ARBITRATING SOCIAL MEDIA CONTENT: A FRAMEWORK FOR BANNING HIGH-PROFILE USERS THROUGH THIRD-PARTY ARBITRATION

*Rachel Gershengoren**

I. INTRODUCTION

Social media websites are defined as "relatively inexpensive and widely accessible electronic tools that facilitate anyone to publish and access information."[1] Nowadays, social media encompasses a significant portion of an individual's daily life. Roughly seven-in-ten Americans use some form of social media,[2] and 82% of the United States (U.S.) population used social media in 2021.[3] Think about your own daily life—how much of your day is consumed by scrolling through different social media apps on your phone? Do you check the news on Twitter, scroll through Instagram the moment you wake up, or fall asleep watching TikTok videos? I know I do all that and more. For better or worse, the internet is intertwined with everything we do. From shopping online to staying up to date with current news, education, and business tools,[4] and content creation, "social media plays a vital role in transforming people's lifestyle."[5] However, as these social media platforms have come to dominate the socio-political landscape,

* Senior Notes Editor, *Cardozo Journal of Conflict Resolution*; J.D. Candidate 2023, Benjamin N. Cardozo School of Law. B.A. Lafayette College 2020. I would like to thank Professor Felix Wu for his insights and review during the development of this Note and the Staff Editors and Editorial Board of the *Cardozo Journal of Conflict Resolution* for their help on editing this Note. Lastly, a special thank you to my fiancé, Ethan, and my family and friends for their unconditional support and encouragement throughout the entire Note process and throughout Law School.

1 John Samuel & S. Shamili, *A Study on Impact of Social Media on Education, Business and Society*, 4 INT'L J. RSCH. IN MGMT. & BUS. STUD. 51, 51–53 (2017).

2 Brooke Auxier & Monica Anderson, *Social Media Use in 2021*, PEW RSCH. CTR. (Apr. 7, 2021), https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [https://perma.cc/9NDW-A276].

3 *Percentage of U.S. Population Who Currently Use Any Social Media From 2008 to 2021*, STATISTA (Nov. 3, 2021), https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/ [https://perma.cc/RJA7-WZNA].

4 Samuel & Shamili, *supra* note 1.

5 *Id.*

questions have arisen as to whether and how we should regulate such content.

On January 6, 2021, the world witnessed one of "the worst example(s) of the impact digital platforms can have on society with the debacle at the U.S. Capitol."[6] When President Trump called to the crowd at his rally that day to march on the nearby Capital, it surely did help spark the deadly riots,[7] but it was not the first or the only spark. In private Facebook groups, Twitter posts and Parlor messages, extremists had been organizing for months before, getting inspired by Trump's online rhetoric.[8] With every tweet and Facebook post by Trump about how the election was stolen and refusals to acknowledge Biden as the forty-sixth president, these activists were getting inspired to plot their violent strike. They were "discussing not only logistics like hotels and rideshares but also sleeping cars and pitching tents should they need to 'occupy' the city."[9] Hundreds of posts leading up to the riots discussed what ammunition to bring, whether there would be medics in case of emergencies, etc., yet the Capitol riots sent shockwaves to millions of Americans—apparently, nobody saw this coming, despite the abundance of threatening social media posts.

What happened at the Capitol on January 6th is just a sliver of the influence high-profile users on social media have on our society. Social media enables the distribution of fake news, manipulation of digital content for political purposes, and promotion of misinformation on elections, vaccines, health emergencies, etc., and allows users with a high following and significant influence to post this type of content.[10] Yet, platforms have not done enough to stop all of this. Social media platforms, although not liable for content posted by users under Section 230 of the Communications Decency Act, have still tried to regulate content through the creation of community standards as well as oversight boards to keep users engaged on their sites and subsequently keep their profits up. The current system of content regulation includes several sources to flag and remove posts and accounts: (1) users, (2) content modera-

---

6 Michael A. Cusumano et al., *Social Media Companies Should Self-Regulate. Now*, HARV. BUS. REV. (Jan. 15, 2021), https://hbr.org/2021/01/social-media-companies-should-self-regulate-now [https://perma.cc/AC5Y-F3JW].

7 Mark Mazzetti et al., *Inside a Deadly Siege: How a String of Failures Led to a Dark Day at the Capitol*, BRIT. COUNCIL (Jan. 10, 2021), https://www.pqblackburn.com/C2/TheCapitolSiege/Reading.pdf [https://perma.cc/GX46-SY9C].

8 *Id.* at 3.

9 *Id*.

10 Cusumano, *supra* note 6.

tors, and (3) automated systems.[11] Nevertheless, these current models of content regulation create issues of bias, lack of context, lack of due process and transparency, among many more. The automated AI systems lack the ability to understand context, often arbitrarily disciplining users, while human content moderators incorporate their own subjective bias when removing posts/accounts, often over-blocking certain groups of users that content moderators are biased toward. Moreover, the criteria for content regulation is unclear for the moderators tasked with removing such content and also for users, which leads to transparency issues and, subsequently, a lack of due process as users get removed without any rhyme or reason.

This current system provides social medial platforms "the flexibility of removing content as it suits them: in the way that best maximizes their profits."[12] As more information is made available about how exactly content is regulated and the effects this poor regulation has on our society, the more legal scholars, Congress, judges, and the public call for a more aggressive and transparent approach to content regulation. Platforms need to become more aggressive at self-regulation while having the flexibility to evolve with the changing social media presence. It is difficult to create one system of content regulation that can fix all the issues with the current system, but my proposal will address a subsection of the larger issue—making sure high-profile users are appropriately banned. My proposal will mimic a fast-track arbitration system that will be outsourced from a third-party dispute resolution center that the social media company will pay for. A tri-panel of arbitrators will hear and decide on whether a high-profile user should be banned from the site, considering the context of the user's account, and why the social media company removed them. Additionally, the arbitrators will provide a reason for the decision to the user and the public to create more transparency and due process. My proposal will be used in conjunction to the current system of content regulation to help alleviate many of these current issues. The goal of my arbitration system is to ensure high-profile users are not over-banned by other users gaming the system to get an influential

---

[11] Jason A. Gallo & Clare Y. Cho, *Social Media: Misinformation and Content Moderation Issues for Congress*, CONG. RSCH. SERV. (Jan. 27, 2021), https://crsreports.congress.gov/product/pdf/R/R46662 [https://perma.cc/TH7H-D9MB].

[12] Nina Brown, *Regulatory Goldilocks: Finding the Just and Right Fit for Content Moderation on Social Platforms,* 8 TEX. A&M L. REV. 451 (2021).

figure removed or under-banned by giving the influential figure a free pass that is associated with their status or following.

Part I of this Note introduces how social media plays a vital role in society and the issues associated with content regulation. Part II explains in more detail how social media providers can choose whether to regulate speech on their platforms through a further exploration of what gives the platforms their immunity from liability, as well as what the current system of moderation is. Part III discusses why there is a need for improved content regulation and why the current system of content regulation is inadequate, focusing on three main issues—lack of context in AI systems, increased bias on the part of content moderators, and a lack of transparency by the intermediaries coupled with and a lack of due process for users. Part III.C explores proposed ideas for content regulation and explains why such ideas of government regulation and amending Section 230 are insufficient to combat the deficiencies in the current model of content regulation. Part IV proposes an arbitration system that increases context and transparency and reduces bias in the current system, with a discussion of the incentives for social media companies to incorporate this system.

II. BACKGROUND

A. *Section 230 of the Communications Decency Act*

Prior to the development of the internet, mass information was spread through newspapers, radios, and broadcast networks where the publisher had the discretion of what was published.[13] "The internet ended the speaker's reliance on the publisher by allowing the speaker to reach his or her audience directly."[14] This broad freedom given to users of social media platforms is based on

---

[13] *The Evolution of the Media*, LUMEN, https://courses.lumenlearning.com/atd-baycollege-americangovernment/chapter/the-evolution-of-the-media/ [https://perma.cc/P5EW-64W8] (last visited Feb. 11, 2022).

[14] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1603–04 (2018), https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf [https://perma.cc/7GET-NC9S] (Prior to the passage of Section 230, cases such as *Cubby Inc v. CompuServe, Inc.* and *Stratton Oakmont, Inc v. Prodigy Services Co.,* "suggested that intermediaries would be liable for defamation posted on their sites if they actively exercised any editorial discretion over offensive speech.").

section 230 of the Communications Decency Act (CDA).[15] Section 230 immunizes websites from legal liability for anything that is posted on their platforms by users.[16] Section 230(c)(1) provides the social media company with immunity from certain lawsuits that are pursued against said provider. Section 230(c)(2) provides immunity to the providers who want to take good faith actions to restrict access to content that they deem "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable."[17] It is important to note that while the immunity Section 230 provides is broad, it is not absolute.[18]

B. *First Amendment Implications*

Despite such exceptions to Section 230, CDA has helped define and expand the impact social media has on our society. By allowing social media providers to be free from liability regarding the user-generated content posted on their sites, it has encouraged the "unfettered and unregulated development of free speech on

---

15 47 U.S.C. §230 (2006) (states, "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."); *see generally* Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293 (2013). *See also* David S. Ardia, *Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act*, 43 LOY. L.A. L. REV. 373, (2010), https://digitalcommons.lmu.edu/cgi/viewcontent.cgi?article=2685&context=LLr [https://perma.cc/V5PZ-GMNZ].

16 *Communications Decency Act Section 230*, ACLU, https://www.aclu.org/issues/free-speech/internet-speech/communications-decency-act-section-230 [https://perma.cc/LR7Q-HS2X] (last visited Sept. 14, 2021).

17 47 U.S.C. §230 (2006); Valerie C. Brannon, *Free Speech and the Regulation of Social Media Content*, CONG. RSCH. SERV. (Mar. 27, 2019), https://www.everycrsreport.com/files/20190327_R45650_9f272501744325782e5a706e2aa76781307abb64.pdf [https://perma.cc/VP6Z-FDQ7] [hereinafter *Free Speech*] (§230(c)(2) is known as the 'Good Samaritan' provision).

18 Social media companies are not immune from liability if they help to facilitate such problematic content. Valerie C. Brannon, *Liability for Content Hosts: An Overview of the Communication Decency Act's Section 230*, CONG. RSCH. SERV. (June 6, 2019), https://sgp.fas.org/crs/misc/LSB10306.pdf [https://perma.cc/9NA5-2Z2E] [hereinafter *Liability for Content Hosts*]. In 2018, President Trump signed into law the "Allow State and Victims to Fight Online Sex Trafficking Act of 2017" (FOSTA) that aims to bar sex trafficking online by making social media providers liable for keeping such posts up on their platforms. Tom Jackman, *Trump Signs 'FOSTA' Bill Targeting Online Sex Trafficking, Enables States and Victims to Pursue Websites*, WASH. POST (Apr. 11, 2018), https://www.washingtonpost.com/news/true-crime/wp/2018/04/11/trump-signs-fosta-bill-targeting-online-sex-trafficking-enables-states-and-victims-to-pursue-websites/ [https://perma.cc/SQA8-6G33]. Section 230(e) provides that immunity does not apply in certain types of lawsuits including federal criminal law, intellectual property laws, and the Electronic Communications Privacy Act of 1986. 47 U.S. Code §230 (2006).

the Internet."[19] Given that social media platforms are privately owned forums of speech, they are not constrained by the First Amendment the same way public state actors are. Thus, if they chose to regulate content, they can freely do so without violating any freedom of speech. However, the purpose of social media is to be an open platform for unregulated free speech, thus, in the past, companies such as Twitter and Facebook have been reluctant to censor posts. The Supreme Court has recognized social media as an important avenue for people to speak and listen to one another and form integral relationships. Justice Kennedy, in an opinion described social media as "perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard."[20]

The purpose of Section 230 is to incite providers to be "Good Samaritans" by removing offensive and harmful content while attempting to reduce over-censorship to prevent free speech encroachments.[21] However, the combination of Section 230 immunity and the lack of regulatory oversight due to free speech concerns has instead enabled these social media platforms to profit off of harmful and unregulated posts.

C. *Current System of Content Regulation*

There are many ways for platforms to regulate content, and they are often used in conjunction, relying on automated systems, human content moderation, and user flagging. The first is what Kate Klonick calls Ex Ante content moderation—it is the regulation of content "in the moment between upload and publication" that is automatically registered and removed through an automated algorithm.[22] this automatic process uses a system that, in seconds, filters and flags posts to be removed by applying the same set of rules to all content.[23] Ex Post proactive manual content

---

19 Klonick, *supra* note 14, at 1608.

20 137 S. Ct. 1730 (2017).

21 Klonick, *supra* note 14, at 1602.

22 *Id.* at 1636.

23 *See* James Grimmelmann, *The Virtues of Moderation*, 17 YALE J. OF L. & TECH. 42, 67 (2015); *see also How Automated Tools are Used in the Content Moderation Process*, NEW AM., https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/how-automated-tools-are-used-in-the-content-moderation-process/ [https://perma.cc/VF65-9WLT] (last visited Feb. 11, 2022) (for examples of different automated tools used in content moderation, such as PhotoDNA that reliably identifies child pornography through a picture-recognition system).

moderation[24] is the regulation of content done proactively by the company itself, but is currently limited to moderation of extremism.[25] Ex Post reactive manual content moderation[26] is the process where user-content is reviewed through flagging by other users that is then reviewed by trained moderators who make decisions based on a set of internal rules the company enforces.[27] This is the most common type of content regulation and from January to June 2020, Twitter suspended nearly one million accounts for rules violations,[28] through this type of content regulation.

Facebook, in particular has a tier system for their human content moderation—tier three moderators do the daily reviewing of posts in call centers outsourced all over the world, tier two moderators supervise tier three moderators and review more controversial content, and tier one moderators are based in Facebook's headquarters.[29] Tier three moderators can confirm or deny that the content they are reviewing violates their community standards or move it up to a tier two moderator to review.[30] If the moderator confirms the post to be an abuse, the post is automatically removed, and the user is given a message that says the post violated Facebook community standards and to review the standards if they want further information.[31] A user can be banned if their content is repeatedly flagged and removed.[32]

Social media companies have also branched out and begun to outsource their content moderation to third parties. In 2018, Facebook created an oversight board that would act as a quasi-judiciary[33] to review some of its high-profile decisions. Facebook

---

24 Klonick, *supra* note 14, at 1638.

25 *Id.*

26 *Id*.

27 *Id.* at 1632 (The "rules and standards are rooted in the social norms and values of a community."). *See generally* Klonick, *supra* note 14, at 1631 (for an analysis on how content moderation developed from standards to a more intricate set of rules).

28 Michael Luca, *Social Media Bans are Really, Actually, Shockingly Common*, WIRED (Jan. 20, 2021, 9:00 AM), https://www.wired.com/story/opinion-social-media-bans-are-really-actually-shockingly-common/ [https://perma.cc/AT5R-J3VP] (Twitter suspended 925,000 accounts in the first half of 2020).

29 Klonick, *supra* note 14, at 1639–40.

30 *Id.* at 1647.

31 *Id.*

32 *Id.*

33 Kate Klonick, *Inside the Making of Facebook's Supreme Court*, NEW YORKER (Feb. 12, 2021), https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court [https://perma.cc/UY8E-5NRH] [hereinafter *Making of Facebook's Supreme Court*] (Harvard Law School Professor, Noah Feldman proposed the creation of a quasi-judiciary on Facebook–also known as the 'Facebook Supreme Court').

elected a panel of twenty former political leaders, human rights activists, journalists, and scholars to act as judges, deliberating and deciding on what to take down and what to leave up. Facebook's intention for setting up the board was to get an outside third-party to review judgements Facebook content moderators make, and subsequently enforce binding decisions when a user requests an appeal.[34]

The boards powers have been criticized as inadequate, however. They can only rule on whether posts have been wrongly taken down, as a sort of appeals process, but not on posts that remain up, and in making these decisions they must follow Facebook's current rules, rather than utilize their own set of standards.[35] One member of the Real Oversight Board said, "This is a Facebook-funded, Facebook-appointed body that has no legitimacy to make real decisions. . .but, rather was designed to deflect attention from Facebook's. . .for-profit business model."[36] Criticism against the oversight board became rife when the oversight board was supposed to make a final decision on Trump's status on Facebook, but rather they deflected and said Facebook itself must decide whether to ban Trump permanently or set a time frame for the suspension.[37] The board noted that Facebook continues to make errors in valuing the "substance of people's messages, and not the context" by treating users with a few followers the same as users like Trump with millions of followers.[38] Despite the board recognizing the flaws of Facebook and its content moderation, the board lacks the power to change the system without more transparent and unarbitrary policies and accountability on the part of Facebook itself.[39]

---

34 Steven Levy, *Oversight Board to Facebook: We're Not Going to Do Your Dirty Work*, WIRED (May 5, 2021, 1:34 PM), https://www.wired.com/story/oversight-board-to-facebook-not-going-to-do-your-dirty-work/ [https://perma.cc/D54C-6W32].

35 Billy Perrigo, *Facebook's Oversight Board is Reviewing Its First Cases, Critics Say It Won't Solve the Platform's Biggest Problems,* TIME (Dec. 7, 2020, 5:39 AM), https://time.com/5918499/facebook-oversight-board-cases/ [https://perma.cc/MF7X-SL5E].

36 Kari Paul, *Facebook Ruling on Trump Renews Criticism of Oversight Board*, THE GUARDIAN (May 5, 2021, 11:39 AM), https://www.theguardian.com/technology/2021/may/05/facebook-oversight-board-donald-trump [https://perma.cc/BQ78-THLD] (the Real Oversight Board is a "group of activists formed as a critique of Facebook's oversight board.").

37 *Id.*

38 Shira Ovide, *The Limits of Facebook's 'Supreme Court'*, N.Y. TIMES (May 5, 2021), https://www.nytimes.com/2021/05/05/technology/facebook-oversight-board-trump.html [https://perma.cc/M8Q8-HPLL].

39 *Id*.

III. Discussion

Given that virtually all speech that is broadcasted on the internet stems from and is facilitated by private companies, they have extraordinary power to regulate free speech and exercise authority over "wrongdoers" who may not otherwise be reachable due to the power of the internet to hide or fake identities.[40] Yet, one of the reasons social media providers are so hesitant to regulate content on their platforms is because of the inherent difference from traditional media that platforms pride themselves on. Traditional news "is defined by limited bandwidth. . .in contrast, social media platforms offer essentially infinite bandwidth."[41]

However, the benefits of unlimited free speech come with consequences. In 2020, the New York Times podcast series "Rabbit Hole"[42] discussed the extremist effects YouTube has on its users. Critics have said that "YouTube has inadvertently created a dangerous on-ramp to extremism by combining two things: a business model that rewards provocative videos with exposure and advertising dollars and an algorithm that guides users down personalized paths meant to keep them glued to their screens."[43] In 2016, 90% of extremists were radicalized, at least in part, by social media.[44]

---

40 Ardia, *supra* note 15, at 378.

41 Dipayan Ghosh, *Are We Entering a New Era of Social Media Regulation?*, Harv. Bus. Rev. (Jan. 14, 2021), https://hbr.org/2021/01/are-we-entering-a-new-era-of-social-media-regulation [https://perma.cc/F4E4-RT34].

42 *Rabbit Hole*, N.Y. Times (May 28, 2020), https://www.nytimes.com/2020/04/22/podcasts/rabbit-hole-prologue.html [https://perma.cc/N86R-35NR]. See Klonick, *supra* note 14, at 1626 ("[T]he mission of Facebook — 'to make the world more open and connected' and found that it often aligned with larger American free speech and democratic values. These philosophies were balanced against competing principles of user safety, harm to users, public relations concerns for Facebook, and the revenue implications of certain content for advertisers.").

43 Kevin Roose, *The Making of a YouTube Radical*, N.Y. Times (June 8, 2019), https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html [https://perma.cc/URM4-MF4F].

44 Michael Jensen et al., *The Use of Social Media By United States Extremists*, The Nat'l Consortium for the Study of Terrorism and Responses to Terrorism, https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf [https://perma.cc/83N3-TAGW]; Lieven Pauwels et al., Explaining and Understanding the Role of Exposure to New Social Media on Violent Extremism. An Integrative Quantitative and Qualitative Approach, *Radimed* (2014), https://orfeo.belnet.be/bitstream/handle/internal/4197/synTA043_en.pdf?sequence=1 [https://perma.cc/P84F-WD76] (defines violent extremism as taking it one step further from radicalism by fully denouncing pluralism and using violent and oppressive methods to achieve their political goals) (this data was collected from 479 extremist's social media activities in the PIRUS dataset between 2005 and 2016).

The internet provides a safety net for extremists to hide behind their radical ideologies which only intensifies their hate and the consequences that follow it. As Rachel Hatzipanagos stated in the Washington Post, if you start speaking abhorrent and hateful speech in the middle of a grocery store, the likelihood you initiate a physical response from someone(s) is very likely but online, you can escalate your rage until it turns into physical violence without any threats to yourself.[45] Not only can extremists hide behind their posts, but they are also influenced and encouraged by others' content. Online radicalization to violence does not happen after viewing just one post, but, is a process that occurs gradually as users continue to immerse themselves in extremist content.[46] Thus, if you are consistently surrounded by hateful words, slurs, and ideas, it becomes the norm and "norms are powerful because they influence people's behavior."[47] This is what happened to Dylan Roof, the man who killed nine black parishioners in a South Carolina church in 2015. Federal prosecutors stated that Dylan became self-radicalized online by absorbing other violent white supremacist posts[48] teaching him to believe violent action against black people was the necessary step in achieving white supremacy.[49] Dylan is an example of how an individual can become gradually indoctrinated by an overwhelming consumption of extremist ideologies online.

Extremists themselves understand the power that social media plays in radicalization and, thus, exploit this easy access social media provides since they know their posts will not be taken down, to recruit, reform, and groom users who would normally be unreachable.[50] Prominent terrorist groups such as foreign jihadists have adapted their recruitment tactics by inspiring users online through "a steady infusion of propaganda videos and call-to-action messages circulating via social media platforms," such as blogs,

---

45 Rachel Hatzipanagos, *How Online Hate Turns Into Real-Life Violence*, WASH. POST (Nov. 30, 2018), https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/ [https://perma.cc/WCX3-4F7X].

46 *Online Radicalization to Violent Extremism*, INT'L ASS'N OF CHIEFS OF POLICE (2014), https://www.theiacp.org/sites/default/files/2018-07/RadicalizationtoViolentExtremismAwareness-Brief.pdf [https://perma.cc/4AJL-7XDN] [hereinafter Awareness Brief].

47 Hatzipanagos, *supra* note 45.

48 Mark Berman, *Prosecutors Say Dylan Roof 'Self-Radicalized' Online, Wrote Another Manifesto in Jail*, WASH. POST (Aug. 22, 2016), https://www.washingtonpost.com/news/post-nation/wp/2016/08/22/prosecutors-say-accused-charleston-church-gunman-self-radicalized-online/ [https://perma.cc/XCE5-ZCWR].

49 *Id.*

50 Awareness Brief, *supra* note 46.

Facebook, YouTube, and Twitter.[51] They understand that the key to instilling this mindset into users is by flooding them with an excess of violent extremist information.

Social media allows users not only get indoctrinated by the sheer volume of extremist posts they can consume, but also by influential figures who have a powerful impact on society. Jair Bolsonaro's spread of misinformation online is just one example on the power high-profile users can have on society when they are unregulated. Bolsonaro, President of Brazil, has continuously spread misinformation about COVID-19 across social media through his reoccurring live-steaming and tweeting as part of his campaign strategy to discredit his democratic opposers and keep Brazil's economy running.[52] He has encouraged people to not wear masks, linked COVID-19 to the flu, and made no plans to get the vaccine.[53] Twitter and YouTube have deleted a few of his posts but this campaign to spread misinformation has hindered efforts to minimize COVID-19's impact[54] and has led Brazil to have one of the highest death tolls from COVID-19 in the world.[55]

In situations like this, a single high-profile user can influence a large group of people with relative ease simply because of their popular status and credibility, while an extremist group requires an average user to consume a large amount of their content in order to become indoctrinated. Therefore, an average user needs to be exposed to fewer posts from a high-profile user to be manipulated, thus, having a method of banning high profile users appropriately can have a larger impact on the average user.

---

51 Joseph Kunkle, *Social Media and the Homegrown Terrorist Threat*, THE POLICE CHIEF (June 6, 2012), https://newspunch.com/wp-content/uploads/2014/12/Police-Chief-Magazine-View-Article.pdf [https://perma.cc/V9BC-5GFB].

52 Julie Ricard, *Using Misinformation as a Political Weapon: COVID-19 and Bolsonaro in Brazil*, HARV. KENNEDY SCH. MISINFO. REV. (Apr. 17, 2020), https://misinforeview.hks.harvard.edu/article/using-misinformation-as-a-political-weapon-covid-19-and-bolsonaro-in-brazil/ [https://perma.cc/57MS-GGZX].

53 Adam Satariano, *YouTube Pulls Videos by Bolsonaro for Spreading Misinformation on the Virus*, N.Y. TIMES (July 24, 2021), https://www.nytimes.com/2021/07/22/world/youtube-bolsonaro-covid.html [HTTPS://PERMA.CC/5AJP-XY6B].

54 Ricard, *supra* note 52; Satariano, *supra* note 53.

55 *Covid: Brazil Hits 500,000 Deaths Amid 'Critical' Situation*, BBC NEWS (June 29, 2021), https://www.bbc.com/news/world-latin-america-57541794 [https://perma.cc/BH45-6FJY].

A. *Incentives to Regulate Content*

Due to the nature of Section 230, content providers are "free to choose which values they want to protect—or to protect no values at all."[56] Given the alarming effects social media has on violent and non-violent extremism and misinformation, platforms became motivated to regulate content while simultaneously balancing principles of free speech and democracy. This is due to a "sense of corporate social responsibility, but also because their economic viability depends on meeting users' speech and community norms."[57] When it comes to social accountability, social media companies felt it was their responsibility to combat extremism and misinformation without silencing free speech—a staple of what their platforms are supposed to provide.[58] Thus, by initially implementing tools and policies for users to filter out and hide violent and false content, platforms could balance free speech by leaving the content up while also satisfying safety concerns.[59] The primary reason, however, for companies to regulate obscene and violent content is to keep users on their sites to increase revenue. When users feel uncomfortable by content and leave the site, the company loses that revenue.[60] However, there is a balance to be struck. If providers take down too much content, they risk losing the users' trust and opportunity for interaction.[61] User posting, commenting, liking, sharing, etc., are how companies like Facebook and Twitter make their money.[62] These initial motivations to regulate content stemmed from the theory that a little would go a long way. However, as societal and government pressure intensified for social media providers to do more to combat the spread of misinformation, hate speech and violent content, companies began facing the dilemma of whether to choose social responsibility over profit. Facebook has chosen profit over safety according to Facebook whistleblower, Frances Haugen.[63]

---

56 Klonick, *supra* note 14, at 1617.

57 *Id.* at 1625.

58 *Id.*

59 *Id*. at 1625–26.

60 *Id*. at 1627.

61 *Id.*

62 Leslie K. John et al., *What's the Value of a Like?*, HAR. BUS. REV. (Apr. 2017), https://hbr.org/2017/03/whats-the-value-of-a-like [https://perma.cc/TAP8-H6SV].

63 David Bauder & Michael Liedtke, *Whistleblower: Facebook Chose Profit Over Public Safety*, AP NEWS (Oct. 4, 2021), https://apnews.com/article/facebook-whistleblower-frances-haugen-4a3640440769d9a241c47670facac213 [https://perma.cc/4FYB-PEJJ].

In 2017, insiders in Facebook became panicked that users would stop using the app altogether after noticing that key measures of engagement[64] were falling. So, in 2018, Facebook decided to modify their algorithm to prioritize meaningful social interactions (MSI's),[65] which in essence would encourage people to interact more with their friends and family, rather than randomly scrolling. Publicly, Facebook Chief Executive Officer (CEO) and co-founder, Mark Zuckerberg, said this change was to improve user's mental health,[66] but after Facebook researchers told Zuckerberg that the MSI's were promoting misinformation and hate speech and proposed ways to change it, Zuckerberg shut them down.[67] His reasoning was that these divisive posts were creating more sensation and driving up user traffic.[68] Facebook recognized that they were making more money by keeping people's attention on its own platform by showing harmful and problematic content—choosing profit over people. After a broad push by the public in the wake of the January 6th riots and Haugen's whistleblower statements exposing Facebook for spreading misinformation and hate speech, Zuckerberg pledged to start reducing the amount of political misinformation circulating and putting emphasis on content that gets attention,[69] along with a greater push on content regulation through their "Supreme Court."[70] Facebook's decision to pick profit over safety has in turn made them lose profit as Facebook's image and trust from users began chipping away.[71] As more light is shed on the harm that "Zuckerbergs" across the world can cause, social media providers need to, now more than ever, turn to more efficient content regulations to keep users happy and subsequently keep their profit up.

---

64 Key measures of engagement are defined as likes, comments, and shares.

65 Ryan Mac, *Engagement Ranking Boost, M.S.I., and More.*, N.Y. TIMES (Oct. 5, 2021), https://www.nytimes.com/2021/10/05/technology/engagement-ranking-boost-msi-facebook.html [https://perma.cc/9LDZ-HBYZ].

66 WSJ Tech News Briefing, *Zuckerberg Resisted Fixes for Facebook's Divisive Algorithm*, WALL ST. J. (Sept. 16, 2021), https://www.wsj.com/podcasts.apple.com/us/podcast/wsj-/technews-briefing/id74844126?i=1000535503809 zuckerberg-resisted-fixes-for-facebook-divisive-algorithm/4526aa57-21ae-4f2b-91de-cfef7f143301[https://perma.cc/GB8A-6BRP].6DTM-9GM6].

67 *Id.*

68 *Id.*

69 *Id.*

70 Klonick, *supra* note 33.

71 Mike Isaac et al., *After Whistle-Blower Goes Public, Facebook Tries Calming Employees*, N.Y. TIMES (Nov. 2, 2021), https://www.nytimes.com/2021/10/10/technology/facebook-whistleblower-employees.html [https://perma.cc/4PDY-4UVT].

B. *Is the Current System of Content Regulation Enough?*

1. Issues with the Automated System of Content Regulation

One reason platform struggle with content moderation is because it is in fact difficult to moderate. There are over 2.85 billion monthly active users on Facebook alone,[72] and over a three-month span in 2020, Facebook disabled over 1.3 billion accounts.[73] There are just too many posts and accounts to effectively regulate without creating an automated system. With an increase in misinformation being spread online because of COVID-19, upticks in violence due to the circulation of hate speech, and employed content moderators not being able to do their job from home during the pandemic,[74] social media companies have moved away from human content moderation and toward automated tools. The issue has become that these algorithms cannot handle complicated, personal, and context dependent[75] posts which leads to an over-censorship of accounts with more false positives and false negatives. Facebook and Twitter admitted that as they move toward a more automated system, they expect more mistakes.[76] Further, with increased pressure by the government to eliminate Section 230 and pressure from the public to remove content, platforms tend to remove all speech that has any inkling of violating their guidelines.[77]

The predominant issue with algorithms is their inability to comprehend context. They cannot identify and remove hate speech without racial bias. In a study done by the Allen Institute for Artificial Intelligence, researchers found that when a tweet is written by an African American it is 150% more likely to be flagged as offensive or hateful by an AI algorithm that detects hate speech.[78] For example, slurs such as the "n-word" or "queer" to an

---

72 *Leading Countries Based on Facebook Audience Size as of July 2021 (in millions)*, STATISTA RSCH. DEP'T (Sept. 10, 2021), https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ [https://perma.cc/TZ4J-2Z39].

73 Melissa Holzberg, *Facebook Banned 1.3 Billion Accounts Over Three Months to Combat 'Fake' and 'Harmful' Content*, FORBES (Mar. 22, 2021, 10:05 AM), https://www.forbes.com/sites/melissaholzberg/2021/03/22/facebook-banned-13-billion-accounts-over-three-months-to-combat-fake-and-harmful-content/?sh=6530217f5215 [https://perma.cc/6C7T-D5GG].

74 Evelyn Douek, *Covid-19 and Social Media Content Moderation*, LAWFARE (Mar. 25, 2021, 1:10 PM), https://www.lawfareblog.com/covid-19-and-social-media-content-moderation [https://perma.cc/EVC9-BGAB].E6LP-8CWX].

75 Brown, *supra* note 12, at 456.

76 Douek, *supra* note 74.

77 Brown, *supra* note 12, at 475.

78 Shirin Ghaffary, *The Algorithms That Detect Hate Speech Online are Biased Against Black People*, VOX (Aug. 15, 2019, 11:00 AM), https://www.vox.com/recode/2019/8/15/20806384/social-

algorithm are offensive, but in the setting in which it is said, that is not always the case. In this scenario, the algorithm is creating too many false positives, but the algorithm also has issues with too many false negatives, particularly with failing to intercept and remove violent content. For instance, the 2019 Christchurch massacre was streamed across Facebook for seventeen minutes before it was taken down.[79] The algorithm allowed the video to stay up, and it was only removed due to abundant user complaints.[80]

As the Coronavirus pandemic settled in, social media companies sent their workers home, including content moderators. These providers turned to AI to monitor online posts with the hope that this could be the new future. However, they quickly learned that algorithms might struggle to differentiate between illicit and licit posts. "While far more content was flagged and removed for allegedly breaking the companies' rules on what could be posted online, in some areas dangerous and possibly illegal material was more likely to slip past the machines."[81] Social media sites nearly doubled in removable content, but it was predominately attributed to false positives that was in hindsight damaging to remove,[82] while actual harmful content removal fell by 40% in the second half of 2020 "because of a lack of humans to make the tough calls about what broke the platform's rules."[83]

2. Issues with Human Content Moderators

Given that algorithms are far from perfect, society has encouraged further human content moderation to avoid false positives and negatives.[84] Even as automated systems grow and develop, the need for human effort to handle difficult situations is

---

media-hate-speech-bias-black-african-american-facebook-twitter [https://perma.cc/WDU7-8T74]; Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, UNIV. OF WASH., https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf [https://perma.cc/EH2S-M4RE] (last visited Sept. 15, 2021).

79 Brown, *supra* note 12, at 478.

80 *Id.*

81 Mark Scott & Laura Kayali, *What Happened When Humans Stopped Managing Social Media Content*, POLITICO (Oct. 21, 2020), https://www.politico.eu/article/facebook-content-moderation-automation/ [https://perma.cc/SV4X-RZHY].

82 *Id.* ("In Syria, where campaigners and journalists rely on social media to document potential war crimes, scores of activists' accounts were closed down overnight—often with no right to appeal those decisions. Other content, including news articles and health information linked to the coronavirus, was similarly scrubbed from the internet as the machines got to work.").

83 *Id.*

84 Minna Ruckenstein, et al., *Re-humanizing the Platform: Content Moderators and the Logic of Care*, 22 CTR. FOR CONSUMER SOC'Y RSCH. AND HELSINKI CTR. FOR DIGITAL HUMANITIES 1026, 1027 (2020), https://journals.sagepub.com/doi/pdf/10.1177/1461444819875990 [https://

still present.[85] However, despite humans being in a better position to recognize context than an AI machine, human content regulation inevitably comes with implicit bias for three main reasons—(1) there is a lack of clear guidelines to rely on that leads to subjective bias, (2) quick turnaround for decisions on whether to remove a post or not, and (3) human content moderators lack cultural knowledge to be able to discern the meaning behind a post.

Human moderators still rely on a set of community guidelines, which not only are outdated and vague, but allow the moderator to apply the standards inconsistently through broad discretion, which encourages subjective bias.[86] Regardless, no matter how up-to-date the standards are, there is no safety feature that ensures moderators apply the standards consistently. Each moderator appears to find different genres of content offensive and worthy of being removed, depending on their own worldview. One can be surprised by videos of animal abuse, and another can be traumatized by racists posts relating to the KKK.[87] Based on their own subjective views of what they find harmful or offensive, they will implement the community guidelines accordingly, which leads to inconsistent results. Despite some platforms having a multi-layered system of checks and reviews to ensure the accuracy of moderation decisions,[88] this system does not address the issue that

---

perma.cc/CG6M-CEGE] ("Human moderators are involved in designing and implementing moderation software, also training machines and making decisions about online content.").

85 *See generally* Ruckenstein, *supra* note 84 (there is a current move to rehumanize content moderators by cultivating discussions and open communication about the work content moderators do and the future of online culture, rather than training moderators to become versions of algorithms.). *See also* Miriah Steiger et al., *The Psychological Well-Being of Content Moderators*, ACM DIGIT. LIBR. (May 13, 2021), https://crowd.cs.vt.edu/wp-content/uploads/2021/02/CHI21_final__The_Psychological_Well_Being_of_Content_Moderators-2.pdf [https://perma.cc/2Q43-PXSW] (discussion about the effects of violent and hateful content moderators sift through daily and the impact that has on their mental health).

86 Brown, *supra* note 12, at 479; Daisy Soderberg-Rivkin, *When it Comes to Content Moderation, We've Been Focusing on the Wrong Type of Bias*, MORNING CONSULT (Dec. 5, 2019, 5:00 AM), https://morningconsult.com/opinions/when-it-comes-to-content-moderation-weve-been-focusing-on-the-wrong-type-of-bias/ [https://perma.cc/PR4D-5AKF]. *See also* Ruckenstein, *supra* note 84 (sharing detailed information about how content moderators make decisions and their work conditions are silenced, because knowledge of this information can harm the social media providers) ("[T]hese moderators keep a low profile not only because of the non-disclosure agreements (NDAs) they have signed but also because they face threats both online and office.").

87 Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads'*, GAWKER (Feb. 16, 2012, 3:45 PM), https://www.gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads [https://perma.cc/B8X4-C5T8].

88 *See* discussion *supra* Section II.C.

causes bias in the first place—ambiguous rules and the ability for bias to seep into moderation no matter what tier the moderator is.

The key to reducing this bias comes down to training on how to uniformly apply the provider's rules.[89] Despite the trainings and the guidelines to base decisions on, "cultural biases still crept into moderation, especially when judging subjective content."[90] The challenges companies face include not just the speed[91] at which moderators must sift through posts, but also the need to train thousands of low-paid moderators to apply a single set of rules despite the daily changes to such rules.[92] Additionally, issues arise such as: "a lack of cultural or political context on the part of the moderators; missing context in posts that makes their meaning ambiguous; and frequent disagreements among moderators about whether the rules should apply in individual cases."[93] Even after being trained, content moderators attempt to apply the policies set out by the social media companies even though it does not make sense to them and are often pressured to make quick decisions due to the high quota placed on them.[94] In a matter of seconds, they must consider the context of the post and the user's profile against the publicly posted community guidelines and the internal guidelines. Moreover, when the community standards they are trained to apply are not directly applicable, moderators must invent policies on the spot.[95] Often these split-second decisions are based off instinct, which is inherently linked to bias.[96] Additionally, moderators must also consider unexpected policy changes that often come after breaking news events. One former moderator said, "during

---

89 Klonick, *supra* note 14, at 1642.

90 *Id.* ("Content moderators act in a capacity very similar to that of judges: (1) like judges, a judge: moderators are trained to exercise professional judgment concerning the application of a platform's internal rules; and (2) in applying these rules, moderators are expected to use legal concepts like relevancy, reason through example and analogy, and apply multifactor tests.").

91 Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/TVZ2-RG7Y] (social media providers have turned these humans into "human processors." Facebook expects their content moderators to sift through hundreds of reports per hour and are trained to work in these fast-paced environments making decision on whether to keep a post-up, remove it, or send it to a higher authority within seconds after enduring long hours with little pay).

92 *Id.*

93 *Id.*

94 *Id.*

95 *Id.*

96 Stephanie Vozza, *5 Common Unconscious Biases That Lead to Bad Decisions*, FAST COMPANY (April 16, 2015), https://www.fastcompany.com/3045035/5-common-unconscious-biases-that-lead-to-bad-decisions [https://perma.cc/C3K3-C5VX].

times of national tragedy . . . managers would tell moderators to remove a video—and then, in a separate post a few hours later, to leave it up."[97]

Even though content moderators are supposed to understand context better than an AI machine, that often falls short of reality. Social media providers like Facebook rely heavily on contract labor to promote efficiency and cut costs.[98] Although some content moderators are based in the United States, a large portion are outsourced in the Philippines, Mexico, and other foreign countries.[99] The consequence of outsourcing these jobs is the disconnect between the moderator and the culture and language of the content they're moderating.[100] Often, moderators are asked to make decisions about posts that come from an entirely different place in the world with a different set of cultural norms and political views, and require a context that the moderator may not possess.[101]

When forming the Facebook Supreme Court, Facebook set up global workshops where they invited individuals to come, look at a scenario where a post was taken down, and decide whether that was the proper decision or not. The guests were asked, "Is this hate speech? What does that mean? And should that be up on Facebook or not?"[102] and after much debate, the room did eventually come to a consensus. However, these were Americans in that room, but when the same test was done in Berlin or Singapore, individuals came up with very different results.[103] As Simon Adler said, "when you talk to people from different parts of the world . . . there's not universal agreement on this."[104] Berhan Taye added, "content moderation is a very difficult task, one that's being done by people that have no freaking idea about our way of life . . . . ."[105] After these workshops, Facebook landed on the solution to have forty members that are representing every continent and from a

---

97 Newton, *supra* note 91.

98 *Id.*

99 Isaac Chotiner, *The Underworld of Online Content Moderation*, NEW YORKER (July 5, 2019), https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation [https://perma.cc/6G29-L43S].

100 *Id.*

101 Newton, *supra* note 91.

102 Simon Adler, *Facebook's Supreme Court*, RADIOLAB (Feb. 12, 2021), https://www.wnycstudios.org/podcasts/radiolab/articles/facebooks-supreme-court [https://perma.cc/V9MS-RZE9].

103 *Id.*

104 *Id*.

105 *Id*.

wide array of backgrounds, including gender, political views, and occupations to minimize the bias of human content moderation.[106]

3. Issues with Transparency[107] and Due Process

Content regulation becomes murky as details are often hidden from the public[108] and community standards that are available to the public are actually not the same rules moderators use when regulating content.[109] Content moderators apply their own professional judgement in deciding whether a post or account should be removed regarding a violation of these internal rules and details are hidden from the user as to why they are being banned.[110] Whenever Facebook removes an account/post, it refers the user to its policy guidelines which are intentionally vague because if a user knew "what criteria was being used to judge their content, they could hold Facebook to them. It would be clear what Facebook was choosing to censor . . . ."[111]

Additionally, users who wish to speak to human content moderators to understand why their post/account got taken down and appeal the decision, are left with few answers because the content moderators themselves cannot provide rationale responses. The moderators often give cookie-cutter answers taken from the guidelines, but who knows why they actually decided to remove a post or ban an account. In the words of one user:

> After I appealed I received an email from someone called Ron at Facebook's Pages Support section saying, "I'm here to help." I emailed Ron explaining that I didn't understand why the page had been unpublished and I asked him to say which post contained (as they claimed) "malicious or misleading content." I offered to comply with Facebook's wishes and delete any post they thought was a problem. His reply didn't provide any detail at all, it simply said "We have a no-tolerance policy concerning

---

106 *Id*.

107 Shagun Jhaver, *Identifying Opportunities to Improve Content Moderation* (May 18, 2020), (Ph.D. dissertation, Ga. Inst. Of Tech.) (on file with the Georgia Tech Library) ("Cornelia Moser defines transparency as opening up "'the working procedures not immediately visible to those not directly involved in order to demonstrate the good working of an institution.").

108 Klonick, *supra* note 14, at 1639. *See also* Catherine Buni & Soraya Chemaly, *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, THE VERGE (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech [https://perma.cc/PDM3-P6YH].

109 Klonick, *supra* note 14, at 1639.

110 Buni & Chemaly, *supra* note 108.

111 Chen, *supra* note 87.

> this infraction and your page is ineligible to be republished." I still don't know which post triggered the censorship.[112]

Many users who seek answers for why this happened to them do so to avoid the experience in the future, but are often left confused and answerless.[113] Not only is there a lack of transparency for why users are being banned or their content removed, but content providers are also masking which users are exempt from content regulation and why they are exempt. Although Mark Zuckerberg has publicly stated that Facebook views their nearly three billion users equally and that the "standards of behavior apply to everyone,"[114] company documents reveal that is far from the truth. A program known as 'X-Check' has given high-profile figures, such as politicians and celebrities, privileges that other users do not receive.[115] What was initially supposed to be a further level of content control for high-profile accounts, is now used to protect these users from the company's current regulation protocols. These users are considered "whitelisted"—immune from content regulation[116]—which often leads to posts containing "harassment or incitement to violence" which would normally be removed by the platform, to stay posted for longer. This allows a certain group of people to violate community standards without any consequences, leading to under-banning. Among the VIP's was soccer star Neymar, who in 2019, posted a nude photo of a woman who accused him of rape.[117] This content which would normally be taken down, was left up by X-Check because X-Check blocked Facebook moderators from attempting to take it down.[118] As the recent Facebook whistleblower, Frances Haugen, said, "the company in-

---

112 Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 NEW MEDIA AND SOC'Y 4366, 4377 (2018), https://journals.sagepub.com/doi/10.1177/1461444818773059 [https://perma.cc/FE8L-ZEFZ].

113 *Id*. at 4377–78.

114 Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt*, WALL ST. J. (Sept. 13, 2021, 10:21 AM), https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=DJemalertNEWS [https://perma.cc/6JE8-628C].

115 *Id.*

116 *Id.*

117 Salvador Rodriguez, *Facebook Shields Millions of VIP Users From Standard Moderation Protocols, per Report*, CNBC (Sept. 13, 2021, 5:35 PM), https://www.cnbc.com/2021/09/13/facebook-shields-millions-of-vip-users-from-moderation-protocols.html [https://perma.cc/6NU8-Q99V].

118 *Id.*

tentionally hides vital information from the public, from the U.S. government, and from governments around the world."[119]

This lack of transparency by Facebook and other content providers leads to a lack of due process.[120] Although social media providers offer an appeals process to discuss whether their post/account were taken down in error, they often run into problems.[121] One user said that when she tried to appeal, she got no response and it was just "speaking into a void", and another said that her response to the appeal was a reiteration of the earlier decision.[122] Other users say providers discuss the opportunity to appeal but provide no avenue to do so.[123] Despite the goal of balancing free speech with reduced violence, misinformation, and hate speech, the current system of content regulation is doing the opposite. By over-blocking users without explanation, and under-blocking high-profile users by letting them slip past the current system, not enough justice is imposed on criminals who are spreading violence rapidly, and those who just want a platform for which to share ideas are being removed without cause or explanation.[124] There is no clear balance of fundamental rights and without a "true day in court,"[125] users are left without essential due process.

---

119 Bobby Allyn, *Here are 4 Key Points From the Facebook Whistleblower's Testimony on Capitol Hill*, NPR (Oct. 5, 2021, 9:30 PM), https://www.npr.org/2021/10/05/1043377310/facebook-whistleblower-frances-haugen-congress [https://perma.cc/5TRJ-JPP2] (Haugen added, "During my time at Facebook, I came to realize a devastating truth: almost no one outside of Facebook knows what happens inside Facebook.").

120 *See generally* Klonick, *supra* note 14, at 1665-–66 ("The internet has been a force for free speech and democratic participation since its inception. The internet has also made speech less expensive, more accessible, more generative, and more interactive than it had arguably ever been before. . .[b]ut the lack of an appeals system for individual users and the open acknowledgment of different treatment and rule sets for powerful users over others reveal that a fair opportunity to participate is not currently a prioritized part of platform moderation systems.").

121 West, *supra* note 112, at 43794378.

122 *Id.*

123 *Id.*

124 Frederick Mostert & Alex Urbelis, *Your Day In Court: Social Media Needs a System of Due Process*, Fin. Times (May 16, 2021), https://www.ft.com/content/48c49453-9a8f-4125-85d7-94220497d13c [https://perma.cc/JCE6-ESYU].

125 This is a metaphorical "true day in court" since there is no day in court given to users in private action.

C. *Proposals for Digital Governance*

1. Congressional Intervention

Content regulation, amidst all the ideals that are partisan in America, has seemingly been an issue criticized on both sides of the political arena with greater push for significant changes to it. Lawmakers have discussed Section 230[126] reform and potential oversight that could come from the FTC or FCC.[127] Conservatives in Washington are advocating for reform of Section 230 that would "constrain platform editorial discretion to create a more favorable climate for their political perspective" and progressives want a platform that will be "less hostile toward speech from marginalized groups struggling for social, economic, and racial justice."[128] Reform of Section 230 is on Biden's current agenda due to the sense of urgency from the Capital riots that took place in early 2021 and the spread of misinformation regarding the 2020 election. Biden has been quoted stating that Section 230 should be revoked immediately.[129] Additionally, after Haugen, the Facebook whistleblower, testified in front of Congress, she called on lawmakers to impose regulations on Facebook, stating that "tweaks to outdated privacy protection or changes to Section 230 will not be sufficient."[130] The increased pressure from Congress has also become a tool of coercion to force social media platforms into reforming themselves. Mark Zuckerberg even told Congress that it "may make [sic] more sense for there to be liability for some of the content" and that Facebook "would benefit from clearer guidance from elected officials."[131]

Some ideas that have circulated amongst legal scholars include adding a 'duty of care' element to Section 230, which would impose an affirmative obligation on social media providers to prevent one

---

126 *See* discussion *supra* Section II.A.

127 Brown, *supra* note 12, at 485.

128 Mark MacCarthy, *Back to the Future for Section 230 Reform*, BROOKINGS (Mar. 17, 2021), https://www.brookings.edu/blog/techtank/2021/03/17/back-to-the-future-for-section-230-reform/ [https://perma.cc/Z9NV-G8UE].

129 Michael D. Smith & Marshall Van Alstyne, *It's Time to Update Section 230*, HARV. BUS. REV. (Aug. 12, 2021), https://hbr.org/2021/08/its-time-to-update-section-230 [https://perma.cc/C33N-E5YU].

130 Lauren Feiner, *Facebook Whistleblower: The Company Knows It's Harming People and the Buck Stops with Zuckerberg*, CNBC (Oct. 5, 2021, 6:32 PM), https://www.cnbc.com/2021/10/05/facebook-whistleblower-testifies-before-senate-committee.html [https://perma.cc/R9E2-3A3W].

131 Smith & Van Alstyne, *supra* note 129.

party from using the platform to harm another party.[132] This would make social media providers liable in the event they create or endorse an unsafe environment through their services.[133] There have also been a few bills introduced into legislation,[134] that seek to amend Section 230 to make social media providers more liable for users' actions.[135]

2. Judicial Intervention

If policymakers are unable to pass any legislation, the courts have indicated that they may step in, in lieu of Congress. In a concurring opinion on the dismissal of a case alleging that President Trump violated the First Amendment by blocking Twitter users,[136] Justice Clarence Thomas criticized Section 230. He indicated that social media companies have "enormous control over speech"[137] and that the court should consider, in an appropriate case, the

---

132 *Id.*

133 *Id.*

134 Meghan Anand et al., *All the Ways Congress Wants to Change Section 230*, SLATE (Mar. 23, 2021, 5:45 AM), https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html [https://perma.cc/7KVD-75QW] (see generally for a list of bills introduced).

135 Democratic Representative Anna Eshoo's bill will hold large social media platforms liable for algorithmic promotion of extremism. *See generally* Anna Eshoo, *Reps. Eshoo and Malinwoski Introduce Bill to Hold Tech Platforms Liable for Algorithmic Promotion of Extremism*, CONGRESSWOMAN ANNA G. ESHOO (Oct. 20, 2020), https://eshoo.house.gov/media/press-releases/reps-eshoo-and-malinowski-introduce-bill-hold-tech-platforms-liable-algorithmic [https://perma.cc/T8AD-SWK9]. The Safe Tech Act introduced by Democratic Senators Mark Warner, Mazie Hirono and Amy Klobuchar which would not hold platforms liable, but rather give individuals a chance to seek redress for harm caused by removing Section 230's legal liability bar. MacCarthy, *supra* note 125. The PACT Act includes procedures that motivate platform providers to remove harmful content and measures that make the moderation systems more accountable to users. The Act would require social media platforms to reveal their policies and practices for content regulation, release consistent statistical reports of content regulated, and explain to users their moderation decision within fourteen days with the ability to appeal such decisions. Additionally, the Bill would reform Section 230 by opening social media companies to civil lawsuits from federal regulators. Makena Kelly, *The PACT Act Would Force Platforms to Disclose Shadowbans and Demonetizations*, THE VERGE (June 24, 2020, 3:36 PM), https://www.theverge.com/2020/6/24/21302170/facebook-google-brian-schatz-john-thune-section-230-content-moderation [https://perma.cc/P9X8-FCGY].

136 Biden v. Knight First Amendment Institute At Columbia Univ., Et Al., 593 U.S. (2021); Rick Rouan, *Fact Check: Justice Clarence Thomas Didn't Say Section 230 Is Unconstitutional*, USA TODAY (Apr. 8, 2021, 8:51 PM), https://www.usatoday.com/story/news/factcheck/2021/04/08/fact-check-post-misrepresents-justice-thomas-section-230/7122886002/ [https://perma.cc/N2RR-N6ZM].

137 Rouan, *supra* note 136.

scope of immunity that Section 230 provides.[138] His concurrence makes clear that the court is deliberating whether the scope of Section 230 should be narrowed, giving social media providers less authority to regulate content online and more power to the government to regulate free speech. In June of 2021, Texas Supreme Court ruled that Section 230 does not protect Facebook from any form of sex-trafficking recruitment that takes place on their platform.[139] The court relied on a duty of care standard[140] that courts and Congress are moving towards implementing.[141]

3. Problems with Congressional and Judicial Intervention

Although government regulation would create more responsibility that is absent from self-regulation currently, this approach would implicate many constitutional limitations, particularly the First Amendment. By either repealing Section 230 or altering it in a way that makes providers liable for users' content, these social media providers lose their central feature—ability to regulate content within their own discretion. Conversely, with government regulation, the government must regulate in accordance with the First Amendment and judicial rulings since the party regulating such content is a public entity. Thus, the government's power to remove content and ban users is exceedingly limited.[142] Most of the proposed bills[143] aimed to alter Section 230 would substantially modify the aspects of social media people love most. Additionally, piecemeal reform of Section 230 is "inevitably underinclusive" because many of these proposed bills seek to amend aspects of Section 230 by providing exceptions whenever a new harmful concept emerges rather than altering Section 230 entirely.[144]

---

138 Wiley Rein LLP, *Justice Thomas Lays Blueprint for Supreme Court to Limit Section 230 in a Future Case*, JDSUPRA (Oct. 15, 2020), https://www.jdsupra.com/legalnews/justice-thomas-lays-blueprint-for-67566/ [https://perma.cc/V4YU-WFLX].

139 In re Facebook, Inc. And Facebook, Inc. D/B/A Instagram, Relators, No. 20-0434625 S.W.3d 80 (2021); Smith & Van Alstyne, *supra* note 126.

140 *See* discussion *supra* Section III.C.1.

141 Smith & Van Alstyne, *supra* note 129.

142 There are nine types of speech not protected by the First Amendment that the government would be able to regulate on social media—obscenity, fighting words, defamation (libel and slander), child pornography, perjury, blackmail, incitement to imminent lawless action, true threats, and solicitations to commit crimes. *Which Types of Speech are Not Protected by the First Amendment?*, FREEDOM F. INST., https://www.freedomforuminstitute.org/about/faq/which-types-of-speech-are-not-protected-by-the-first-amendment/ [https://perma.cc/GG4H-P3CK].

143 *See* discussion *supra* Section III.C.1.

144 MacCarthy, *supra* note 128.

One of the greatest concerns for government regulation is that it could be abused for political exploitation—a partisan government can tilt online discussion to favor its own point of view.[145] A great example is former-President Trump's 2020 executive order aimed at limiting some of Section 230 protections including allowing federal authorities to hold social media companies liable for infringing on user's free speech by removing or modifying the users' posts.[146] This executive order came just days after Twitter added a fact-check label to Trump's tweets about mail-in voting.[147] "The point was to send a clear message to social platforms: any effort to limit or frame the President's speech, even when false and potentially damaging, will be met with aggressive legal action."[148]

Additionally, the government does not have the technological ability to detect and remove posts and accounts at an equivalent or better scale and speed than current platforms are using. The process for creating new policies of content regulation would take longer to institute by a government entity like the FCC or FTC than it would for providers like Facebook and Twitter to implement.[149] Even if the government could regulate content at an equal or better level, similar problems relating to bias, and transparency will continue to exist. The only way to circumvent these issues is to completely alter how social media is used and that would not only implicate sizeable constitutional concerns, but also social media would look more like what traditional publishing is, rather than a platform with infinite bandwidth.

IV. PROPOSAL

A. *Preface*

This proposed arbitration system will not be the final solution to the current issues of content regulation, nor will it replace the

---

145 *Id.*

146 Kim Lyons, *Biden Revokes Trump Executive Order That Targeted Section 230*, VERGE (May 15, 2021), https://www.theverge.com/2021/5/15/22437627/biden-revokes-trump-executive-order-section-230-twitter-facebook-google [https://perma.cc/6S6Y-8YAJ].

147 Brown, *supra* note 12, at 487.

148 *Id.*

149 Drew Desilver, *Congress is Off to a Slow Start in 2021, Much as it Has Been In Previous Years*, PEW RSCH. CTR. (Aug. 13, 2021), https://www.pewresearch.org/fact-tank/2021/08/13/congress-is-off-to-a-slow-start-in-2021-much-as-it-has-been-in-previous-years/ [https://perma.cc/M724-AXWE] (for more information about why the 117th Congress has been slow to enact law).

current system. Rather, this proposal will be used in conjunction with the current system as a third and final step to content regulation to improve the current system and will be used for a subsection of users—high-profile users. High-profile users have both the ability to negatively influence society with minimal effort,[150] while simultaneously being able to scrape by the current system of moderation due to their popularity and the desire for social media companies to keep them active for profit.[151] Thus, social media companies need an arbitration system to appropriately ban these high-profile users; the system will minimize over-banning by keeping a check on users gaming the system,[152] and minimize under-banning by ensuring that companies like Facebook are not blocking the moderation system for these users.[153]

Arbitration will not eliminate content moderators or automated AI systems because there are not enough resources to moderate content without these imperfect systems, and algorithms and content moderators are beneficial in certain circumstances. For example, algorithms are trained to successfully moderate certain content such as child pornography and copyright infringement content with "little evidence of implicit bias."[154] Content moderators can also take down billions of posts that do clearly fit into a category of guidelines/rules. This new arbitration system would be used as the final and third step in the process—the first step would be removing posts through an AI generated system, the second step is having posts and accounts reviewed by content moderators, and then the third step requires that high-profile users, who are flagged to be part of this arbitration system, be reviewed by a third-party neutral tri-panel of arbitrators to determine whether they should in fact be removed.[155]

This proposal will not be able to remove all extremism or misinformation on the internet. There will still be problems with bias, context, and transparency, but the hope is that this system can reduce some of those issues by targeting users who have a large impact on society. Moreover, this proposal can hopefully inspire social media companies to see the benefits of a better method for content regulation and improve their current system.

---

150 *See* discussion *supra* Section III.

151 *See* discussion *supra* Sections III.A, III.B.3.

152 *See* discussion *infra* Sections IV.C, IV.F.

153 *See* discussion *infra* Sections IV.B.

154 Soderberg-Rivkin, *supra* note 86.

155 *See* discussion *infra* Sections IV.B.

B. *Proposed Arbitration System*

Arbitration is a "private process where disputing parties agree that one or more individuals can make a decision about the dispute after receiving evidence and hearing arguments."[156] It is, in essence, a private trial that will be paid for by the social media company.[157] In my proposed system, the social media company, as the disputing party, and the user will contract to send the dispute to arbitration through an arbitration clause that the user must sign when creating an account.[158] The basic overview for my proposed arbitration process is to create a set rule—that X number of posts need to get flagged by Y number of users before the user gets flagged for arbitration. Additionally, the user must meet a certain threshold of followers which would correlate to the influence that the user's account has on the public. Once a user is flagged, my automatic fast-track arbitration process will begin.[159] The social media company cannot override this computerized system that automatically places a user who reaches the required number of flags and followers into arbitration; thus, the issue of providers underbanning high-profile users, as seen in the X-Check system Facebook used, will be eliminated.[160]

All social media companies that are part of this system will contribute a percentage of their yearly revenue to a new third-party Social Media Dispute Resolution Center that is created as

---

156 *Dispute Resolution Process: Arbitration*, A.B.A., https://www.americanbar.org/groups/dispute_resolution/resources/disputeresolutionprocesses/arbitration/ [https://perma.cc/Z57H-25X7] (last visited Sept. 15, 2021).

157 Stephen J. Ware, *Is Adjudication a Public Good?: 'Overcrowded Courts' and the Private-Sector Alternative of Arbitration,* 14 CARDOZO J. CONFLICT RESOL. 899, 905–06 (2013) ("A downside of arbitration for the disputing parties is that they have to pay for it. While litigation receives a sizable government subsidy, arbitration does not. . .parties to arbitration must pay the arbitrator's fee, as well as the administrative costs of the arbitration organization, and any cost of the hearing room."); *See* discussion *infra* IV.F (for incentives on why the social media company will want to pay for this arbitration system).

158 *See generally* Kelsey L. Swaim, *Alternative Dispute Resolution and Social Media: How Mandatory Arbitration Clauses Impact Social Networking*, 5 Y.B. Arb. & Mediation 356 (2013) (for an understanding of Instagram's terms and conditions including a forced arbitration clause).

159 Fast track arbitration is defined as a full arbitration process that is compressed to have a quicker resolution of the dispute. Jus Mundi, *Fast Track Arbitration: A time-efficient procedure that could hinder the award?*, JUS MUNDI BLOG (May 29, 2020), https://blog.jusmundi.com/fast-track-arbitration-a-time-efficient-procedure-that-could-hinder-the-award/ [https://perma.cc/B87N-BHGE]; *see generally Fast Track Administered Arbitration Rules*, INT'L INST. FOR CONFLICT PREVENTION & RESOL. (July 1, 2020), https://www.cpradr.org/resource-center/rules/arbitration/fast-track-administered-arbitration-rules [https://perma.cc/4CZK-KRBV].

160 *See* discussion *supra* Section III.B.3.

part of this process. When a user gets to this stage of arbitration, the case will be automatically referred to this center to begin arbitration. This center will have a list of arbitrators they hired who comply with a certain set of guidelines and are licensed arbitrators with a proven track record of neutrality. Additionally, the list of arbitrators will contain members from a wide variety of cultures, nationalities, and age ranges. Then the center will provide the user and the social media company with their list of arbitrators, and both the disputing party and the user each pick an arbitrator from that list, and those two arbitrators will pick a third arbitrator. This is called a tri-panel arbitration system and is one of the most common forms of arbitration.[161] After the parties are notified of the arbitration process, and the arbitrators are selected, the tri-panel will review the available information and can request more information from the parties as necessary. The arbitrators will then discuss and decide amongst themselves the appropriate action that should be taken while considering the social media's original decision to take down the user's account following the provided guidelines, the user's intent behind their account, and any context relating to who the user is and their following. The three arbitrators will then decide whether the account should be banned within a week's time. This is a binding arbitration system, and decisions cannot be reversed, but a summary of the reasoning is given to the user, including the steps that went into making the decision.[162] The reasoning will not be confidential, and a shortened summary will also be given to the public. The reason for this is to increase transparency to the public and incentivize other users to think about the consequences of their posts. While the arbitration proceeding is happening, the account is considered under review; thus, the user cannot post anything, nor can other users see their past posts. The three main issues with content regulation now are: (1) lack of context, (2) human bias, and (3) lack of transparency. Arbitration resolves all three of those issues.

---

161 *Question: A Single Arbitrator or Three-Arbitrator Panel?: Answer – a Two-Arbitrator Panel*, ABA (Dec. 11, 2020), https://www.americanbar.org/groups/construction_industry/publications/under_construction/2020/winter2020/single-arbitrator-or-three-arbitrator-panel/ [https://perma.cc/3ABK-KU6L].

162 There are three types of arbitration awards: (1) standard award that states the decision without giving any reason for it, (2) reasoned award that explains why and how the arbitrators came to their decision, and (3) an award that includes findings of fact and law. My proposed system of arbitration will include a reasoned award to provide the parties with transparency. *See generally Types of Final Arbitration Awards: Why the Choice Matters*, STRADLEY RONAN STEVENS & YOUNG, LLP (Feb. 2020), https://www.stradley.com/-/media/files/publications/2020/02/adr-advisor—february-2020.pdf [https://perma.cc/985M-Q3HT].

This proposal will be modeled after the Uniform Domain-Name Dispute Resolution Policy (UDRP)—a private arbitration proceeding to recover a domain name if the owner of the domain violated trademark laws.[163] The UDRP requires three elements to begin a complaint, similar to how the arbitration system will require a threshold of followers and flags to begin arbitration; UDRP also has a panel of neutrals from a dispute resolution service provider who considers several non-exclusive factors when making their decision, similar to how the tri-panel of arbitrators outsourced from the dispute resolution center will analyze a multitude of factors including the community guidelines and the user's account, and the UDRP, just like my arbitration system, is a fast track and less costly method of decision making that is published on the internet.[164]

C. *Arbitration & Context*

Given that this arbitration system will be used in conjunction with the automated system, it will not resolve the issues of individualized content blocking but instead will be used as another layer for high-profile user accounts to be viewed within a broader context than an AI machine could. A big issue that content moderation lacks when banning users is a full picture of the user and their activity on the platform. A user typically gets banned after their content has been removed a certain number of times,[165] but that means the current system will not typically evaluate all the removed content to see if it was mistakenly removed or if there is content that should have been removed but was not. When an arbitrator has context, they can ban a user who might not have been banned in the first place because the system missed posts that

---

163 *Uniform Domain-Name Dispute Resolution Policy*, ICANN, https://www.icann.org/resources/pages/help/dndr/udrp-en [https://perma.cc/88RT-KY23] (last visited Feb. 11, 2022); *WIPO Guide to the Uniform Domain Name Dispute Resolution Policy (UDRP)*, WORLD INTELL. PROP. ORG., https://www.wipo.int/amc/en/domains/guide/#e [https://perma.cc/R2KQ-KRUU] (last visited Feb. 11, 2022) (the administrative panel is composed of one or three neutrals appointed by the dispute resolution service provider and they can make one of three decisions–decide in favor of the entity that filed the complaint and order that the disputes domain name be transferred to that entity, decide in favor of entity that filed complaint and order the disputed domain name be cancelled, or decide in favor of the domain name registrant).

164 *Rules for Uniform Domain Name Dispute Resolution Policy (the "Rules")*, ICANN (Sept. 28, 2013), https://www.icann.org/resources/pages/udrp-rules-2015-03-11-en [https://perma.cc/8DBB-CLBM] (for more information about the policies and procedures of the UDRP).

165 Klonick, *supra* note 14, at 1647.

should have been removed or keep a user on the platform that would normally be banned by finding that the system over removed their posts without proper reason for removal.

Having a tri-panel of arbitrators seeing the account themselves and then discussing and debating with two other arbitrators about such evidence allows for a more individualized and more accurate determination of whether the account should be removed.[166] The arbitrators' review considers the context of all aspects of a user's account. By relying not only on the social media providers' community guidelines to observe what the company deems to be a violation but also observing who the user is, their background, the amount and types of posts that were constantly being flagged, who their followers are, etc., can provide a better analysis on whether this user should be banned to prevent the influence of hate speech and violence or whether other users are simply gaming[167] the system to remove someone, they do not like. Users can trigger this arbitration system by continuously flagging the high-profile user they want to be removed, but the benefit of arbitration as compared to the current system is that this gaming tactic will not automatically result in getting a user banned. Arbitrators will appropriately review all the posts, reasons for the flagging, etc., and be able to detect this tactic of gaming.

The tri-panel of arbitrators will discuss amongst themselves the weighting of factors in each decision. An issue with the current system of content regulation is that every disciplinary action is evaluated using the same set of rules through their community guidelines, and context is not given to each user. Thus, a standardized weighting system for each of the factors will not be provided, especially because the users that would be participating in this arbitration process will be high-profile and will require individualized

---

166 Adrian Bastianelli, *Question: A Single Arbitrator or Three-Arbitrator Panel? Answer: A Two-Arbitrator Panel*, A.B.A. (Dec. 11, 2020), https://www.americanbar.org/groups/construction_industry/publications/under_construction/2020/winter2020/single-arbitrator-or-three-arbitrator-panel/ [https://perma.cc/S6UF-G2BF] ("There is a perception among many in the industry that three-arbitrators are likely to reach a more informed, accurate, and balanced award than a single arbitrator, and the outrageous or extreme result is less likely to occur. . .Some of a lawyer's best thinking and analysis is accomplished through discussion with another lawyer. Just verbalizing an argument can provide a significant benefit in developing a cogent, well-reasoned analysis. With a single arbitrator, there is no one for the arbitrator to use as a sounding board and discuss the issues with. The award must be developed in a vacuum, which in many cases diminishes its quality. Maybe the greatest benefit of a three-arbitrator panel is the availability of other arbitrators to confer with.").

167 Gaming is defined as multiple users ganging up on a single user by consistently flagging their posts disingenuously in order to remove them from the platform.

and unique evaluations.[168] Additionally, by limiting the arbitration system to high-profile users, the arbitrators can understand the scope and influence such users have on society. "Context matters when assessing issues of causality and the probability and imminence of harm," the Facebook Supreme Court Board wrote; "What is important is the degree of influence that a user has over other users."[169] Furthermore, having arbitrators with different cultural backgrounds and nationalities can help solve some of the many issues with the current system, including an insufficient understanding of linguistical and cultural nuances.[170] The arbitration system would also provide the arbitrators with time to process and understand the situation, rather than having to decide whether to remove a user within a millisecond of time.[171] Users should anticipate the process to take a week, but in more difficult cases, an exception can be made to prolong the arbitration process.

Facebook spokesman Andy Stone said, in defense of their XCheck system,[172] that the system was designed "to create an additional step so we can accurately enforce policies on content that could require more understanding."[173] Content providers are aware that there are scenarios in which content regulation needs more individualized attention, and often these scenarios come up with high-profile users who are on the verge of being banned, which are the users that this arbitration system seeks to regulate.

D. *Arbitration & Bias*

With a tri-panel arbitration system, there is an increased probability of a more balanced approach to the decision-making

---

168 Bastianelli, *supra* note 166 ("A mistake is less likely to make it into the award where there are three arbitrators providing input.").

169 Ovide, *supra* note 38.

170 *See* discussion *supra* Section III.B.2; *see also*, Heidi Tworek et al., *Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective*, TRANSATLANTIC WORKING GRP. (Jan. 14, 2020), https://cdn.annenbergpublicpolicycenter.org/wpcontent/uploads/2020/05/Dispute_Resolution_TWG_Tworek_Jan_2020.pdf [https://perma.cc/GKN7-VL2Z] (compare the European press council that applies national codes of ethics to this arbitration system that will apply national freedom of expression standards and democratic ideals to better conceptualize context of the user, their posts and who is flagging them).

171 *See* discussion *supra* Section III.B.2.

172 *See* discussion *supra* Section III.B.3.

173 Horwitz, *supra* note 114.

process,[174] and tri-panels are better used when an all-or-nothing decision need to be rendered, such as to ban or not ban an account. The arbitrators, just like the human content moderators, will still have to make decisions as to what, for example, hate speech means in the context of the companies' guidelines, but unlike the human content moderators who decide on an individual basis, having three arbitrators allows for a multitude of perspectives and the arbitrators' individual opinions would not be able to break through and influence the final decision. Moreover, the list of arbitrators that my proposed Dispute Resolution Center provides are independent licensed arbitrators who adhere to not only international arbitration guidelines but also to the center's own guidelines. Impartiality is essential, and arbitrators, like mediators and conciliators, are third-party neutrals within the Alternative Dispute Resolution ("ADR") system.

Although there is concern that party-appointed arbitrators are not neutral, having the social media companies outsource their high-profile decisions to a third party would not bias the process. The social media provider will have to sign a contract with the Dispute Resolution Center that states the social media provider is bound to pick an arbitrator from the center. Additionally, the user will be bound by the same terms when they sign the terms of service, which prevents one party from having an advantage over the other. Although the social media company is paying the center, the center has full autonomy to find arbitrators who would adhere to their established guidelines, as well as guidelines from the American Arbitration Association (AAA), International Centre for Dispute Resolution (ICDR),[175] and the Code of Ethics for Arbitrators in Commercial Disputes (Code of Ethics).[176]

---

174 Hon. John DiBlasi, *The Commercial Arbitration: The Single Arbitrator Versus the Tri-Panel*, JDSUPRA (July 10, 2018), https://www.jdsupra.com/legalnews/the-commercial-arbitration-the-single-63029/ [https://perma.cc/2Y9M-VKF9].

175 *About the American Arbitration Association (AAA) and the International Centre for Dispute Resolution*, AM. ARB. ASS'N, https://www.adr.org/about [https://perma.cc/B4TX-8A6X] (last visited Jan. 20, 2022) (the AAA and the ICDR are not-for profit organizations in the United States and internationally that assists in appointment of arbitrators and helps "move cases through arbitration. . .in a fair and impartial manner until completion."); *see also Commercial Arbitration Rules and Meditation Procedures*, AM. ARB. ASS'N (Oct. 1, 2013), https://www.adr.org/sites/default/files/CommercialRules_Web-Final.pdf [https://perma.cc/X4U6-DG7L] (for a look at arbitration guidelines in the AAA).

176 *Code of Ethics for Arbitrators in Commercial Disputes*, FINRA, https://www.finra.org/arbitration-mediation/code-ethics-arbitrators-commercial-disputes [https://perma.cc/66ZV-W4R3] (last visited Jan. 20, 2022) (The Code of Ethics for Arbitrators in Commercial disputes was first proposed in 1977 by the AAA and the American Bar Association (ABA). The code "provides

The Dispute Resolution Center will follow guidelines similar to the Center for Public Resources Institute for Dispute Resolution (CPR).[177] Under CPR Arbitration Rule 5.1, the tri-panel will consist of two arbitrators, one appointed by each of the parties and a third arbitrator who will chair the tribunal.[178] Moreover, the Dispute Resolution Center will mandate a similar rule to CPR Rule 5.4(d), which advises neither party to disclose information to the arbitrators as to which party selected them, removing any sign of party-designated arbitrators:[179] "[n]o party, or anyone acting on its behalf, shall have any ex parte communications relating to the case with any arbitrator."[180] Modeling Rule 5.4(a) of CPR, the Dispute Resolution Center will provide each party with a list of the arbitrators and disclose any circumstances that can give rise to any degree of doubt regarding the arbitrators' impartiality, and then each party will send back a list of their top three choices ranked.[181] These rules reduce the influence users can have on arbitrators by isolating them from the involved parties and making sure there are no conflicts of interest that might arise—imposing the highest degree of neutrality.

Although there is a greater degree of likelihood that the arbitrators know the user being adjudicated due to the user's high follower count, the Code of Ethics for Arbitrators requires arbitrators to disclose "any interest or relationship likely to affect impartiality or which might create an appearance of partiality,"[182] including

---

ethical guidance for many types of arbitration. . .and includes annotations on how courts have interpreted the riles of the Code of Ethics.").

177 *2019 Administered Arbitration Rules*, CPR INT'L INST. FOR CONFLICT PREVENTION & RESOL. (Mar. 1, 2019), https://www.cpradr.org/resource-center/rules/arbitration/administered-arbitration-rules-2019 [https://perma.cc/YML3-6QA8].

178 Seth H. Liberman, *Something's Rotten in the State of Party-Appointed Arbitration: Healing ADR's Black Eye That is "Nonneutral Neutrals"*, 52 CARDOZO J. OF CONFLICT RESOL. 215, 230 (2004).

179 *Id.*

180 *Id.*

181 If one or both of the parties fail to pick an arbitrator in the allotted time given to pick one, under rule 6.1–6.5, the Center for Dispute Resolution will provide the party(s) with one. *2018 CPR Non-Administered Arbitration Rules*, CPR INT'L INST. FOR CONFLICT PREVENTION & RESOL. (Mar. 1, 2018), https://www.cpradr.org/resource-center/rules/arbitration/non-administered/2018-cpr-non-administered-arbitration-rules [https://perma.cc/Y73R-4YAV] [hereinafter CPR Rules].

182 CODE OF ETHICS FOR ARBS. IN COM. DISPS., Canon II(A) (AM. ARB. ASS'N 2004).; *see* Mitchell Zamoff & Leslie Bellwood, *Proposed Guidelines for Arbitral Disclosure of Social Media Activity*, 23.1 CARDOZO J. CONFLICT RESOL. 1 (2022) (for more information about disclosing social media activity to prevent conflicts of interest); *see also* CPR Rules, *supra* note 181 (for CPR arbitration rule 7.3).

bias toward the user or the social media provider. Additionally, the parties in the arbitration proceeding would be entitled to disclosures of the arbitrator's social media activity "so they can realize one of the primary benefits of arbitration–the ability to meaningfully participate in the selection of an impartial arbiter."[183]

Moreover, as Facebook acknowledged in its creation of the Facebook Supreme Court, having a diverse group of arbitrators is integral to creating a bias-free environment and decision-making process.[184] The diversity criteria should include not only geographical, ethical, and religious differences but also the age of the arbitrators. Studies have shown that age is a bias factor in a judge's decision-making.[185] In a study that analyzed the influence of age on judicial decision-making in age discrimination cases, the results indicated that the youngest judges were less sympathetic than older judges to victims of age discrimination.[186] The conclusion: a judge's decision-making changes over time depending on age. The age of the arbitrator can subsequently have a drastic effect on how they view certain social media posts and users.[187] An older judge may not understand slang language used by the younger generations, for example, how the term "queer," a former slur, is now reclaimed by the LGBTQIA+ population.[188] The answer to this issue is to have a variety of arbitrators from different age groups, that way the parties have the ability to choose an arbitrator they feel might understand them better. A younger user can pick an arbitrator of similar age out of the list, for example, so that they won't feel the process has been biased against them and that the arbitrators understand the context of the user's posts fully.

---

183 Zamoff & Bellwood, *supra* note 182, at 2.

184 *See* discussion *supra* Section III.B.2.

185 Kenneth Manning et al., *Does Age Matter? Judicial Decision Making in Age Discrimination Cases*, 85 Soc. Science Q. 1, 1 (2004), https://www.jstor.org/stable/42955923 [https://perma.cc/22F9-B6AX].

186 *Id.*

187 *See Id.*

188 Safe Zone Project, *Isn't "Queer" a Bad Word?*, https://thesafezoneproject.com/faq/isnt-queer-a-bad-word/ [https://perma.cc/VK29-JWDA] (Queer is now used as a term of pride for the younger generation, while for the older generation queer can still be considered a bad word) (last visited Jan. 6, 2022).

E. *Arbitration & Transparency with Due Process*

In arbitration, the arbitrators will not solely rely on a set of guidelines or rules to determine if an account violates those rules; rather, they will decide on an individual basis and provide clear results and reasons for the outcome to the affected user and the public, removing many of the due process concerns. Instead of being kicked off a platform with no sufficient reason or ignored by the process, without the ability to appeal, users who are eligible for arbitration would get the transparency they deserve. Arbitration is comparable to litigation and can provide that "day in court" users often complain is lacking. By having a tri-panel of arbitrators make binding decisions, they act as a quasi-judiciary hearing both sides of a case.[189] This, of course, cannot be provided for every user who has been banned and seeks an explanation, but at the same time, this will be a step in the right direction in committing to due process principles. Additionally, if the social media providers see profit from such transparency and due process, they themselves might be more willing to disclose how they moderate content, publish transparency reports, and provide a better appeals process.

F. *Incentives to Use Arbitration*

Although arbitration will be more expensive and time-consuming than the current system, it will only be used on a limited number of accounts with a high threshold of followers that have been flagged by a high threshold of users. Applying these parameters will not only reduce the expense of an arbitration by lowering the number of times arbitration is used, but it will have more substantial results because the decision to ban an account will be made in cases of highly controversial and well-known users that have greater potential for harm. Additionally, this system will be far cheaper and more efficient than litigating the decisions while maintaining many of the same aspects of litigation.[190] In order to prevent users from gaming the system and flooding the arbitration system, the automated AI system and the use of human content moderators will place precautions to watch out for this gaming tac-

---

189 Emily Holland, *Arbitration vs. Litigation: What is the Difference?*, ADR TIMES (Mar. 25, 2021), https://www.adrtimes.com/arbitration-vs-litigation/ [https://perma.cc/R7YP-D2XQ].

190 Liberman, *supra* note 178, at 219 ("Arbitration provides impartial decisions from a third party that can be achieved expeditiously and for less money than traditional litigation.").

tic. The automated system and human content moderators are still able to detect clear violations of community guidelines and often do so accurately. Thus, the use of the current system, alongside arbitrators being able to spot gaming tactics when analyzing context,[191] will hopefully minimize the effects of such gaming tactics.

Social media companies would be incentivized to pay for this outsourced decision-making process because they want their platforms to be clean from hate speech, misinformation, and violent content, but they do not want to take responsibility for cleaning it up.[192] Arbitration would provide this cleaner system while allowing companies to wash their hands of it. Moreover, companies themselves are beginning to call for government/third-party intervention.[193] Furthermore, social media companies have in the past worked together to establish procedures for content moderation,[194] so there is already precedence to work together to create an industry-standard practice of using this third-party arbitration system. Social media providers also know that this system is feasible, given eBay's success in implementing a similar system of online dispute resolution ("ODR"). Under eBay's "independent feedback review,"[195] a seller can challenge a posted review by a buyer, and through eBay's ODR process, which includes an impartial third-party outsourced from a dispute resolution service, the neutral can determine whether to affirm, withdraw, or take no action on the post.[196] Additionally, through the eBay "money back guarantee" system, a buyer has a right to file a complaint if an item has not

---

191 *See* discussion *supra* Section IV.C.

192 Adam Satariano & Mike Isaac, *The Silent Partner Cleaning Up Facebook for $500 Million a Year*, N.Y. TIMES (Oct. 28, 2021), https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html [https://perma.cc/7W99-9M4W] (Behind the scenes, Facebook has outsourced its content regulation responsibility. "Since 2012, the company has hired at least 10 consulting and staffing firms globally to sift through its posts. . .in an effort. . .to distance itself from the most toxic part of its business.").

193 Amanda Macias, *Facebook CEO Mark Zuckerberg Calls for More Regulation of Online Content*, CNBC (Feb. 15, 2020, 1:36 PM), https://www.cnbc.com/2020/02/15/facebook-ceo-zuckerberg-calls-for-more-government-regulation-online-content.html [https://perma.cc/Y57J-2ZNL] (for more information about Zuckerberg's call for government regulation to help with the "growing problem of harmful online content.").

194 Tworek, *supra* note 170, at 7 ("An example of industry cooperation is the Global Internet Forum to Counter Terrorism (GIFCT), established in 2017 by Facebook, Twitter, YouTube and Microsoft. It enables coordination between social media companies seeking to remove "terrorist" content, and the GIFCT houses a hash database of "terrorist images.").

195 Amy J. Schmitz, *Expanding Access to Remedies Through E-Court Initiatives*, 67 BUFF. L. REV. 89, 99–100 (2019), https://digitalcommons.law.buffalo.edu/cgi/viewcontent.cgi?article=4724&context=Buffalolawreview [https://perma.cc/D5CR-KCDU].

196 Tworek, *supra* note 170, at 12.

been received or the item came not as promised and if the seller does not respond or provide an adequate remedy within a certain time period, the buyer can ask eBay to assign an ODR neutral to review the case and make a binding determination.[197]

A social media provider's main concern is profit, and they gain profit when users are engaged on their site.[198] Thus, moderation can affect whether users choose to continue being a member of their community.[199] With users getting increasingly frustrated with the lack of transparency on content regulation, users are more likely to disengage. Thus, when explanations for content removal are available to users, the users may be more likely to change their behavior and become more proactive members of that community.[200] Furthermore, if users find that the explanations the platform gives are not justified, it may discourage them from posting on that platform again, or it could provoke recalcitrant behavior.[201] An arbitration system will increase transparency by providing users not only with explanations but justified and detailed explanations for why they were removed, which would encourage users to then stay on the platform. Although increased transparency would require more effort on the part of these neutral arbitrators and thus more cost, implementing transparency through arbitration is worth this additional cost, especially in lieu of the pressure social media providers are getting from Congress and the public alike. Given that this arbitration system will be geared toward high-profile users, there is a lower likelihood that this user will then retaliate against X company by posting on Y social media site misinformation and hate speech about X company. This would decrease the chance their millions of followers would go along with retaliating against X company. By maintaining their public image, the social media providers will profit.

---

197 Schmitz, *supra* note 195, at 98.

198 Social Media companies' main source of revenue is through targeted advertisements that the users engage with. Thus, if users disengage from the social media platform, advertisers will subsequently follow, and the social media company will lose profit. Greg McFarlane, *How Facebook (Meta), Twitter, Social Media Make Money From You*, INVESTOPEDIA (Nov. 4, 2021), https://www.investopedia.com/stock-analysis/032114/how-facebook-twitter-social-media-make-money-you-twtr-lnkd-fb-goog.aspx [https://perma.cc/PN25-QMQ5].

199 Jhaver, *supra* note 107, at 8.

200 *Id.* at 8–9.

201 *Id.*

V. Conclusion

With the increase in online extremism, hate speech, and misinformation during the COVID-19 Pandemic and Donald Trump's presidency, there has been a persistent public outcry for a more aggressive response to content regulation on social media. Social media providers have not ignored these signs, in fact, many, like Twitter and Facebook, have committed to doing better in halting the spread of such dangerous content, but while they have outwardly committed to fixing the problem, these content providers, particularly Facebook, have continuously chosen profit over safety. These promises to do better "have largely been reactive: social platforms have not embraced the concept of proactively reducing abusive content. Only after the most egregious abuses, and particularly following threatened legal action or loss of advertisers, have social platforms responded . . . making (often minor) policy changes."[202] Social media companies should have a third-party outsourced Dispute Resolution Center conduct arbitration proceedings that may be invoked to suspend high-profile accounts. These proceedings would lead to better outcomes than the current system of content regulation by reducing a lack of context and bias in the moderation system and increasing transparency and due process for users.

---

[202] Brown, *supra* note 12, at 455.