertheless, most platforms rely heavily on other types of moderation.

Automatic detection, a method that relies exclusively on AI, is extensively utilized across platforms for screening content after it is posted.[47] AI works well for certain types of content and policies where there is a consensus about what content constitutes a rule violation, such as child pornography.[48] Further, to ensure continued accuracy, AI is regularly evaluated and updated through software updates and machine learning.[49] To demonstrate its success, for example, Facebook found that "[f]or every category, except bullying, harassment, and hate speech," over ninety-five percent of the content removed as a violation of its community standards, "before it was reported by a user, was in large part because of its AI."[50]

However, AI has significant limitations when it assesses issues that rely heavily on context and the nuances of language, such as hate speech.[51] Specifically, AI cannot identify content as problematic if it has not already been identified as being so and, even when it has been recognized, challenges remain because AI detection can lead to false positives and negatives when trying to interpret words.[52] For instance, Facebook determined that AI detects only about thirty-eight percent of the hate-speech-related posts that the platforms removes.[53] Compared

---

[44] *See* GILLESPIE, *supra* note 1, at 86.

[45] *See id.* at 82.

[46] *See id.* at 79.

[47] *See* Klonick, *supra* note 7, at 1636.

[48] *See id.*; *see also* GILLESPIE, *supra* note 1, at 100.

[49] *See* Klonick, *supra* note 7, at 1637.

[50] Evelyn Douek, *Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & TECH. 1, 13 (2019).

[51] *See* Klonick, *supra* note 7, at 1634–35.

[52] *See* GILLESPIE, *supra* note 1, at 99–100.

[53] Jason Koebler & Joseph Cox, *The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People*, VICE (Aug. 23, 2018, 10:15 AM), https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works.

to the ninety-five percent earlier,[54] this lower success rate demonstrates the constraints of AI in moderating nuanced, language-specific issues. Nevertheless, for other, bright-line issues, AI is an important aspect of a platform's moderation strategy and will continue to be critical as platforms boost their AI use for content moderation.[55]

In addition to automatic detection, platforms significantly rely upon their communities to flag posts that violate their community guidelines, which are then removed or unflagged by human moderators employed by platforms.[56] Most online platforms have tools to allow users to notify them about content that is illegal and/or that their community guidelines prohibit.[57] This method places platform users as the forerunning moderators, reporting content for review.

Community flagging provides both significant advantages and challenges for platforms. Specifically, it gives platforms legitimacy and cover, as any changes within the guidelines or mass banning may be framed as what the users want, based on the flagging.[58] Conversely, this method can be and has been hijacked by users to harass others.[59] Nevertheless, platforms rely upon this method extensively, as it is a practical means of reviewing a substantial amount of content and creates a collection of posts for human moderators to review.

In conjunction with community flagging, human moderators are employed by platforms or contracted by third parties, both in the US and around the world, to apply the community guidelines and detailed internal rules to judge content that users flag.[60] For example, at Facebook, moderators are split into three tiers.[61] Tier one consists of lawyers, engineers, or policymakers at company headquarters who create the detailed rules.[62] Tier two moderators, working remotely or in the "call centers," review prioritized content, consisting of imminent threats of violence and self-harm, and supervise tier three moderators.[63] Lastly, tier three moderators, working in "call centers," typically review material that is flagged as lower priority; such as nudity or porn, insults or attacks based on a protected class; and inappropriate or annoying

[54] Douek, *supra* note 50, at 13.

[55] *See* GILLESPIE, *supra* note 1, at 107–08.

[56] *See id.* at 87–88.

[57] *See id.* at 87.

[58] *See id.*

[59] *See id.* at 93; *see also* Robinson Meyer, *The Primary Way to Report Harassment Online Is Broken*, THE ATLANTIC (Aug. 21, 2014), https://www.theatlantic.com/technology/archive/2014/08/the-way-we-report-harassment-on-the-social-web-is-broken/378730/ (discussing the general defects of community flagging).

[60] *See* GILLESPIE, *supra* note 1, at 138–39.

[61] Klonick, *supra* note 7, at 1639–40.

[62] *See* Klonick, *supra* note 7, at 1639–41.

[63] *See id.* at 1640–41.

content.[64] Even though tier three moderators conduct the bulk of the moderation, moderators working in lower tiers can escalate content up to tier one for further evaluation when moderators below are unsure.[65] Of course, this human moderator structure is only representative of Facebook, but it illustrates how human moderation of content users flag practically occurs.

By explaining the public content moderation policies across platforms, along with the three approaches to employing content moderation online, this section explains the content moderation systems platforms created over decades. Largely hidden from users, content moderation plays a critical role in the continued success of platforms. Due to its importance and wide-reaching implications for speech, many across the world have called for its regulation. The following section describes the proposed and implemented laws within the US and Europe.

IV. CURRENT US AND EUROPEAN REGULATIONS

The US and Europe have significantly different approaches to regulation, the former utilizing a free-market, noninterventionist approach, the latter utilizing a heavy-handed approach, requiring platforms to adopt certain policies and imposing large sanctions if platforms fail to abide by the statues. In fact, as the US lags in its regulatory response, debating and deadlocked on party-lines, Europe is taking the lead on governing platforms, quickly becoming the world leader on regulating the modern Internet. This section describes the approaches taken by both the US and Europe.

A. US Content Moderation Regulations

Unlike other countries, the US, on a federal level, does not have regulation specifically targeting the content moderation practices and policies of online platforms. Instead, the First Amendment of the United States Constitution and Section 230 of the Communications Decency Act indirectly affect content moderation efforts.

The First Amendment of the Constitution, also known as the Free Speech Clause, clearly states that "Congress shall make no law . . . abridging the freedom of speech.[66] By virtue of its diction, the First Amendment only implicates state action, leaving private entities, such as online platforms, free to restrict speech. Without this limitation, online content moderation in the US could not exist.

[64] *See id.* at 1639–40.
[65] *See id.* at 1641.
[66] U.S. CONST. amend. I.

Complimenting the First Amendment's grant of power, section 230 of the Communications Decency Act, a landmark Internet regulation, further empowers and protects online platforms. Specifically, its provisions ensure that (i) platforms, as intermediaries, cannot be held liable for the speech of their users, and (ii) if a platform does police what its users say, through content moderation policies, it still cannot be deemed liable.[67] By preventing the government and other private parties from threatening to hold platforms liable, based on the content users post, this statute provides substantive legal protections to platforms and grants permission to establish rules and policies to moderate user content.[68]

Together, these laws create a legal environment in the US that both protects online platforms and empowers them to develop and police content moderation policies online. Unlike most European regulations, the law does not compel platforms to retain specific content moderation policies. Instead, US law utilizes a laissez-faire approach, granting platforms full reign over all aspects of content moderation.

B. European Moderation Regulations

In a stark contrast to the US law, several European countries and the European Union (EU) itself have promulgated laws directly regulating the content moderation policies of online platforms. Specifically, this section discusses (1) Germany's Netzwerkdurchsetzungsgesetz (NetzDG), (2) the United Kingdom's (UK) proposed policies in its Online Harms White Paper, (3) France's election and Avia law, (4) the EU's Code of Conduct for Countering Illegal Hate Speech Online, and (5) the EU's proposed Terrorism Content Regulation.[69] Together, these statutes illustrate Europe's heavy-handed approach to content moderation regulations.

NetzDG was adopted and implemented to hold large social media companies responsible for policing content on their platforms.[70]

---

[67] 47 U.S.C.A. § 230 (2018).

[68] 18 U.S.C.A. § 2421A (2018). There is one exception to this general grant of legal protection: Section 230 does not apply to civil and criminal charges of sex trafficking or to conduct which "promotes or facilitates" prostitution, as stated in the Allow States and Victims to Fight Online Sex Trafficking Act. *Id.*

[69] As of May 2020, the proposed policies by the EU and Britain have not been implemented.

[70] *See* Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law*, 1, 9 (Transatlantic High Level Working Grp. On Content Moderation Online and Freedom of Expression, Working Paper), https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf. Although many consider NetzDG as a drastic, burdensome regulation, it is important to note that Germany did hold

Applying to platforms with more than two million users located in Germany, the statute both shapes how platforms moderate content and requires transparency regarding moderation.[71] First, it is important to note that the law does not create new categories of illegal content, rather it requires the enforcement of twenty-two German statutes—whose substance includes insult, defamation, incitement to hatred, etc.— by platforms.[72] Specifically, the statute requires content, once reported, reviewed, and deemed "manifestly unlawful" to be removed within twenty-four hours of notice, along with mandating that other illegal content be deleted within one week of notice.[73] To incentivize compliance, the statute allows platforms to be fined up to fifty million Euros.[74] Second, the law requires platforms to publish semi-annual reports detailing their content moderation policies, allowing the public to see how the platforms are meeting the law's requirements.[75]

Following in Germany's regulatory footsteps, the UK is also considering promulgating its own heavy regulations of online content moderation, as described in its Online Harms White Paper.[76] The paper proposes establishing a new statutory duty of care for platforms to take reasonable steps to keep users safe, and tackle illegal and harmful activity, along with ensuring its compliance by allowing an independent regulator to enforce the statue through sanctioning platforms.[77] This duty of care includes, but is not limited to, illegal material like terrorism and child pornography, along with activity that is deemed harmful in the UK, labeled "threats to our way of life," which comprises of inciting violence, encouraging suicide, disinformation, inappropriate material being accessed by children, and cyber bullying.[78] Further, it proposes that companies will have to publish annual transparency reports, similar to NetzDG, on the amount of harmful content on their platforms and

companies to a voluntary compliance system beginning in 2015 and, only after finding that platforms were insufficiently compliant, did it implement NetzDG in 2018. *Id.*

[71] *Id.* at 2.

[72] *Id.*

[73] *Id.*

[74] *Id.*

[75] *See id.* (stating that more specifically, the law requires platforms who receive more than 100 complaints per year to publish reports, which, of course, includes most platforms).

[76] John Naughton, *The White Paper on Online Harms is a Global First. It Has Never Been More Needed*, THE GUARDIAN (Apr. 14, 2019, 2:00), https://www.theguardian.com/commentisfree/2019/apr/14/white-paper-online-harms-global-first-needed-tech-industry-dcms-google-facebook.

[77] *Id.*

[78] Natasha Lomas, *UK Sets Out Safety-Focused Plan to Regulate Internet Firms*, TECH CRUNCH (Apr. 8, 2019, 6:45 AM), https://techcrunch.com/2019/04/08/uk-sets-out-safety-focused-plan-to-regulate-internet-firms/.

how they are combating it.[79] The proposed law seems to apply to any company that allows users to share or discover user generated content or interact with each other online, thereby including a broader range of websites than simply social media platforms.[80]

Since the initial proposal was released, the UK has assigned Ofcom, its media regulator, as the entity to monitor internet content and imbued it with the power to issue penalties against platforms that do not meet the duty of care.[81] Specifically, it will oversee two areas: illegal and harmful content. For illegal content, Ofcom will ensure that companies quickly remove such material and prevent it from being published in the first place.[82] For harmful content, Ofcom will make sure social networks enforce their own terms and conditions.[83] Contrary to prior speculation, Ofcom will not have the power to remove specific posts from social media platforms.[84] For now, the UK continues to debate and slowly implement its policies.

Like Germany and the UK, France continues the trend of regulating content moderation through timelines and mandatory requirements imposed upon platforms, as demonstrated by two statutes. First, in 2018 France passed a law which bans fake news during election campaigns and carries the potential penalties of one year in prison and fines up to seventy-five thousand Euros.[85] Second, France adopted the Avia Law, which is similar to NetzDG, as it requires platforms to delete content the French Government deems hate speech within twenty-four hours, and carries fines up to 1.25 million Euros.[86] Hate speech is defined in

---

[79] *See id.*

[80] *See id.*

[81] Adam Satariano, *Britain to Create Regulator for Internet Content*, N.Y. TIMES (Feb. 12, 2020), https://www.nytimes.com/2020/02/12/technology/britain-internet-regulator.html.

[82] Alex Hern, *What Powers Will Ofcom Have to Regulate the Internet?*, THE GUARDIAN (Feb. 12, 2020), https://www.theguardian.com/media/2020/feb/12/what-powers-ofcom-have-regulate-internet-uk.

[83] *Id.*

[84] *See* Satariano, *supra* note 81.

[85] Michael-Ross Fiorentino, *France Passes Controversial 'Fake News' Law*, EURONEWS (Sept. 5, 2019), https://www.euronews.com/2018/11/22/france-passes-controversial-fake-news-law; Eliza Mackintosh, *European Union Task Force Holds its First Summit on Fighting Russian Disinformation*, CNN (Jan. 30, 2020, 9:04 AM), https://www.cnn.com/2020/01/30/europe/european-union-task-force-russian-disinformation-intl/index.html.

[86] Makena Kelly, *France Wants to Fine Facebook Over Hate Speech*, THE VERGE (July 4, 2019, 3:41 PM), https://www.theverge.com/2019/7/4/20682513/french-parliament-facebook-google-social-network-hate-speech-removal; Aurelien Breeden, *France Will Debate a Bill to Stop Online Hate Speech. What's at Stake?*, N.Y. TIMES (July 1, 2019), https://www.nytimes.com/2019/07/01/world/europe/france-bill-to-stop-online-hate-speech.html (noting that as of July 2019, the law passed France's lower parliament; in December 2019, it was being debated in the French Senate).

this context as content which is "manifestly unlawful on grounds of race, religion, sex, sexual orientation or disability."[87] Together, these two regulations govern the platform's content moderation policies and processes, shaping internal decisions in accordance to government authority.

Paralleling the direction of individual European countries, the EU itself has promulgated its own regulation: The Code of Conduct for Countering Illegal Hate Speech Online.[88] This Code reflects the commitment companies made to the EU to remove illegal hate speech within twenty-four hours of notice.[89] The Code was the result of discussions between European Commissioners, the major platforms, EU member states, and civil society organizations. Yet, despite its passage and language, it primarily publicizes and formalizes elements of the platforms' internal content moderation policies.[90] Thus, in practical terms, it essentially amounts to a public relations exercise for platforms, without adding any real regulatory substance.

On the other hand, the EU's proposed Terrorism Content Regulation (TERREG) is more substantial and has the potential to be quite impactful on platforms' content moderation efforts.[91] Proposed to tackle the availability of terrorism content online and prevent radicalization and support for terrorism, TERREG recommends stringent duties for platforms.[92] These include suggestions that terrorist content should be removed within an hour of it being flagged by local law enforcement or Europol, the EU's police agency.[93] Further, it suggests requiring platforms to develop upload filters that capture

---

[87] EURACTIV Network, *France Adopts Tough Law Against Online Hate Speech*, EURACTIV (July 10, 2019), https://www.euractiv.com/section/politics/news/france-adopts-tough-law-against-online-hate-speech/.

[88] *See* Barbora Bukovská, *The European Commission's Code of Conduct for Countering Illegal Hate Speech Online*, 1, 3 (Transatlantic High Level Working Grp. on Content Moderation Online and Freedom of Expression, Working Paper No. 19), https://www.ivir.nl/publicaties/download/Bukovska.pdf.

[89] *Id.* at 5; Dawn Carla Nunziato, *The Marketplace of Ideas Online*, 94 NOTRE DAME L. REV. 1519, 1532 (2019).

[90] Bukovská, *supra* note 88, at 3, 10.

[91] *See* Joris Von Hoboken, *The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications*, 1 (Transatlantic High Level Working Grp. On Content Moderation Online and Freedom of Expression, Working Paper), https://www.ivir.nl/publicaties/download/TERREG_FoE-ANALYSIS.pdf.

[92] *Id.* at 3.

[93] *See* Dustin Volz, *Tech Firms Tout Progress on Scrubbing Online Terror Content*, THE WALL STREET JOURNAL (June 19, 2018, 7:00 AM), https://www.wsj.com/articles/tech-firms-tout-progress-on-scrubbing-online-terror-content-1529406000.

illegal content before users can post it online.[94] However, the regulation, unlike the laws of other European countries, does not provide a standard to determine what content should be removed, instead TERREG allows platforms to remove content based on the terms of each individual platform and to refer violations to law enforcement based on these policies.[95] Along with these requirements, TERREG proposes an appeal system, where users can complain if they believe their content has been unjustifiably removed.[96]

Overall, the US remains liberal in its approach to regulating platforms, without any indication of new regulation. In contrast, as pressure for regulation mounts against platforms worldwide, Europe is quickly setting precedent through the policies it is enacting. However, as admirable and well-intentioned as Europe's regulatory response is to issues of content moderation, its over-regulatory approach is too harsh. Nevertheless, both the US and European approaches are either too lenient or too harsh, another approach to regulation is needed. To understand what regulation is required, it is first important to recognize that social media platforms are best suited to create, enforce, and apply content moderation policies themselves.

V. SELF-EXECUTING CONTENT MODERATION BY PLATFORMS IS THE BEST OPTION

This section proposes that platforms are best suited to create and implement content moderation policies for three reasons. First, American law is limited in creating regulations pertaining to content moderation. Second, Europe's regulatory approach is too harsh, as demonstrated by its negative effects. Third, the global implications of any individual government's policies outweigh any benefits from such regulation.

A. American Law May Not Allow Content Moderation Regulation

From a global perspective, the First Amendment of the Constitution is hailed as one of the strongest protections of free speech. Consequently, its security for individuals prevents the government from passing regulations restricting online speech and regulating platforms'

---

[94] *See* Giovanni Sartor, *The Impact of Algorithms for Online Content Filtering or Moderation*, EUROPEAN PARLIAMENT 9-10 (Sept. 2020), https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf.

[95] *See* Hoboken, *supra* note 91, at 7.

[96] *See id.*

content moderation policies. Moreover, based on current First Amendment jurisprudence, it is unlikely that platforms themselves will be considered state actors.[97] Such limitations on the state disallow a regulatory approach similar to that of Europe, instead it supports the argument that platforms should continue to create and implement content moderation policies themselves.

The Constitution prevents the government from censoring speech. The First Amendment clearly states, "Congress shall make no law . . . abridging the freedom of speech," and, jurisprudentially, this ban is subject only to a few discrete exceptions, such as direct incitement to violence.[98] This means that the government cannot censor speech online, even if the speech, such as fake news or hate speech, may, from a policy perspective, seem harmful. However, as explained previously, since the amendment does not apply to private parties, platforms may regulate and outlaw speech that harms individuals and protected groups.[99]

Further, even if Congress found a workable approach to regulate platforms' content moderation operations, this approach may still be problematic. Statutes that attempt to limit content moderation may be challenged as abridging the platforms' own speech rights.[100] Specifically, platforms may redefine their content moderation procedures and algorithms as editorial choices and argue that any regulation of them should receive strict scrutiny.[101] Current doctrinal trends seem to favor this type of protection for platforms.[102] In fact, under the First Amendment, the Supreme Court of the United States tends to treat all forms of protected communication with equal consideration, the First Amendment is viewed as deregulatory in nature, and the Court does not usually dilute First Amendment jurisprudence because the cases arise in a business setting.[103] Thus, even if a regulation passes under the First Amendment, it faces an uphill battle. As a result, the US may be left with laissez-faire policies in which private platforms remain in charge of content moderation.

---

[97] *See* Jed Rubenfeld, *Are Facebook and Google State Actors?*, LAWFARE (Nov. 4, 2019, 8:20 AM), https://www.lawfareblog.com/are-facebook-and-google-state-actors.

[98] U.S. CONST. amend. I.

[99] *See* John Samples, *Why the Government Should Not Regulate Content Moderation of Social Media*, CATO INSTITUTE (Apr. 9, 2019), https://www.cato.org/publications/policy-analysis/why-government-should-not-regulate-content-moderation-social-media.

[100] *See* Langvardt, *supra* note 10, at 1364.

[101] *See id.* at 1365.

[102] *See id.*

[103] *See id.*

Consequently, parties who, rightly, want the government to help moderate harmful speech online have argued to apply the First Amendment to platforms themselves, essentially making them state actors. In order to do so, platforms would need to be categorized under the existing jurisprudence of the First Amendment as state actors, either as company towns, broadcasters, or editors.[104] However, even if this occurred, such categorization is itself limiting because the First Amendment would disallow any content moderation policies platforms created that are allowable under the First Amendment, such as hate speech or fake news.[105] Further, the only punishment platforms can implement upon users, as state actors under the First Amendment, is restricting access to the site, either temporarily or permanently, and removing users' offensive speech.[106] As a state actor, such approaches are likely problematic under the prior restraint doctrine, as it means denying access to a public forum.[107] Moreover, if the user's speech consists mostly of protected speech, then platforms acting to remove that content from the site may lead to serious First Amendment issues.[108] Hence, the First Amendment's application to platforms would make these sites less valuable to end users, along with making the application of content moderation policies difficult and inefficient.

The First Amendment limitations are two-fold: it prevents direct government censorship and regulation governing platform's content moderation policies; and it limits platforms' authority as state actors. Consequently, any American regulation regarding the problems with content moderation should not arise from new interpretations of the First Amendment or following Europe's regulatory footsteps. Instead, platforms should continue to create and implement content moderation policies.

B. Europe's Regulatory Approach is too Stringent[109]

Unlike the United States, Europe does not have similar constitutional limitations preventing heavy regulation. Instead, Europe can address the many issues caused by social media, ranging from amplifying disinformation to providing a conduit for radicalization,

---

[104] *See* Klonick, *supra* note 7, at 1609.

[105] *See id.* at 1659; *See* Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2026 (2018).

[106] *See* Balkin, *supra* note 105, at 2026–27.

[107] *See id.* at 2027.

[108] *See id.*

[109] Here, I will be discussing the diction, structure, and implications of the European regulations, but these critiques apply to any proposed and future regulations which are written and structured similarly as these.

head-on by directing platform policies. However, the consequences of Europe's regulations are steep as their statues, lead to over-censorship and privatized enforcement of individuals' speech, provide permission for authoritarian governments to implement oppressive regulation, and lack any substantive guidance for platforms. Such issues demonstrate why the US should not follow Europe's lead and, further, supports the argument that platforms should continue to create and implement content moderation policies themselves.

1. Over-Censorship

Foremost, European policies likely result in the over-censorship of speech, caused by ambiguous statutory language, steep monetary penalties, and rigid timeframes.

First, ambiguous statutory language, which may be either too broadly or narrowly drawn, will likely lead to overregulation of user speech by platforms tasked with implementing the specifics of the statute.[110] In fact, without nuance in the statutes themselves, which carve exemptions for keeping up legitimate content like satire or political commentary, the regulations encourage the censorship of speech. NetzDG, the UK's proposed rules, and TERREG best demonstrate this trend.

NetzDG will result in over-censorship for multiple reasons. Specifically, the speech prohibitions in the law, which substantively reflect twenty-two German criminal statutes and includes categories like hate speech and blasphemy, are too broadly defined.[111] These categories may be interpreted broadly or narrowly, however, for social media platforms that are subject to steep fines, it is safer to engage in over-enforcement than under-enforcement to avoid monetary penalties. Further, unique to NetzDG, the statute relies heavily upon other German laws, so overregulation will likely occur because platforms do not have the expertise or time to assess every German precedent in detail, especially since making a legal assessment as to whether the content is prohibited requires expertise in the German language and the law itself.[112]

Similarly, the proposals presented in the UK's White Paper on Online Harm are too vague. The duty of care required by platforms

---

[110] *See* Kyle Langvardt, *A New Deal for the Online Public Sphere*, 26 GEO. MASON L. REV. 341, 352 (2018).

[111] *See* Tworek & Leerssen, *supra* note 70, at 3; Linda Kinstler, *Germany's Attempt to Fix Facebook Is Backfiring*, THE ATLANTIC (May 18, 2018), https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/.

[112] *See* Tworek & Leerssen, *supra* note 70, at 3; Kinstler, *supra* note 111.

focuses on harmful content, but it never defines the term "harmful content." Instead, it provides a list of behavior within the scope of being harmful, as briefly explained above, that is non-exhaustive, and this is an issue because the term "harm" is highly subjective with no identifiable boundaries.[113] Again, as a result, platforms are incentivized to be overcautious in their approach to content moderation when faced with the risk of potential sanctions imposed by the UK government.

Moreover, the EU's proposed regulation, TERREG, also promotes censorship through its imprecise language. Its definition of "terrorism content," as currently drafted, is too broad and ripe for abuse through the implementation of such a definition.[114] Here, the definition's weak diction creates more space for establishing greater restrictions than necessary—where posts that are not clearly intended to support or incite terrorism may be labeled as terrorist content.[115] For example, these may include posts that are responding to or quoting from terrorist propaganda.[116] Even if there is no such intention by EU regulators to create this type of policy, the potential for such interpretation and, consequently, overregulation is present.

However, even when a regulation's language is less ambiguous, platforms will likely overregulate for two reasons: the penalties imposed on platforms for underenforcement and the required rigid timeframes for removal.[117] First, the penalties for noncompliance of removing prohibited content in a timely manner, such as NetzDG's fifty million Euro fine, are quite steep. Therefore, the statutes incentivize the removal of online content on platforms. Second, the timeframes present in almost all the European regulations discussed—NetzDG's and the French Avia law's twenty-four-hour deadline, along with TERREG's

---

[113] *See* Paris Martineau, *The UK's Tech Backlash Could Change the Internet*, WIRED (Apr. 9, 2019, 7:00 AM), https://www.wired.com/story/uk-tech-backlash-could-change-internet/.

[114] *See* Hoboken, *supra* note 91, at 3–4, 6 (defining "terrorism content" as meeting one or more of the following:

> (a) inciting or advocating, including by glorifying, the commission of terrorist offences, thereby causing a danger that such acts be committed;
> (b) encouraging the contribution to terrorist offences;
> (c) promoting the activities of a terrorist group, in particular by encouraging the participation in or support to a terrorist group within the meaning of Article 2(3) of Directive (EU) 2017/541;
> (d) instructing on methods or techniques for the purpose of committing terrorist offences.)

[115] *See id.* at 4, 6.

[116] *See id.* at 6.

[117] *See* Langvardt, *supra* note 110, at 352; *see* Hoboken, *supra* note 91, at 8.

one-hour timeframe—are simply too rigid.[118] It is highly unlikely that moderation of such flagged content can be well-addressed within such a short window. Instead, these time frames encourage platforms to minimize the review of such content and err on the side of caution by choosing to simply remove the content.

Further, such over-censorship, caused by ambiguous regulations and implemented by platforms will, arguably, backfire as prominent cases of deletion may fuel anti-government sentiment and/or publicize the deleted material broadly.[119] For instance, in January 2018, Twitter removed the post of a notable, far right German politician under NetzDG, leading to wide-spread media coverage, which included the post's content and potential illegality of the law itself.[120] Thus, the content of the tweet, which was illegal under NetzDG, was widely advertised and it provided the politician a platform to criticize the government.

Overall, the reasons provided here demonstrate why the EU's policies, as currently written, will likely lead to over-censorship of individual speech. The next section explains a more troubling critique of Europe's over-regulatory policies: privatization of statutory enforcement.

2. Privatized Enforcement[121]

In each of the European regulations, the law asks private corporations to stand-in for courts and public prosecutors in addressing

---

[118] *See* Hoboken, *supra* note 91, at 8.

[119] *See* Tworek & Leerssen, *supra* note 70, at 3.

[120] *See* Ivana Kottasova, *Twitter Blocks Far-Right Leader as Germany Tightens Hate Speech Law*, CNN BUSINESS (Jan. 2, 2018, 12:10 PM), https://money.cnn.com/2018/01/02/technology/twitter-facebook-germany-hate-speech/index.html; Cristina Maza, *Twitter and Facebook Shut Down Anti-Muslim Posts by Far-Right Party in Germany*, NEWSWEEK (Jan. 2, 2019, 11:18 AM), https://www.newsweek.com/twitter-facebook-anti-muslim-posts-germany-afd-767869; Jill Petzinger, *A Far-Right Politician Was the First to Fall Foul of Germany's Online Hate-Speech Laws*, QUARTZ (Jan. 2, 2018), https://qz.com/1169423/far-right-alternative-for-germany-politician-beatrix-von-storch-had-her-posts-deleted; Philip Oltermann & Padraig Collins, *Two Members of Germany's Far-Right Party Investigated by State Prosecutor*, THE GUARDIAN (Jan. 2, 2018, 8:46), https://www.theguardian.com/world/2018/jan/02/german-far-right-mp-investigated-anti-muslim-social-media-posts.

[121] Though this critique may seem contrary to argument that self-execution by platforms is the best option, this Article later discusses the ways in which government regulation may be implemented alongside platforms creating and applying content moderation policies to combat the critiques described in this section.

the legality of online content on their platforms.[122] Such expansive and harrowing designation of power to private platforms is problematic as it delegates executive and judiciary functions to companies, creating significant issues for individual rights, such as a lack of public accountability and due process.[123]

Specifically, under European regulations, platforms' executing content moderation decisions and appeals procedures of individual grievances serve at the forefront of government regulation of online speech and conduct.[124] Specifically, platforms are tasked with applying criminal and other law under short deadlines and threat of heavy sanctions and this incentivizes faster and greater takedowns of individual speech.[125] Further, these regulations provide individual users little or no remedy to address these removals, such as an appeals system, and, if there are some protections, they are insignificant guarantees for protecting individual freedom.[126] Essentially, arbitrary regulatory language leads to governments delegating greater power to platforms over individual speech, providing them with a legal source for violating individual rights, all without any real public accountability.

TERREG is a concrete example of privatized enforcement and its consequences. In TERREG, the referral procedure requires platforms to address law enforcement notifications, which essentially codifies law enforcement's informal referrals and subsequent extralegal removal of information online based on platforms' community guidelines.[127] This undermines due process safeguards for individuals because community guidelines are broad and more loosely enforced than the law.[128] Due process is further thwarted by the proposal requiring platforms to set up automated tools to proactively screen, filter, and refuse/permit individuals to post content before it is uploaded onto the platform.[129] In fact, such pre-emptive censorship would be based on AI, which has issues assessing contextual posts as discussed above, and, thus, creates serious issues for due process, in terms of unwarranted censorship. Moreover, the proposal does not provide procedures for individuals to make referrals outside of the platforms themselves and, consequently,

---

[122] *See* Peter Pomerantsev, *How (Not) to Regulate the Internet*, THE AMERICAN INTEREST (June 10, 2019), https://www.the-american-interest.com/2019/06/10/how-not-to-regulate-the-internet/; Tworek & Leerssen, *supra* note 70, at 3; Hoboken, *supra* note 91, at 7; Balkin, *supra* note 105, at 2029.

[123] *See* Pomerantsev, *supra* note 122; Balkin, *supra* note 105, at 2031.

[124] *See* Balkin, *supra* note 105, at 2029.

[125] *See* Bukovska, *supra* note 88, at 2.

[126] *See* Pomerantsev, *supra* note 122; Tworek & Leerssen, *supra* note 70, at 3; Hoboken, *supra* note 91, at 7–8; Balkin, *supra* note 105, at 2030–31.

[127] *See* Hoboken, *supra* note 91, at 7.

[128] *See id.*

[129] *See id.* at 8–9.

disallows an independent assessment of private content moderation decisions.[130] Such proposed regulation, which allows for broad delegation in these aforementioned ways, demonstrates the dangers of privatized enforcement: undermining due process and accountability. Overall, as regulations delegate the actual moderation decisions to private platforms, governments are essentially empowering platforms with executive and judiciary functions, creating serious questions when it comes to the effects on individual rights. In fact, together, such privatized enforcement, breadth and vagueness of these policies, and their capacity to overregulate individual speech, creates a wonderful combination ripe for adoption by authoritarian governments

3. Permission for Authoritarian Governments

As previously stated, regulating the modern internet is a recent development, and Europe, as a representative of democratic states, is taking the lead. Consequently, as with any regulatory pioneer, other countries are taking their cues from Europe, including, most worryingly, authoritarian governments.

These autocratic countries are using the loose regulatory actions and surrounding rhetoric regarding these regulations, among European democracies, to fuel their repressive agendas.[131] Specifically, European statutes are being adopted as precedent for authoritarian regimes to repress online speech, using these statutory schemes to take control of the online platform space and punishing individuals for speech they disagree with online.[132] Demonstrating this trend, in mid-2017 Russia copied passages directly from NetzDG for its own online anti-terror law, which required internet providers to save the content of all communications for six months.[133] This illustrates the very real consequences of passing such vague regulations, which provide few protections for individuals, while justifying those very regulations with rhetoric that these statues are stopping harms and mitigating dangers.

Even though authoritarian governments would find a way to repress speech without these regulations from Western countries, the over-regulatory, vague language derived from the statutes and rhetoric surrounding their passage allow authoritarian regimes to utilize the same terminology to justify their own harsh, repressive political

[130] *See id.* at 7.

[131] *See generally* Tworek & Leerssen, *supra* note 70, at 4.

[132] *See id.*; *see also Social Media: How Do Other Governments Regulate It?*, BBC: TECH (Feb, 12, 2020), https://www.bbc.com/news/technology-47135058.

[133] *See* Tworek & Leerssen, *supra* note 70, at 4.

agendas.[134] In addition to the aforementioned critiques of European regulation, a broad characteristic of such legislation, arising as a result of vague language and privatized enforcement and utilized by authoritarian governments, is that the statutes fail to guide platforms in their regulation of online speech.

4. Lack of Substantive Guidance

Content moderation policies are not easily developed; instead, it is a complex process. The European regulations implicitly recognize this complexity by utilizing vague language and delegating the creation of specific moderation policies to platforms, as explained above.[135] Thus, the regulations do not provide any substantive guidance to platforms, rather they only succeed in exerting pressures, which leads to over-censorship, and broadcasting their regulatory agendas. This lack of benefit begs the question as to why such over-regulatory policies are needed in the first place, especially since platforms are already moderating such content.

Foremost, when developing policies that govern content moderation, drawing lines as to when, how, and why platforms should intervene is difficult.[136] Specifically, societal expectations and norms change all the time and trying to figure out these norms, in addition to creating a policy moderating them, is highly challenging.[137] For instance, ascertaining what is "harmful" or "hate speech" or "terrorist content" is dependent on context and how a platform's communities, much less society itself, defines these terms. Of course, there is certain content which is easily identifiable as "hate speech," such as calling for genocide of a specific race or religion, but most content is not so clear-cut. Hence, ascertaining what appropriately falls under these categories and creating content moderation rules surrounding them is a burdensome task, which can lead to high-profile mistakes. Such intricacy is a significant reason why European governments, which have passed regulations calling for the ban of "hate speech," "harmful content," or "blasphemy," fail to define these terms and how such terms should be moderated online.

Despite these challenges, a law or regulation's failure to define terms that it bans, such as "harmful content," creates its own set of

---

[134] *See* Pomerantsev, *supra* note 122. A contrary proposal to prevent authoritarian governments from using Western regulatory statues is by framing regulation in terms of the rights of people on the internet, instead of framing it in terms of stopping harms and mitigating dangers. *Id.*

[135] *See id.*

[136] *See* GILLESPIE, *supra* note 1, at 5.

[137] *See* Klonick, *supra* note 7, at 1627–28.

issues, as explained above, but the bottom-line is that European governments are essentially leaving it to platforms to define these terms and implement them through policies.[138] So, in actuality, platforms are not provided any set standards as to the policies, such as banning terrorist content, they are required to enforce. Instead, the statutes simply tell platforms that regulation of these categories of speech is important and force the platforms to do something about it. Thus, the only practical effect the regulations have is to demonstrate that governments want platforms to moderate these areas: showcasing their public agenda. The statutes, proposed or passed, do not have real substantive guidance.

Despite this critique, it is important to recognize why such regulations were passed. The European governments felt the platforms were not listening to their concerns and any regulation, without subsequent sanctions, were simply publicity stunts. These concerns are valid, but the ways in which European countries have approached regulation is problematic.

Overall, despite Europe's efforts to make platforms safer for individuals, their over-regulatory approach is problematic for multiple reasons, including over-censorship, the privatization of regulation, providing permission to authoritative governments, and lacking substantive guidance for platforms. Hence, such consequences, illustrate that platforms should create and implement content moderation operations themselves.

C. Global Implications of Content Moderation

Along with the concrete limitations of American law on potential regulation and the consequences emanating from Europe's regulatory approach, the global nature of platforms demonstrates that they should be left to self-execute content moderation policies. This is because confining platforms' content moderation policies to the law of one nation or union is both impractical and unreasonable, as such regulation affects more than the country promulgating it.

Platforms are global enterprises that must manage multiple communities, across numerous nations and cultures, all of which mingle with each other.[139] To demonstrate the scale of each platforms' reach, it is illustrative to assess the platforms' number of active, monthly users:

[138] *See* Balkin, *supra* note 105, at 2031.
[139] *See* GILLESPIE, *supra* note 1, at 76.

Facebook has 2.6 billion users, YouTube retains 2 billion users, and Twitter, albeit smaller but still global, boasts 326 million users.[140]

Further demonstrating their global nature, it is important to note that platforms are subject to a wide variety of laws and competing values. This is best shown in how platforms are held responsible for the content their users post across the globe. In the US, platforms are subject to section 230 of the Communications Decency Act, whose liberal approach does not create any liability for platforms.[141] In contrast, European countries apply conditional liability, where platforms are not liable for users "as they have no 'actual knowledge' of and did not produce or initiate the illegal or illicit material, [but] they must respond to requests from the state or courts to remove illicit third-party content."[142] Even further, China and multiple nations in the Middle East utilize strict liability, which requires platforms to proactively remove or censor content, often while cooperating with the government.[143] Thus, platforms are definitively global enterprises, and, as such, their content moderation policies, which determine what content can be accessed on their platforms, impact users across the world.

Consequently, regulations that limit the substance of platforms' content moderation policies do not simply affect the country passing such legislation. Instead, such regulations affect platforms' approach to content moderation for all. For instance, when Facebook removes content under a country's statute, it does not geoblock the content, i.e., simply removing the content within a certain geography; rather, it removes the content globally.[144] Consider NetzDG, which applies only to users in Germany or German citizens. However, in its application, it is difficult to implement the policy to ensure that it does not affect others, especially when it is hard to determine whether a user is a German citizen or when German citizens are in another country. This means that other users' speech will be subject to the substantive parameters found in NetzDG. In fact, the substantive restrictions provided by the European regulations, both those of the individual countries and the EU, will most likely affect populations and citizens other than those countries and the EU itself, causing gross over-censorship. Hence, these regulations do not work in isolation, limited in application to one country; instead, they bleed into affecting others as well.

---

[140] J. Clement, *Most Popular Social Networks Worldwide as of July 2020, Ranked by Number of Active Users*, STATISTA (Aug. 21, 2020), https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[141] *See* 47 U.S.C. § 230 (2018).

[142] *See* GILLESPIE, *supra* note 1, at 33.

[143] *See id.*

[144] *See* Klonick, *supra* note 7, at 1651.

Further, even if platforms are managing to follow content moderation regulations promulgated by different counties, each with its own standards and law, the question remains as to whether this is the best approach for globally situated platforms.[145] With multiple countries passing legislation to shape platform moderation policies, it is an ultimate act of hubris to try to shape global moderation policy based on one country's substantive interpretations of what content moderation should look like. It assumes that the users in different countries share the values touted by the countries passing legislation.

Platforms have approached content moderation in this way, and it has been problematic. Specifically, platforms value American norms of free speech and civil discourse, which require people to tolerate a great deal of speech. When applying American-based content moderation policies internationally, platforms found that users across the world did not share their values.[146] As a result, platforms developed content moderation policies that restrict speech further than the First Amendment. Hence, it is similarly unreasonable to try to shape platforms' moderation policies on the interpretations of countries passing such regulations: doing so raises serious international concerns.[147]

Platforms are global enterprises; their decisions change the speech of billions online and any regulations affecting the substance and application of content moderation alter more than just the population of the government passing it. So, to prevent one country from attempting to hijack platform moderation policy, it is better to leave the creation and application of content moderation to platforms themselves.

Overall, this section argued that platforms are best suited to create and implement content moderation policies by explaining the limitations of American law, which prevents any European-like legislation; the unwanted consequences of heavy European regulations; and the global implications of one country or union regulating platforms' content moderation policies. However, even though platforms may be best suited to handle the particulars of content moderation, this does not mean that governments should not regulate

---

[145] *See id.* at 1650–51. For instance, YouTube only removes material when it violates the law of the country and limits the removal to the geographical bounds of that country. *Id.*

[146] *See* Klonick, *supra* note 7, at 1623; GILLESPIE, *supra* note 1, at 11; Balkin, *supra* note 105, at 2030–31.

[147] *See* Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939, 964 (2020) (provides a response to the concern raised by countries' effecting global platform content moderation policies by suggesting an alternative source of law for content moderation: utilizing a human rights framework in creating such policies.)

platforms. Rather, regulation should be shaped by the nuances of the issues presented by content moderation.

VI. COMPLIMENTARY GOVERNMENT REGULATION TO PLATFORMS' CONTENT MODERATION EFFORTS

Platforms should self-regulate and exercise the creation and implementation of content moderation policies. However, this does not mean platforms themselves should not be regulated at all, instead there are other ways to redefine the relationships between governments and platforms. This section endeavors to demonstrate that when platforms create content moderation policies, they are opaquer about their processes and, in the application of such policies, offer less protection for individuals than governments. Thus, governments should compel transparency regarding both the decision-making and rulemaking processes of platforms, along with the policies themselves, and implement procedural protections, as well as a clear appeal process for users.[148]

A. Why Legislation is Needed

So far, platforms have neither disclosed their decision-making and rulemaking procedures, nor the exact content moderation policies moderators utilize. Further, the procedural protections platforms provided for users, whose speech is moderated, do not necessarily encompass proper due process. This needs to change: the public must see these details in order to both advocate for the policies they believe are best and police the behavior of platforms.[149] Additionally, governments must compel platforms to employ greater procedural due process mechanisms throughout the moderation process. Specifically,

[148] *See* SARAH T. ROBERTS, BEHIND THE SCREEN (2019); Casey Newton, *The Trauma Floor*, VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Another area of legislation is to regulate the working conditions of content moderators, those housed in the United States, by platforms who are both directly employed by platforms and through third parties. *See* ROBERTS, *supra* note 148; *see also* Newton, *supra* note 148. Specifically, the psychological effects of content moderation have been found to be extremely debilitating for moderators. *See* ROBERTS, *supra* note 148; *see also* Newton, *supra* note 148.

[149] *See* GILLESPIE, *supra* note 1, at 44, 199; Koslov, *supra* note 37, at 208–09; Klonick, *supra* note 7, at 1665; Kate Klonick, *The Most Important Lesson from the Leaked Facebook Content Moderation Documents*, FUTURE TENSE (June 29, 2017), https://slate.com/technology/2017/06/the-most-important-lesson-to-learn-about-facebook-content-moderation.html.

to demonstrate why legislation is required, it is important to assess the power of platforms and the non-neutral utilization of that power.

With much of society's public discourse, cultural production, and social interaction occurring online, the power which a handful of large, privately owned platforms retain is enormous.[150] To illustrate, Facebook's terms of service govern over one-fourth of the world's population online, YouTube streams a billion hours of video daily to two billion users worldwide, and Twitter is the primary medium by which the President of the US uses to communicate with the nation.[151] Hence, each decision surrounding the moderation process exhibits substantial influence over individual speech and public conversation.[152]

The dangers of consolidating this power on platforms is best demonstrated by the fact that platforms do not use their content moderation policies neutrally, instead their economic motivations shape their moderation choices. Naturally, platforms, as private companies, are economically motivated to make a profit and reassure advertisers.[153] As explained above, this is a large reason as to why content moderation systems were established in the first place. However, in addition to ensuring user safety and engagement, platforms shape their moderation decisions to protect their brand, ensure growth in other countries, and mitigate their liability in various countries; thereby cumulatively using their content moderation policies as a gatekeeping mechanism to control the nature of user-generated content.[154]

To demonstrate, take for instance Facebook's actions. Initially, when it launched its nudity policy, it banned breastfeeding photos in order to protect its brand and to ensure it remained user-friendly for all populations. However, when faced with the reality that many on Facebook wanted to share their breastfeeding photos, it took the breastfeeding community years of activism to change Facebook's policies.[155] This type of gatekeeping directly contradicts Facebook's externally advertised commitment to free speech and broader commitment to its user base; instead, it showcases that Facebook did not want to alienate potential advertisers and users, whose data they mine. Moreover, Facebook has warned moderators in Pakistan against creating a "PR fire" and cautioned against doing anything that may have

---

[150] *See* GILLESPIE, *supra* note 1, at 6.

[151] *See* Annemarie Bridy, *Remediating Social Media: A Layer-Conscious Approach*, 24 B.U. J. SCI. & TECH. L. 193, 195 (2019).

[152] *See id.* at 195–96.

[153] *See* GILLESPIE, *supra* note 1, at 11.

[154] *See* Koebler & Cox, *supra* note 53; Aodhan Beirne, *5 Takeaways From Facebook's Leaked Moderation Documents*, N.Y. TIMES (Dec. 27, 2018), https://www.nytimes.com/2018/12/27/world/facebook-moderators-takeaways.html.

[155] *See* GILLESPIE, *supra* note 1, at 168.

a "negative impact on Facebook's reputation or put it at legal risk."[156] Such actions demonstrate that platforms use content moderation to promote their economic goals. Although these motivations may be natural, by placing profits at the forefront of moderation decisions, considerations of the rights of users always remain secondary, leading to the development of moderation policies and decision-making that is not always conducive to free speech.

Further demonstrating platforms' biased use of moderation is the fact that moderation policies are applied inconsistently across both individuals and countries—waived whenever necessary or convenient.[157] In fact, platforms are increasingly providing preferential treatment to some users over others where they are creating special rules for public figures, such as powerful actors and organizations, or newsworthy events.[158] For instance, when India militarized against protestors in Kashmir, Facebook posts of Kashmir activists were being deleted and members of a group, called Kashmir Solidarity Network, found that their Facebook accounts had been blocked on the same day, without any explanation as to why.[159] Hence, it is clear that platforms are unafraid to change the way in which they apply moderation policies in a partisan way, through unexplained internal decisions.

Similarly, platforms are biased in their approach to how they apply moderation policies among countries. This is demonstrated by the manner in which platforms behave in the EU and the US compared to how they act in developing countries.[160] In fact, the only way countries outside the US receive sustained attention from platforms is if those countries have a large market, like those of China or the EU; if journalists uncover significant human rights violations; or if countries threaten platforms with sanctions.[161] For instance, several European countries have created agency relationships with platforms because they brought about legislation to hold platforms responsible for their actions. Specifically, Germany initially tried out a voluntary compliance system for platforms' compliance with their concerns about online speech and found the results to be ineffective, as platforms were not taking its

---

[156] Beirne, *supra* note 154.

[157] *See* Balkin, *supra* note 105, at 2025.

[158] *See* Klonick, *supra* note 7, at 1665.

[159] *See* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

[160] *See* Davey Alba, *How Duterte Used Facebook To Fuel the Philippine Drug War*, BUZZFEED NEWS (last updated Sept. 4, 2018, 2:19 PM), https://www.buzzfeednews.com/article/daveyalba/facebook-philippines-dutertes-drug-war.

[161] *See* Tworek & Leerssen, *supra* note 70, at 9.

concerns seriously.[162] Ultimately, Germany chose to enact stringent legislation, NetzDG, and only then did Germany successfully gain platform compliance.[163]

In contrast, in developing countries, where platforms have no competition and governments have little leverage, platforms act without much consequence.[164] For example, in countries such as Myanmar, Sri Lanka, and the Philippines, Facebook usage and delayed platform response to government concerns has led to rampant violence.[165] Though nefarious use of Facebook is not limited to developing countries, Facebook consistently has not acted or responded to government concerns, stating that the violence-inducing content did not violate community standards or acting far too late to moderate the content, unlike in the EU or the US.[166] The reality is that developing countries are small markets and retain little leverage over Facebook or any other platform. Therefore, local governments cannot successfully address such lack of or delayed response to real issues with platforms.

Such inequality in the application of content moderation policies demonstrates that platforms act differently within different countries, for reasons hidden in internal, private decisions. The amount of power platforms retain across the world and the uneven application of moderation among users and countries illustrates why legislation is needed, both in terms of transparency and procedural due process.

B. Legislation Mandating Transparency

This section advises that legislation should compel two types of transparency: disclosure of the processes behind creating and making decisions in the use of content moderation policies, and the exact policies platforms use internally. The section further clarifies how decision-making and rulemaking occurs today; first, through small, homogeneous teams at platforms' headquarters and, second, via content moderators who are subjected to both complex, confusing moderation policies and communication from platforms regarding the use of moderation policies. This discussion endeavors to show that

---

[162] *Id*.

[163] *Id.*

[164] *See* John Naughton, *Facebook's Global Monopoly Poses a Deadly Threat in Developing Nations*, THE GUARDIAN (Apr. 29, 2018), https://www.theguardian.com/commentisfree/2018/apr/29/facebook-global-monopoly-deadly-problem-myanmar-sri-lanka.

[165] *See* Alba, *supra* note 160; Naughton, *supra* note 164; Amanda Taub & Max Fisher, *Where Countries are Tinderboxes and Facebook Is a Match*, THE NEW YORK TIMES (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

[166] *See* Taub & Fisher, *supra* note 165.

transparency is required to ensure accountability to users, to demonstrate why and how platforms arrived at their rules and changes as they self-regulate, to create consistency in application of such policies, and to allow the public to informally critique such decision-making and the policies themselves.[167]

Generally, relatively small internal teams within platforms oversee the rulemaking and decision-making processes governing content moderation.[168] These teams have an enormous influence, as they decide how the proverbial lines of content moderation policies are drawn, the punishment for users who violate the policies, and the overall philosophical approaches platforms take to governance.[169] Yet, who these people are, how they do their work, and how users can reach them is deliberately obscure; instead the policies they create or change are often released in the platform's voice.[170] All the public knows is that the policies, affecting people's speech worldwide, and the implementation of such policies is "overseen by a few hundred, largely white, largely young and tech-savvy Americans who occupy a small and tight social and professional circle."[171]

The creation of such universal policies in the hands of the few, without public supervision and accountability, is problematic for the following reasons: a small team of homogenous people unilaterally create the moderation policies, the decision-making process is opaque and likely inconsistent, and the internal, detailed policies moderators use are not publicly available.

First, the development of these policies depends on different values, ideologies, and various, competing politics of culture; thus, leaving the creation of such rules to a small group of people is problematic.[172] There are many questions that factor into figuring out where and how to moderate online speech, such as "what [content] is unacceptable . . . [,] reconciling competing value systems[,] mediating when people harm one another, . . . [and] grappling with [social inequalities.]"[173] When these rules, which require difficult, nuanced decisions that affect the speech of populations worldwide, are created by small homogenous teams that share a particular worldview, it is likely that they do not apply well to "different experiences, cultures, or value systems."[174]

---

[167] *See* GILLESPIE, *supra* note 1, at 169–70; Koslov, *supra* note 37, at 208–11.
[168] GILLESPIE, *supra* note 1, at 117.
[169] *See id.*
[170] *See id.* at 117, 119.
[171] *See id.* at 119.
[172] *See id.* at 10.
[173] *Id.*
[174] *See id.* at 8.

This is an issue as an absence of culturally and politically specific considerations can create biased moderation policies.[175] For instance, Facebook's 2017 training material instructed moderators to identify hate speech using a formula: protected category and an attack equals hate speech.[176] The protected categories mirrored those of American law, but the guidelines provided greater leeway for posts that only refer to subsets of protected categories, which, in reality, meant that white men received greater protection than female drivers or black children.[177] This occurred because gender and race were both protected, but occupation and age were not. Such issues with the homogeneity of the internal platform team and the biases written into the rules themselves may be lessened or avoided with public critique of the rulemaking and decision-making process and the actual, internal policies themselves. This would be possible through mandating transparency as to the process and rules themselves. Even though the individuals creating these policies within platforms may be impressive, they are playing a legislative role for billions of individuals who do not currently have a say or ability to monitor rule development, creating a real lack of accountability. Again, this can be countered through mandating transparency.

Second, for high-profile moderation cases where the higher-up, internal teams become involved in the decision-making, instead of leaving it to the every-day content moderators, issues may arise regarding the consistency of decision-making. For instance, at Facebook, when making high-profile decisions about moderation in a situation in which there is no policy to fall back on, internal decisionmakers may sometimes make "spirit of the policy" decisions that fit the contours of other decisions the company has made.[178] Such higher-up decisions are not only opaque, but also may be inconsistent with the platforms' internal, current policies.

In addition to the inherent issues with rulemaking and small group decision-making, the dissemination and textual content of these rules, along with the inherent challenges of content moderation, affect how every-day content moderators make decisions. In fact, such trends allow moderators to make biased, inconsistent decisions. Specifically, moderators have several sources with which they may consult when making decisions, that may be ambiguous and confusing. Assessing Facebook, moderators have several overwhelming, sometimes conflicting sources: the community guidelines, both public and internal; additional commentary and guidance by platforms on tough questions

[175] *See* Koslov, *supra* note 37, at 191–92.
[176] *See id.* at 191.
[177] *See* Koslov, *supra* note 37, at 191–92.
[178] *See* Koebler & Cox, *supra* note 53.

of moderation; discussion moderators have amongst themselves; and Facebook's internal tools to distribute information during breaking news events.[179] Consequently, the wide variety of authorities to choose from, buttressed by a lack of required transparency in that choice, creates a great deal of uncertainty as to how content moderation decisions are made and how on-the-spot rulemaking occurs in emergency situations. At the same time, it contributes to dissimilar treatment of similar content posted on the platform.

Further, the dissemination of rules to content moderators, who may be directly employed by the platform or by third-party contractors, is riddled with issues. For example, at Facebook many moderators state that there is a communication disconnect between Facebook's policymakers, employees who are supposed to communicate the policy, and the senior staff at outsourced firms charged with explaining it to moderation teams[180] This is problematic as internal policies change frequently and, in fact, Facebook relies on these changes to address highly contextual or nuanced cases. [181] Moreover, moderators further report that there is no overarching guide to these new changes as they are announced.[182] Sometimes outsourced firms misinterpret an update in policy, Facebook guidance comes too late, or Facebook may issue a policy update that it subsequently changes or retracts entirely.[183] For instance, when there is a breaking news event, such as a mass shooting, managers will often use Facebook's internal tool to distribute information, but they may end up posting conflicting information about how to moderate content.[184] Consequently, such a lack of streamlined communication regarding new rules and changes creates uncertainty that leads to inconsistent decision-making.

Moreover, there are fundamental challenges to content moderation which allow for even more inconsistent decision making. Specifically, the number of posts moderators must review; the time in which they must review an individual post, ranging eight to ten seconds; daily changes and clarifications to the moderation rules; a lack of cultural and political context, which makes the meaning of posts ambiguous; and frequent disagreements between moderators regarding the application of certain rules all contribute and perpetuate the challenge of making

[179] *See* Newton, *supra* note 148.

[180] Katie Notopoulos, *Burt's Bush and XXXTentacion's Death: Why Facebook Moderators Fail*, BUZZFEED NEWS (last updated Sept. 23, 2019, 5:30 PM), https://www.buzzfeednews.com/article/katienotopoulos/facebook-moderators-are-set-up-to-fail.

[181] *See id.*

[182] *See* Koslov, *supra* note 37, at 190–91.

[183] *Id.*

[184] *See* Newton, *supra* note 148.

consistent, objective decisions for moderators.[185] Such opacity in the implementation of content moderation politics, both on a higher corporate and individual moderator level is problematic as to the biases behind the decisions and inconsistency in the decision-making. Hence, transparency is needed to allow the public to comment on and improve these processes.

These aforementioned issues are exacerbated by the fact that platforms do not share their detailed internal content moderation guidelines with the public, even though they change much more frequently than the community guidelines and are far more comprehensive.[186] Consequently, it is difficult to understand corporate decisions that create or modify moderation rules, except in the most obvious cases.[187] In fact, without assessing the policies themselves, it is challenging for the public to critique platform's decision-making and rulemaking processes. Therefore, mandating transparency of these policies will allow the public to comment upon and, hopefully, inform platforms about the lack of cultural and political context and biases implemented in the rules themselves. In the end, if society wants self-regulation to be a fair and equal process, platforms must allow the public to access the internal rules themselves in order to create a more democratic process.

For these reasons, transparency is needed to keep platforms accountable for the rule creation and moderation decision processes. Transparency ensures these decisions and rules are consistent with one another and that they strike the right balance between nuance and context. Such legislation may take the form of disclosure requirements of how the moderation policies are created by platforms thorough record-keeping and internal memoranda, which track the reasoning behind all platform rulemaking decisions. These may include emergency and high-profile decisions and the manner of policy dissemination to content moderators. This legislation should further be complimented by the disclosure of the detailed, internal moderation policies themselves, allowing the public to critique the rules, which will likely add greater nuance and regional commentary which the internal teams may have missed. In addition to statutes that advocate for transparency, the government should pass legislation that provides greater procedural due process for users.

[185] *See* Newton, *supra* note 148; Beirne, *supra* note 154.
[186] *See* Klonick, *supra* note 7, at 1639.
[187] *See id.* at 1615.

C. Legislation Mandating Procedural Due Process

Platforms retain an immense amount of power over individual speech, yet there is little in the way of due process to help individuals protect their rights. In fact, an individual user's ability to appeal a decision on content takedown, account suspension, or deletion varies widely between the platforms.[188] Consequently, there is a real difference between the rules platforms post and the decisions they make on a case-by-case basis in practice.[189] So, the government must implement legislation that provides individuals with greater procedural protection during the moderation process. This section describes how procedural due process works now, along with its skewed effect on marginalized communities, and supplies suggestions as to what procedural due process may look like in the content moderation process.

1. How Procedural Due Process Works Today

This section describes how platforms lack procedural due process when posts are flagged, assesses specific AI-generated decisions as to whether or not to delete certain posts, analyzes specific critiques surrounding the rules' ambiguity, and describes appeals processes enacted by platforms.

To demonstrate the need of procedural due process through legislation, it is instructive to assess how little process occurs with AI screening of users' posts. As described above, AI is not a neutral tool, instead it is highly subjective, and platforms utilize it on a large scale to scan and moderate posts. Specifically, AI cannot identify and remove content without human bias in any language because the tools and algorithms, regardless of their complexity, are designed, tested, maintained, deployed, and overridden by people.[190] So, AI likely carries forward the biases or assumptions of the person or persons creating the algorithms and conforms to extend those biases to decision-making.[191] However, when it makes these moderation decisions, AI does so automatically, without providing the person who was affected the substantive feeling of due process through an explanation for its decision. This may be sufficient if there were greater procedural process granted after posts are flagged by AI, or through community flagging, but platforms lack this post-flagging procedural process as well.

---

[188] *See* Klonick, *supra* note 7, at 1648; Balkin, *supra* note 105, at 2025.
[189] GILLESPIE, *supra* note 1, at 72.
[190] *See id.* at 97.
[191] *See id.* at 104–05.

In fact, there are several factors that prevent adequate process when content moderators review posts after being flagged by AI or the community. These include ambiguity in the flagging itself and the moderation policies, along with the challenges in the application of such rules by moderators. It is important to note that content moderators, here, act very similarly to judges as they are trained to exercise professional judgement about the application of the rules and apply these rules, which often take the form of multi-factor tests.

Specifically, the moderation guidelines and the user-flagged posts allow platforms room to honor the flagging in some cases and overrule it in others, without explaining why to the user.[192] This phenomenon occurs because the moderation guidelines are ambiguous and, when users flag posts, moderators are not privy to the post's context or the user's degree of concern.[193] Specifically, platforms can take this ambiguity and skew it to help legitimize a decision or explain away discrepancies when the platform wants to make a different decision, if the platform wants to provide an explanation at all.[194] Such opacity makes it difficult to understand why particular decisions are made, why certain rules are inconsistently enforced, and how the process can be improved for both the user who wants to dispute the decision and the platforms themselves.[195]

Moreover, the moderation policies themselves are biased, ambiguous, and open to the moderators' interpretation. First, both the community guidelines and internal policies are open to prejudice, even if platforms try to create objective rules. For instance, Facebook decided years ago that the rules it provides moderators should be objective as often as possible, as the policies are applied globally, but what is considered "objective" remains up to Facebook's interpretation and decision.[196] Second, community guidelines are riddled with general, imprecise language, and the lack of detail leaves content moderators without guidance when trying to navigate difficult moderation questions.[197] Third, internal guidelines remain ambiguous, confusing, and sometimes wrong. The leaked internal documents created for content moderators from Facebook are especially illuminating. They consist of rules that contain ambiguous language, complex charts, and many complicated instructions.[198] The instructions usually consist of formulas, a series of yes or no multiple-choice

[192] *See* GILLESPIE, *supra* note 1, at 96.
[193] *See id.*
[194] *See id.* at 97.
[195] *See id.* at 139.
[196] *See* Koebler & Cox, *supra* note 53.
[197] *See* Koslov, *supra* note 37, at 189.
[198] *See* Notopoulos, *supra* note 180.

questions to follow, which try to encapsulate human communication and instructions dictating when moderators ought to intervene.[199]

Therefore, the degree of ambiguity and opaqueness in community guidelines, internal policies, and decision-making processes leads platforms to make unilateral decisions regarding the removal of speech, without allowing input from users or providing any further explanation. For instance, Facebook states that users who violate its community standards may experience different consequences depending on the severity of the offense and the user's history.[200] Yet, it does not consider whether the history under consideration arose out of erroneous content flagging or removals, which may have been caused by inconsistent enforcement or erroneous judgement.[201] Moreover, Facebook states that platforms may provide content to law enforcement to prevent real world harm, without elaborating as to what such instances may entail.[202] Such actions are contrary to protecting user rights and providing due process.

It is important to acknowledge that platforms have made some effort to create procedural due process by implementing appeals processes, but the current processes are not enough to warrant fair process.[203] Specifically, significant ambiguity remains as to how these appeal processes occur in terms of what evidence these platforms assess, who makes the decision, and how they make the decision. Moreover, platforms may or may not choose to provide an explanation for their appeals decision and, if they do, the platforms determine the level of detail they must provide. Further, the user's interaction with the appeals process is limited to filling out a form on the platforms' website. Without real guidance as to what factors platforms assess during the appeals process, it is difficult to know how to create a persuasive, successful argument and why appeals were denied or accepted. Consequently, users are left in the dark about why their individual speech was tampered with. Though platforms may be experimenting with a greater appeals process, such as Facebook's Oversight Board, currently, users have little power to address the censorship of their

---

[199] *See* Koebler & Cox, *supra* note 53.

[200] *See* Koslov, *supra* note 37, at 189-190.

[201] *Id.* at 190.

[202] *Id.*

[203] *See* Facebook, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, ABOUT FACEBOOK (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/; YouTube, *Appeal Community Guidelines Actions*, YOUTUBE HELP, https://support.google.com/youtube/answer/185111?hl=en; Twitter, *About Suspended Accounts*, TWITTER: HELP CENTER, https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts.

speech online caused by ambiguous, unilateral decision-making by platforms.[204]

Essentially, the lack of procedural due process in all the stages of moderation—flagging, decision-making processes in the removal of certain posts, and appeals—demonstrates that, for this area of platform governance, the user's individual interests are secondary to the protection of the platform. For most users, this is not an issue as their experience with content moderation rules are in the peripheral. However, marginalized voices, unpopular speakers, or international human rights activists have been unduly burdened in their experiences with these arbitrary platform policies.[205]

For instance, consider Facebook's rule that it deletes curses, slurs, calls for violence, and other types of attacks only when they are directed at protected categories, like race or sex.[206] As a result of this rule, Facebook left up a post calling for the murder of radical Muslims while a post from a Black Lives Matter activist attacking whites in general was removed.[207] In fact, though Facebook's policies are designed to defend all races and genders equally, they do not necessarily provide equal protection. Consider the post, "it's not a crime when White freelance vigilantes and agents of 'the state' are serial killers of unarmed Black people, but when Black people kill each other then we are 'animals' or 'criminals.'" Under Facebook's rules, the platform immediately removed this post and disabled the posting account for three days.[208] Facebook provided no explanation for its removal.[209]

These examples demonstrate not only how the problems discussed disproportionality affect certain populations, but they also serve to elaborate on how the lack of procedural due process can be extremely harmful. Without an adequate chance for users to argue their case, their

---

[204] *See* Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, FACEBOOK NEWSROOM (Sept. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure/. It is highly unlikely that the Oversight Board will add procedural due process for users. *Id.* Instead, the sheer volume of content moderation decisions Facebook makes every day means the Court cannot be expected to offer procedural protection or error correction except for the smallest of cases. *Id.* However, the Board may help highlight weaknesses in Facebook's policies and procedural process and, in fact, does provide an independent forum to discuss moderation decisions. *Id.* But even with these positives, Facebook retains the ability to choose which cases it wants to consider; thus, the process is skewed from the beginning. *Id.* Instead, I argue that this is simply a way to legitimate content moderation, state of government regulation, and outsource controversial decisions away from the company. *See generally id.*

[205] *See* GILLESPIE, *supra* note 1, at 8; Langvardt, *supra* note 10, at 1385.

[206] *See* Angwin & Grassegger, *supra* note 159.

[207] *Id.*

[208] *Id.*

[209] *Id.*

speech is censored, and, at times, their accounts are disabled. In a world increasingly reliant on social media as the ultimate space for public debate, this is not a small punishment. Moreover, notwithstanding issues arising from the degree of due process, adequate procedural due process provides an inherent acknowledgement of individual dignity. It communicates that a user has the right to advocate for its speech and deserves an explanation of the platform's moderation decisions. It is what users are owed.

2. Possible Forms of Procedural Due Process

As a result of this failure to provide procedural due process, legislation compelling greater procedural due process by platforms is needed, and this section provides various mechanisms for legislation to impose these minimal protections.

Specifically, legislation ought to require platforms to provide a clear reason as to why a post was removed or an account disabled. Platforms should give users an opportunity to adequately voice their grievances. This should be something more than simply filling out a form online; instead, it may take the form of a hearing. Lastly, platforms should make the factors and processes they use, when deciding a user's appeal, publicly available.

In addition to these suggestions, others have provided further methods to establish procedural due process regarding online speech, in the form of legislation or otherwise. This Article only describes the following two.

First, the model of technological due process advocates for understanding the tradeoffs of automation and human discretion, protecting individual rights to notice and hearings, and providing transparency to rulemaking and adjudication.[210] Second, the Manilla Principles on Intermediary Liability requires, among other things: "(1) clear and public notice of the content-regulation policies companies actually employ; (2) an explanation and an effective right to be heard before content is removed; and (3) when this is impractical an obligation to provide a post facto explanation and review of a decision to remove content as soon as practically possible."[211] These are simply suggestions, albeit slightly different, that maintain the same underlying goal: the assurance of adequate procedural due process.

Overall, platforms lack procedural due process rights to users, which is why legislation is needed to rectify this problem. With issues in each stage of moderation, ranging from the initial flagging to the

[210] *See* Klonick, *supra* note 7, at 1668.
[211] *See* Balkin, *supra* note 105, at 2044–45.

appeals process, due process should be required to ensure that, when individual speech is censored and punishment is provided to users, users have the ability to challenge this decision and receive an adequate explanation.

Today, platforms retain an immense amount of power over creating content moderation policies: they keep the entire moderation process deliberately obscure and provide little protections for users in the form of appeals. This creates a host of issues, from biased rules and decision-making to inconsistent decisions on similar posts. However, legislation mandating transparency regarding the moderation process and greater due process protections for users can address these issues.

VII. CONCLUSION

Today, the United States faces numerous challenges, and so, the issues platforms generated are likely on the legislative backburner. However, in the future, when regulations are being debated, this Article argues that legislators need to approach every issue surrounding social media with nuance, as there is no one solution that addresses the various problems regarding social media platforms.

In order to demonstrate the nuance required, this Article provided an assessment of content moderation. Beginning with explaining current moderation policies used by platforms and various legislative approaches to regulating it, this Article critiqued the current regulations and argued that platforms are best suited to create, apply, and enforce content moderation policies. As a result, this Article further advocated for the promulgation of legislation mandating transparency and procedural due process in order to ensure, overall, that platforms are held accountable for their actions through public supervision and that users are provided their due process. As legislators move forward, this Article cautions against both promulgating any heavy-handed regulation which may lead to over-censorship and, on the opposite side, remaining completely laissez-faire. There is no question that legislation is needed, but it must be the right kind.