Harvard Journal of Law & Technology
Volume 35, Number 1 Fall 2021

LESSONS FROM TRUMP'S SUSPENSION: HOW TWITTER SHOULD CLARIFY AND STRENGTHEN ITS "PUBLIC INTEREST" APPROACH TO MODERATING LEADERS' VIOLENCE-INSPIRING SPEECH

*Erika Suh Holmberg**

TABLE OF CONTENTS

* Harvard Law School, Candidate for J.D., 2022. My sincerest thanks to Professor Martha Minow for her insight and encouragement as she advised me throughout the process of writing a previous iteration of this Note as an independent project. I would also like to thank Evelyn Douek, whose passion for and expertise in content moderation policy sparked my initial interest in this topic, for her guidance and feedback.

I. INTRODUCTION

Twitter's decision to permanently suspend @realDonaldTrump after the January 6, 2021, insurrection marked a significant departure from both its stated policies and its recent practices. In 2019, the company had announced a "public interest framework," formalizing a practice of content moderation leniency toward major world leaders' Tweets.[1] Throughout Trump's presidency, Twitter had held world leaders to a more lenient standard than other users — declining to remove their Tweets or suspend their accounts, even if they violated platform-wide rules, so long as the actual or potential harmfulness of their continued presence on Twitter did not outweigh the presumed inherent value of preserving public access to these leaders' online speech.[2] Trump's suspension marked the first time Twitter banned a head of state since adopting this framework.[3]

Twitter's decision to permanently suspend Trump and its stated justifications for doing so[4] raise important questions regarding how Twitter will moderate the actions of political leaders in the future. For example, why hadn't Twitter banned Trump earlier? Will Twitter revise its existing moderation policies and practices to enforce a similar context-aware approach against other global leaders who might use their accounts to incite violence?[5] Will Twitter apply a similar form of analysis as the one described in its statement on Trump's suspension if

---

1. This Note uses the term "public interest framework" to refer to Twitter's 2019 policies and statements regarding its lenient approach to moderating a defined class of world leaders, set forth in the following sources: Twitter Safety, *Defining Public Interest on Twitter*, TWITTER: BLOG (Oct. 15, 2019), https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html [https://perma.cc/K4QQ-TA6K]; Twitter Inc., *World Leaders on Twitter: Principles and Approach*, TWITTER: BLOG (Oct. 15, 2019), https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html [https://perma.cc/G4HE-9VGF] [hereinafter *Twitter, 2019 World Leaders Policy Statement*]; *General Guidelines and Policies: About Public-interest Exceptions on Twitter*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/public-interest [https://perma.cc/PAQ3-8PJS] [hereinafter *Twitter General Guidelines, Public-Interest Exception*]. This Note also uses the term "framework-eligible" leaders to refer to those leaders whose Tweets qualify for review under the public interest framework in the context of violence-inciting Tweets. The framework encompasses other types of harmful Tweets beyond potentially violence-inciting or threatening Tweets, such as election misinformation and hate speech, that are beyond the scope of this Note's analysis and recommendations.

2. *See* Sara Morrison, *Facebook and Twitter Made Special World Leader Rules for Trump. What Happens Now?*, VOX: RECODE (Jan. 20, 2021, 8:00 AM), https://www.vox.com/recode/22233450/trump-twitter-facebook-ban-world-leader-rules-exception [https://perma.cc/HS7E-DLNZ].

3. *Id.*

4. Twitter Inc., *Permanent Suspension of @realDonaldTrump*, TWITTER: BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/ZC5N-YMAC] [hereinafter *Twitter, Permanent Suspension*].

5. *See* Adam Satariano, *After Barring Trump, Facebook and Twitter Face Scrutiny About Inaction Abroad*, N.Y. TIMES (Jan. 17, 2021), https://www.nytimes.com/2021/01/14/technology/trump-facebook-twitter.html [https://perma.cc/V3UM-XNM2].

other leaders do in fact use Twitter to incite violence?[6] Until and unless Twitter meaningfully answers these open questions, world leaders and their followers will be left guessing as to which leaders Twitter might de-platform next, and the same failures that resulted from Twitter's current content moderation policies may enable another world leader to use Twitter to inspire political violence.

With these issues in mind, this Note argues that Trump's suspension and eventual ban from Twitter illustrate key defects and ambiguities in the policies and practices Twitter uses to moderate political leaders' "dangerous speech"[7] under the public interest framework. This Note proposes mechanisms through which Twitter can ameliorate these weaknesses without undervaluing the public interest benefits and free speech principles that the framework aims to protect. Given the particularly significant impact of world leaders' online speech on real-world safety, this Note offers reforms to the framework's approach to the specific harm of potential violence incitement as a first step in Twitter's broader process of rethinking its lenient approach to moderating world leaders. Though Twitter functions as a public square in some respects, as a private platform, under the current U.S. legal and regulatory landscape, it is free to create and enforce its own rules, suspend users, and remove or leave up Tweets at will. So long as this status quo continues, incorporating the free speech principles and trade-offs into the public interest framework in the manner this Note proposes can better ensure that Twitter will wield its tremendous speech-policing power more responsibly and transparently.[8]

---

6. *See* Tom Phillips, Hannah Ellis-Petersen, Shaun Walker & Julia Carrie Wong, *Trump Social Media Ban Sparks Calls for Action Against Other Populist Leaders*, GUARDIAN (Jan. 17, 2021), https://www.theguardian.com/media/2021/jan/17/trump-social-media-ban-jair-bolsonaro-narendra-modi [https://perma.cc/9R2U-J4PT] (quoting David Kaye); Miriam Berger & Elizabeth Dwoskin, *Trump Ban by Social Media Companies Came after Years of Accommodation for World Leaders Who Pushed the Line*, WASH. POST (Jan. 15, 2021), https://www.washingtonpost.com/world/2021/01/15/world-leaders-facebook-twitter-trump-ban/ [https://perma.cc/R86F-JAE5].

7. This Note uses the term "dangerous speech" to refer to online speech that tends to "increase the risk that its audience will condone or commit violence against members of another group." DANGEROUS SPEECH PROJECT, DANGEROUS SPEECH: A PRACTICAL GUIDE (2021), https://dangerousspeech.org/wp-content/uploads/2020/08/Dangerous-Speech-A-Practical-Guide.pdf [https://perma.cc/46NZ-MU5N].

8. For discussion of the well-established American approach to social media regulation, which allows platforms absolute freedom to regulate users' speech on their platforms without potential liability or legal obligations under the First Amendment, see generally Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018); Daphne Keller, *Six Constitutional Hurdles for Platform Speech Regulation*, STAN. L. SCH. CTR. FOR INTERNET & SOC'Y (Jan. 22, 2021, 6:50 AM), http://cyberlaw.stanford.edu/blog/2021/01/six-constitutional-hurdles-platform-speech-regulation-0 [https://perma.cc/6CWF-6MMH]; Eric Goldman, *An Overview of the United States' Section 230 Internet Immunity*, *in* OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY 155 (Giancarlo Frosio ed., 2020); Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353 (2018).

Part II provides a brief overview of Twitter's publicized moderation policies on content posted by world leaders that potentially incites violence, the evolution of these policies during Trump's presidency, and tensions between the formal policy itself and Twitter's stated reasons for banning Trump.

Part III, Section A argues that whenever framework-eligible leaders Tweet "clear and direct threats of violence against an individual,"[9] Twitter must hold them to the same enforcement standard as the general public without consideration of the possible public interest value of the Tweet(s) due to the particularly high risk of inciting grievous harm and the relative weakness of public interest justifications in this context. In order to resolve existing ambiguity in Twitter's definition of prohibited "clear and direct threats of violence against an individual" that will "result in enforcement action" regardless of the author's status,[10] Twitter should clarify that the only type of "context" that may inform these decisions is the context of on-platform surrounding circumstances. Twitter must also clarify whether similar threats against *groups* are also subject to immediate removal.

Part III, Section B addresses Tweets by framework-eligible leaders that may carry a substantial risk of causing violence even when they do not constitute "clear and direct threats against an individual," as illustrated by Trump's Tweets that emboldened his supporters to violently storm the Capitol but fell short of explicitly threatening or encouraging those specific violent acts. In these myriad "harder cases," accurately assessing the risk of violence requires a more context-aware approach that considers how users interpret and react to both on- and off-platform content. Part III, Section B proposes three ways in which Twitter can and should change its policies and practices on moderating ambiguous but potentially violence-inspiring Tweets by framework-eligible leaders to more effectively gauge the risk of violence to be balanced against the speech's public interest value.

Part III, Section C explores an example of Tweets posted by a particular politician to illustrate how these proposed changes would function in practice. Finally, this Note offers concluding thoughts on

---

Whether this Note's proposed content moderation approaches in the context of violence incitement should also apply to Twitter's moderation of other types of harmful speech, such as misinformation or hate speech, is outside the scope of this Note.

In addition, this Note focuses exclusively on Twitter due to space constraints, the popularity of Twitter among world leaders, and the relative feasibility of adopting context-aware content moderation on the platform given Twitter's existing infrastructure. *See* Thomas Zeitzoff, *How Social Media is Changing Conflict*, 61 J. CONFLICT RESOL. 1970, 1971 (2017) ("Over 75 percent of world leaders have an active Twitter or Facebook account."). *See generally General Guidelines and Policies: Our Approach to Policy Development and Enforcement Philosophies*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/enforcement-philosophy [https://perma.cc/ZU72-HRNV].

9. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

10. *Id.*

solutions to mitigate concerns that this proposed approach grants undue discretion and power to Twitter in policing online speech or heightens the risk of selective enforcement.

II. THE CURRENT LANDSCAPE

A. A Brief History and Overview of the "Public Interest Framework"

Prior to 2019, Twitter responded to criticism of its non-interference with controversial Tweets by pointing to the inherent "newsworthiness" and public interest value of world leaders' speech, emphasizing the "critical role" political leaders play in public conversation.[11] In June 2019, Twitter created a new formalized "notice" system, under which world leaders' Tweets that violate the Twitter Rules remain on the platform, but are shielded behind an interstitial user notice explaining that the shielded Tweet violates a rule,[12] but "it may be in the public's interest for the Tweet to remain available."[13] This system exclusively applies to Tweets from accounts that (1) represent a "government/elected official," a candidate for public office, or a government appointee, (2) have over 100,000 followers, and (3) are verified.[14]

In October 2019, Twitter expanded the notice system by announcing a broader "public interest framework" in an official policy statement (the "World Leaders Policy Statement").[15] The platform also added a provision to its general guidelines on content moderation that codified the new "public interest exception" system.[16] To qualify for framework analysis, an account must meet the same criteria defined

---

11. Klonick, *supra* note 8, at 1665 & nn.465–66 ("In September 2017, Twitter announced that it had a different content-moderation rule set for removing President Trump's tweets. . . . It is important to note that the uses of 'public figure' and 'newsworthiness' here differ from their meanings in the sense of communications or privacy torts.") (citing Arjun Kharpal, *Why Twitter Won't Take Down Donald Trump's Tweet Which North Korea Called a "Declaration of War"*, CNBC (Sept. 26, 2017, 2:56 AM), https://www.cnbc.com/2017/09/26/donald-trump-north-korea-twitter-tweet.html [https://perma.cc/HAE4-5ZY9]); @Twitter, *World Leaders on Twitter*, TWITTER: BLOG (Jan. 5, 2018), https://blog.twitter.com/en_us/topics/company/2018/world-leaders-and-twitter.html [https://perma.cc/W6Z8-TH2S] (Twitter stating in 2018 that "[b]locking a world leader from Twitter or removing their controversial Tweets would hide important information people should be able to see and debate," which would "not silence that leader" but would "certainly hamper necessary discussion around [leaders'] words and actions.").

12. *General Guidelines and Policies: Notices on Twitter and What They Mean*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/notices-on-twitter [https://perma.cc/85KQ-MJQQ].

13. Twitter Safety, *Defining Public Interest on Twitter*, *supra* note 1. Shielded Tweets also are not algorithmically amplified. *Id.*

14. Twitter Safety, *Defining Public Interest on Twitter*, *supra* note 1.

15. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

16. *See Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

under the notice system.[17] Under the framework, Twitter's Enforcement Team evaluates reported Tweets from framework-eligible leaders against Twitter's rules, focusing on the plain language of the Tweet without "attempt[ing] to determine all potential interpretations of the content or its intent."[18] If the Enforcement Team determines the Tweet violates a rule, the Trust & Safety Team then shares a recommendation on "whether or not continued access to the Tweet is in the public interest" with a "cross-functional set of leaders across different internal teams with diverse and multidisciplinary backgrounds . . . as well as in-market teams with an understanding of the cultural context in which the Tweet was posted."[19] After receiving the stakeholders' feedback, Trust & Safety Team senior leaders decide whether to leave the Tweet up behind a notice — thereby implementing the "public interest exception" to standard enforcement of the Twitter Rules — or remove the Tweet, balancing "potential risk and severity of harm" against public interest value.[20]

According to its general guidelines provision on the public interest exception, Twitter is "less likely to make [public interest] exceptions" when a Tweet "threatens or glorifies violence,"[21] and will "especially err on the side of removal . . . where there is evidence the content may be leading to actual or likely offline harm," though in "rare instances," the exception may apply "if there is a more attenuated connection to actual violence."[22] Twitter is also "more likely" to remove a Tweet after public interest review if it includes a "declarative call to action that could harm a specific individual or group."[23]

The World Leaders Policy Statement further clarifies that Tweets containing "clear and direct threats of violence against an individual" will result in "enforcement action . . . *without consideration of the potential public interest value* in allowing the Tweet to remain visible

17. The framework also clarified the definition of "government/elected official," to mean current holders of and candidates or nominees for an elected or appointed membership in a local, state, national, or supra-national governmental or legislative body. *See id.*

18. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

19. *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

20. *Id.*

21. Twitter Safety, *Defining Public Interest on Twitter*, *supra* note 1 (citing *General Guidelines and Policies: Violent Threats Policy*, TWITTER: HELP CTR. (Mar. 2019), https://help.twitter.com/en/rules-and-policies/violent-threats-glorification [https://perma.cc/94PP-LPZS] [hereinafter *Twitter General Guidelines, Violent Threats Policy*]).

22. *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

23. *Id.* Note that Twitter updated its "glorification of violence policy" and "violent threats policy" in the general guidelines in March 2019, the former of which Twitter cited as the basis of its decision to permanently suspend President Trump. *See General Guidelines and Policies: Glorification of Violence Policy*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/glorification-of-violence [https://perma.cc/7QCH-BXPV] [hereinafter *Twitter General Guidelines, Glorification of Violence Policy*]; *Twitter General Guidelines, Violent Threats Policy*, *supra* note 21; Twitter, *Permanent Suspension*, *supra* note 4.

behind a notice."[24] However, "context matters": direct threats made during "interactions with fellow public figures" or when commenting "on political or foreign policy issues would likely not result in enforcement."[25] In other words, the framework specifically groups "direct threats" of violence with certain other types of speech — such as posting someone's private contact information or promoting terrorism — as areas that do not qualify for the public interest exception: leaders making Tweets that fall into these categories are subject to the same enforcement standards as the general public. But when it comes to all other violations of the platform-wide Twitter Rules, including indirectly advocating or glorifying violence, framework-eligible leaders enjoy the increased leniency of analysis under the framework's balancing test.

B. Trump and the Suspension

Prior to the January 6, 2021 insurrection, Twitter had only taken enforcement action against a single Trump Tweet on violence-related grounds: in May 2020, the platform applied a notice to a Trump Tweet warning protestors that "[w]hen the looting starts, the shooting starts."[26] On the day of the Capitol insurrection, Twitter removed three of Trump's Tweets as "severe" violations of Twitter's civic integrity policy, then immediately following the insurrection, "as a result of the unprecedented and ongoing violent situation," Twitter imposed a twelve-hour lock on Trump's account, warning that future rule violations would result in permanent suspension.[27]

---

24. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1 (emphasis added).

25. *Id.* ("We want to make it clear today that the accounts of world leaders <u>are not</u> above our policies entirely. The below areas will result in enforcement action for <u>any</u> account on our service" (emphasis in original)).

26. *See* Twitter Comms (@TwitterComms), TWITTER (May 29, 2020, 3:17 AM), https://twitter.com/TwitterComms/status/1266267446979129345 [https://perma.cc/HG9D-87PX] (stating that the decision to place a public interest notice on the Tweet, which Twitter found violated the Glorification of Violence Policy, was "based on the historical context of the last line, its connection to violence, and the risk it could inspire similar actions today"); Jon Porter, *Twitter Restricts New Trump Tweet for "Glorifying Violence"*, VERGE (May 29, 2020), https://www.theverge.com/2020/5/29/21274323/trump-twitter-glorifying-violence-minneapolis-shooting-looting-notice-restriction [https://perma.cc/2XZE-YL9S]. *See generally* @AngelSDiaz_, *Given Last Week's Escalated Tension Between Platforms and Trump, @laur_hf and I Analyzed Twitter's Public Interest Exception and Facebook's Newsworthiness Policy*, THREAD READER (Jun. 1, 2020), https://threadreaderapp.com/thread/1267462126022676487.html [https://perma.cc/QB8J-NQFJ].

27. Twitter Safety (@TwitterSafety), TWITTER (Jan. 6, 2021, 7:02 PM), https://twitter.com/TwitterSafety/status/1346970430062485505 [https://perma.cc/2MG4-UXWP] (citing Civic integrity policy, TWITTER: HELP CTR. (Oct. 2021), https://help.twitter.com/en/rules-and-policies/election-integrity-policy [https://perma.cc/PHL4-2E5R]); *see also Twitter General Guidelines, Public-Interest Exception*, *supra* note 1 ("Where the risk of harm is higher and/or more severe, we are less likely to make an exception [to the policy]."); Kate Conger & Mike Isaac, *Inside Twitter's Decision to*

Two days later, Twitter followed through on its warning, stating that "after close review of recent Tweets" by Trump "and the context around them — specifically how they are being received and interpreted on and off Twitter — we have permanently suspended the account due to the risk of further incitement of violence."[28] The offending Tweets included one in which Trump celebrated his supporters and one in which Trump announced he would not attend the presidential inauguration of Joe Biden.[29] Twitter claimed these "Tweets must be read in the context of broader events in the country" and how they could be mobilized by supporters "to incite violence . . . in the context of the pattern of behavior from [Trump's] account in recent weeks."[30] Twitter further explained that the Tweets had been evaluated under its Glorification of Violence Policy. The platform stated that its decision had been based on "a number of factors," including indications that the Tweets were being interpreted "on and off Twitter" to encourage further armed protests, plans for which had already started to proliferate both on and off the platform.[31]

Although the decision to suspend Trump was justifiable in order to prevent future political violence in light of the contexts Twitter cited,

---

Cut off Trump, N.Y. TIMES (Jan. 16, 2021), https://www.nytimes.com/2021/01/16/technology/twitter-donald-trump-jack-dorsey.html [https://perma.cc/NET6-DB5J] ("Mr. Dorsey repeated that Twitter should be consistent with its policies. But he said he had drawn a line in the sand that the president could not cross or Mr. Trump would lose his account privileges . . . ."); Brian Heater, *Twitter Locks Trump Out of His Account for at Least 12 Hours*, TECHCRUNCH (Jan. 6, 2021, 7:12 PM), https://techcrunch.com/2021/01/06/twitter-locks-trump-out-of-his-account-for-at-least-12-hours/ [https://perma.cc/G6JU-C4BA]; Adi Robertson, *Twitter Says Trump's Account is Locked, and He's Facing a Ban*, VERGE (Jan. 6, 2021, 7:13 PM), https://www.theverge.com/2021/1/6/22217686/trump-twitter-account-locked-capitol-hill-riot-tweets-policy-violations [https://perma.cc/FT7V-N8LM].

28. Twitter, *Permanent Suspension*, *supra* note 4. As will be further discussed *infra* in Section III.B, this statement apparently conflicts with the text of Twitter's World Leaders Policy Statement, which emphasized that when reviewing world leaders' posts, Twitter focuses on the language itself and immediately surrounding on-platform context, and *not* on how the Tweet is or could be interpreted on Twitter, let alone off-platform. *See, e.g.*, Jacob Schulz, *Twitter Puts an End to Trump's Rhetorical Presidency*, LAWFARE (Jan. 11, 2021, 1:35 PM), https://www.lawfareblog.com/twitter-puts-end-trumps-rhetorical-presidency [https://perma.cc/S767-9LYJ] ("It's wise for Twitter to look to context in making such an important decision, but it's also not consistent with previous interpretive techniques favored by the platform.").

29. Twitter, *Permanent Suspension*, *supra* note 4. As will be discussed *infra* Section III.B, these Tweets were relatively anodyne compared to past Trump Tweets.

30. *Id.*

31. *Id.*; *see* Jack Dorsey (@Jack), TWITTER (Jan. 13, 2021, 7:16 PM), https://twitter.com/jack/status/1349510769268850690 [https://perma.cc/4SW8-5XJ8] (justifying the decision to permanently suspend Trump by pointing to on and off-platform evidence of ongoing threats to physical safety); *see also* Conger & Isaac, *supra* note 27 (detailing CEO Jack Dorsey's reported personal hesitance to ban Trump and internal decision-making processes between January 6 and January 8, including reports from anonymous internal Twitter sources indicating that Twitter's Safety Team had found evidence that immediately after Trump's Tweets early on January 8, supporters were posting plans for further unrest on Twitter and on Parler). *See generally* Satariano, *supra* note 5.

the reasons Twitter provided are in tension with the current framework, which does not mention the relevance of off-platform context or address whether, and under what circumstances, a framework-eligible leader might face account suspension as a consequence of posting potentially violence-inciting Tweets. Therefore, "the question going forward," as law professor David Kaye noted, "is whether this is a new kind of standard [social media platforms] intend to apply for leaders worldwide."[32] At time of publication, Twitter has not published a revision to its public interest framework. However, in March 2021, Twitter solicited public comments on whether users believe leaders "should be subject to the same rules as others on Twitter" and "what type of enforcement action is appropriate" when leaders violate a rule, signaling the company's intent to make "forthcoming revisions" to the framework.[33]

III. PROPOSED CHANGES

Twitter's March 2021 statement evinces awareness that some form of revision to the framework is necessary. As Twitter considers how to proceed, this Part offers initial steps Twitter can and should take to clarify and revise its approach to potential violence incitement by framework-eligible leaders.

---

32. Satariano, *supra* note 5.

33. Twitter Safety, *Calling for Public Input on Our Approach to World Leaders*, TWITTER: BLOG (Mar. 18, 2021), https://blog.twitter.com/en_us/topics/company/2021/calling-for-public-input-on-our-approach-to-world-leaders.html [https://perma.cc/TK2W-AFWS]; *see also* Mitchell Clark, *Twitter Wants to Know if You Think World Leaders Should Get Special Treatment*, VERGE (Mar. 19, 2021, 5:09 PM), https://www.theverge.com/2021/3/19/22340643/twitter-public-survey-world-leader-rules-enforcement [https://perma.cc/6YX2-P8SS].

Facebook's recent decision to formally end its controversial policy of leniency toward leaders' posts might also strengthen the pressure on Twitter to revise or even abandon the public interest framework. In June 2021, Facebook ended its "newsworthiness" policy, which was roughly analogous to Twitter's public-interest framework. *See* Alex Heath, *Facebook to End Special Treatment for Politicians After Trump Ban*, VERGE (June 3, 2021, 4:23 PM), https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy [https://perma.cc/BGP6-QWTA]; *see also* Evelyn Douek, *Facebook's Responses in the Trump Case Are Better than a Kick in the Teeth, but Not Much*, LAWFARE (June 4, 2021, 4:32 PM), https://www.lawfareblog.com/facebooks-responses-trump-case-are-better-kick-teeth-not-much [https://perma.cc/P2TC-6G7V] (noting that although the end of the newsworthiness policy seems like a major reversal at first glance, "Facebook's decision on this is not at all surprising and could result in little substantive change"). *See generally* Thomas E. Kadri & Kate Klonick, *Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech*, 93 S. CAL. L. REV. 37 (2019); Kate Klonick, *Facebook v. Sullivan: Investigating Facebook's Use of the "Public Figure" and "Newsworthiness" Concepts in Content Moderation Decisions*, KNIGHT FIRST AMENDMENT INST. (Oct. 1, 2018), https://knightcolumbia.org/content/facebook-v-sullivan [https://perma.cc/Y43B-7XHB].

A. Continue and Clarify Treatment of Direct Threats as "Exceptions to the Exception"

Twitter's statement on Trump's suspension emphasized that notwithstanding the framework's goal of protecting users' "right to hold power to account in the open," world leaders "cannot use Twitter to incite violence."[34] This apparently refers to the framework's treatment of direct threats of violence against individuals as one of the enumerated exceptions to framework analysis. But the 2019 World Leaders Policy Statement's actual definition of the type of speech that is categorically ineligible for framework leniency contains more nuance. According to that statement, the types of otherwise framework-eligible Tweets that are subject to prompt removal without consideration of their public interest value include "[c]lear and direct threats against an individual (context matters: as noted above, direct interactions with fellow public figures and/or commentary on political and foreign policy issues would likely not result in enforcement)."[35] In order to address such situations and effectively curtail violence, Twitter should revise the "context matters" parenthetical to more clearly define the on-platform context that would warrant applying public interest analysis to an otherwise "clear and direct threat."[36]

Clear-cut direct threats against individuals should remain categorically ineligible for leniency because the framework's balancing test will virtually always weigh in favor of removal. First, the risk of harm is particularly high. When evaluating the risk of real-world harm in a given context, one must consider not only the content of the message itself, but also the speaker's influence, the method and reach of transmission, the audience's susceptibility to persuasion, and any historical violence against the target of the speech.[37] When a framework-eligible leader's Tweets violates the platform wide Violent Threats policy's prohibition on "clear and direct threat against an individual," the message, speaker, and medium automatically produce such a high risk of harm that evaluating off-platform context or audience dynamics and interpretations becomes less essential.[38]

---

34. Twitter, *Permanent Suspension*, *supra* note 4; *see also* Twitter Safety (@TwitterSafety), TWITTER (Jan. 8, 2021, 6:21 PM), https://twitter.com/TwitterSafety/status/1347684879526481925 [https://perma.cc/LT9V-GWXU].

35. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

36. *Id.*

37. *See* DANGEROUS SPEECH PROJECT, *supra* note 7.

38. Consider, for example, the impact of directly targeted threatening Tweets such as "I will give $20,000 to any brave patriot who takes out [named political opponent] once and for all!" or "If [named activist] organizes another rally against me, I will make sure she never walks another day again." These statements have a particularly high risk of causing real harm when they are posted by a framework-eligible leader, who by definition has a large audience and can draw on the influence inherent in political office. *See* Vicki Jackson & Martha Minow, *Facebook Suspended Trump. The Oversight Board Shouldn't Let Him Back*, LAWFARE (Mar.

Secondly, the principles typically invoked to justify ensuring public access to world leaders' speech apply poorly in the context of direct threats of violence. When a leader uses his platform to express intent to commit serious violence against an individual or offers to reward anyone who does, the public interest value[39] of preserving constituents' ability to engage in counter-speech and hold elected leaders accountable ("the accountability principle") dims in comparison to the harm that may result. Even Jameel Jaffer, who directs the Knight First Amendment Institute and believes that the American public had a First Amendment right to fully access and engage with Trump's presidential Twitter account, argued in the context of Twitter's January 6 twelve-hour lock on Trump's account that although typically "the public needs to know what leaders say, even when — and perhaps especially when — what those leaders are saying is wrong or offensive . . . . [t]here are limits to this principle."[40] Namely, "[a] political leader who uses his account to incite violence is causing harms that can't be countered by speech and can't be undone by a future election."[41] In other words, these threats

---

8, 2021, 11:02 AM), https://www.lawfareblog.com/facebook-suspended-trump-oversight-board-shouldnt-let-him-back [https://perma.cc/R2XQ-2QWQ] ("Although what leaders of government have to say may be of unusual public interest, their words can have much greater influence by virtue of their positions of power. Social media platforms work as a megaphone for those already famous, potentially amplifying the instantaneous reach and effect of their speech to the entire world.").

39. *See, e.g.*, *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1 ("[W]e limit exceptions to one critical type of public-interest content — Tweets from elected and government officials — given the significant public interest in knowing and being able to discuss their actions and statements . . . . Twitter stands for the value of direct access to powerful figures, and maintaining a robust public record provides benefits to accountability."); *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1 ("Our mission is to provide a forum that enables people to be informed and to engage their leaders directly."); Twitter Safety, *Defining Public Interest on Twitter*, *supra* note 1 ("By nature of their positions these leaders have outsized influence and sometimes say things that could be considered controversial or invite debate and discussion. A critical function of our service is providing a place where people can openly and publicly respond to their leaders and hold them accountable.").

40. Jameel Jaffer, *Knight Institute Comments on Suspension of Trump's Social Media Accounts*, KNIGHT FIRST AMENDMENT INST. (Jan. 7, 2021), https://knightcolumbia.org/content/knight-institute-comments-on-suspension-of-president-trumps-social-media-accounts [https://perma.cc/2XZX-6LTQ]; *see also* Andrew Marantz, *The Importance, and Incoherence, of Twitter's Trump Ban*, NEW YORKER (Jan. 15, 2021), https://www.newyorker.com/news/daily-comment/the-importance-and-incoherence-of-twitters-trump-ban [https://perma.cc/HM4H-ZD69]. *See generally* Knight First Amendment Inst. v. Trump, 928 F.3d 226 (2d Cir. 2019); Lincoln Caplan, *Near and Distant Objectives*, HARV. MAG. (Sept. 2020), https://harvardmagazine.com/2020/09/features-noah-feldman [https://perma.cc/DT2Y-M4RY] (Jaffer discussing his argument that President's Trump Twitter is a public forum covered by the First Amendment, in contrast with Professor Noah Feldman's opposing view).

41. Jaffer, *supra* note 40; *see also id.* ("When the platforms reasonably conclude that a political leader is engaged in this kind of activity, they're justified in taking his posts down — and in suspending his account . . . . [I]t's the responsible exercise of a First Amendment right."); Evelyn Douek, *Facebook Has Referred Trump's Suspension to Its Oversight Board. Now What?*, LAWFARE (Jan. 21, 2021), https://www.lawfareblog.com/facebook-has-referred-trumps-suspension-its-oversight-board-now-what [https://perma.cc/FSF7-9ZLN] ("A good

carry a particularly high risk that the leader himself or one of his followers will inflict serious irreparable physical harm.[42] Swiftly removing reported Tweets containing clear threats of violence avoids a time-consuming evaluation of contextual factors that would almost surely result in removal, but by which time the leader or one of his followers may have already committed the threatened violence against the individual.[43]

In addition to the relative inapplicability of the accountability principle in extreme cases, excepting leaders from immunity in this context also finds support by analogy to *Nixon v. Fitzgerald*, which held that the President retains "absolute immunity from damages liability for acts within the 'outer perimeter' of his official responsibility" and found such immunity "a functionally mandated incident of the President's unique office, rooted in the constitutional tradition of the separation of powers."[44] Just as law professor Douglas McKechnie argued that President Trump engaging in malicious defamation via Tweet would fall outside the *Nixon* "outer perimeter" of the President's official duties[45] (and thus, holding him liable for those Tweets would not interfere with his ability to faithfully execute the office or raise separation-of-powers concerns), Tweeting a "clear and direct threat" of interpersonal violence clearly does not fall within a leader's official duties.

For these reasons, Twitter must ensure that this narrow category of individualized threats remains an exception to the public interest exception. That said, the "context matters"[46] parenthetical following the "clear and direct threat" language in the World Leaders Policy Statement requires further clarification. The text heavily implies that the "context" that "matters" is limited to readily apparent on-platform context, rather than off-platform factors. This distinction should be made more explicit, distinguishing from the consideration of off-platform

---

argument can be made . . . that democracy requires voters to know who their candidates really are and what they believe, even (or, perhaps, especially) when those beliefs are abhorrent. (This does not and should not apply with respect to incitements to violence.)").

42. *See, e.g.*, *supra* text accompanying note 41.

43. The problem of time-consuming reviews of context — whether on- or off-platform — postponing action until real-world harm is already committed could be at least partially resolved through improvements in technology and human reviewers that monitor the limited class of framework-eligible leaders' accounts more robustly and preemptively, as discussed in detail *infra* Section III.B. But for the less common instances when a Tweet is by its terms a more clear-cut threat against an individual, as this Part addresses, subjecting it to public interest exception analysis would still likely be unnecessary even if that review process were less time-consuming, given the inherent dangerousness of face-value individualized threats when posted by a framework-eligible leader.

44. Nixon v. Fitzgerald, 457 U.S. 731, 749, 756 (1982).

45. *See* Douglas B. McKechnie, *@POTUS: Rethinking Presidential Immunity in the Time of Twitter*, 72 U. MIAMI L. REV. 1 (2017).

46. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1 ("Clear and direct threats of violence against an *individual (context matters: as noted above, direct interactions with fellow public figures and/or commentary on political and foreign policy issues would likely not result in enforcement)* . . . .") (emphasis added).

context when evaluating rule-violative Tweets that do not fall into the enumerated categories (including direct individualized threats) exempt from public interest analysis.[47]

The World Leaders Policy Statement should also clarify whether "clear and direct threats against an individual" *or against a specific group* will be removed without consideration of their public interest value. The statement itself only lists such threats *against an individual*, and the public interest provision of the general guidelines lists "declarative calls to action against an individual" as subject to framework analysis but "more likely" to be removed rather than qualify for the exception.[48] But the policy statement hyperlinks to the Violent Threats policy, which by its terms applies to threats against individuals *and* groups.[49] This distinction is significant: for example, in July 2018, Indian legislator and Hindu nationalist incendiary T. Raja Singh posted a video on Facebook stating that if Rohingya immigrants did not leave India, they should be shot.[50] Had Singh posted the video to his framework-eligible Twitter account, under the current framework the video would be "less likely" to remain online, and would at the very least

47. Twitter should also clarify examples of how a clear and direct threat against an individual in clear violation of the Violent Threats policy might constitute "commentary" on political or foreign policy issues, since the term is not further defined in the public-interest exception guidelines.

48. The 2019 World Leaders Policy Statement limits the scope of Tweets by framework-eligible leaders that will result in "enforcement action" without analysis of its potential public interest value to threats against an *individual*, without mentioning threats directed against a group. In contrast, the public interest provision of the general guidelines lists a "declarative call to action that could harm a specific individual *or group*" (emphasis added) among the types of Tweets that will be evaluated under public-interest exception analysis but are "more likely" to not survive that analysis and be removed rather than remain online behind a notice. *Compare Twitter, 2019 World Leaders Policy Statement*, *supra* note 1, *with Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

49. *Compare Twitter General Guidelines, Public-Interest Exception*, *supra* note 1, *with Twitter General Guidelines, Violent Threats Policy*, *supra* note 21.

50. Ashish Pandey, *Shoot Rohingya, Bangladeshi Immigrants, Says Controversial BJP MLA*, INDIA TODAY (July 31, 2018), https://www.indiatoday.in/india/story/shoot-rohingya-bangladeshi-migrants-says-controversial-bjp-mla-1301394-2018-07-31 [https://perma.cc/B3ZS-NG6W]; *see also Assam NRC: BJP MLA Raja Singh Says Illegal Immigrants Refusing to Go back Should Be Shot*, TIMES OF INDIA: VIDEOS (July 31, 2018, 4:17 PM), https://timesofindia.indiatimes.com/videos/news/assam-nrc-bjp-mla-raja-singh-says-illegal-immigrants-refusing-to-go-back-should-be-shot/videoshow/65213498.cms [https://perma.cc/4YT3-3VD6]. Other dangerous statements from Mr. Singh that Mr. Singh Tweeted will be discussed *infra* Section III.B. Note that in 2020, after Indian political pressure and a Wall Street Journal article that exposed the pro-BJP biases and lobbying for a light-handed approach to moderating Singh by Facebook's top public policy executive in India, in 2020 Facebook permanently banned Singh's account on the basis that he was a "dangerous individual," citing his comments about Rohingya Muslims. *See* Newley Purnell & Rajesh Roy, *Facebook, Under Pressure in India, Bans Politician for Hate Speech*, WALL ST. J. (Sept. 3, 2020, 8:30 AM), https://www.wsj.com/articles/facebook-under-pressure-in-india-bans-politician-for-hate-speech-11599105042 [https://perma.cc/3PRP-GLYY]; *see also* Newley Purnell & Jeff Horwitz, *Facebook's Hate-Speech Rules Collide with Indian Politics*, WALL. ST. J. (Aug. 14, 2020, 12:47 PM), https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346 [https://perma.cc/5E2H-8AKP].

likely be placed behind a notice because it "threatens or glorifies" violence "against an individual or group of people."[51] But if Singh had gone so far as to Tweet "If Rohingya Muslims don't leave India, *I* will kill them," it remains unclear under the current framework's text, when read against the Violent Threats policy it references, whether this threat would be removed without consideration of its public interest value. Thus, Twitter should clarify that "clear and direct" threats against *a group* that obviously violate the Violent Threats policy will be analyzed under the framework but are "less likely" to survive the analysis. Alternatively, Twitter should add the words "or group" after "clear and direct threat against an individual" to the World Leaders Policy Statement.

B. Revise "Public Interest Framework" Analysis to Explicitly Consider Off-Platform Context

Most of the dangerous speech Tweeted by political figures does not rise to the level of "clear and direct threats against an individual" that would constitute a clear violation of the Violent Threats policy.[52] President Trump's Tweets, including the two cited as the basis for his suspension, did not facially violate Twitter's rules or constitute categorically framework-ineligible direct individualized threats.[53]

---

51. *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1. Note that the public interest framework and the violent threats policy are not limited to any particular *type* of group, such as protected groups under the hate speech rules.

52. *See* Susan Benesch, *The Insidious Creep of Violent Rhetoric*, DANGEROUS SPEECH PROJECT (Mar. 8, 2021), https://dangerousspeech.org/noema-the-insidious-creep-of-violent-rhetoric/ [https://perma.cc/2GW6-SE4J] ("The words [of a politician posting online speech that increases the risk of real-world violence] are typically equivocal, as the politician's . . . followers know just as well as the moderators."); *see generally* Dia Kayyali, *If Trump Can Be Banned, What About Other World Leaders Who Incite Violence?*, VICE (Jan. 19, 2021, 9:00 AM), https://www.vice.com/en/article/93wz4z/if-trump-can-be-banned-what-about-other-world-leaders-who-incite-violence [https://perma.cc/CNX6-TMG5]. Twitter's Violent Threats rule, which the "clear and direct threats of violence against an individual" provision of the world leaders statement hyperlinks, defines threats of violence as "statements of an intent to kill or inflict serious physical harm on a specific person or group of people." *Twitter General Guidelines, Violent Threats Policy*, *supra* note 21. "Intent" is defined as including statements like "I will," "I'm going to," or conditional statements like "If you do X, I will [violent act]"; violations also include "asking for or offering a financial reward in exchange for inflicting violence on a specific person or group of people." *Id.*

53. *See, e.g.*, Benesch, *supra* note 52 ("[Trump] typically used ambiguous language, just as most inciters do . . . ."); Fabiola Cineas, *Donald Trump is the Accelerant: A Comprehensive Timeline of Trump Encouraging Hate Groups and Political Violence*, VOX (Jan. 9, 2021, 11:04 AM), https://www.vox.com/21506029/trump-violence-tweets-racist-hate-speech [https://perma.cc/VD56-YED7]; Kim Wright, *Blocking the President*, HARV. L. TODAY (Jan. 13, 2021), https://today.law.harvard.edu/blocking-the-president/ [https://perma.cc/K7AW-7SKJ] (quoting Evelyn Douek: "[T]he actual content the president posted last week was not materially different from content he has posted before. Most famously, Trump's accounts survived posting that 'when the looting starts, the shooting starts' during the summer's Black Lives Matter protests. By contrast, the tweets that Twitter cited as leading to its decision were fairly anodyne."); Morrison, *supra* note 2 ("His posts, on their face, were actually fairly tame

Trump incited the insurrection not only through his speech at the Ellipse that morning, but also indirectly through repeated dangerous speech that increased the risk of future violence.[54] After the violence had already taken place, Twitter scrambled to de-platform President Trump in the face of evidence that his post-insurrection statements were inspiring plans of future violence on both Twitter and other platforms.[55] The best argument Twitter could muster within the confines of its framework was to cite two relatively innocuous Tweets as violations of the rules against glorifying violence.[56] To support permanent suspension, Twitter cited evidence that the Tweets were being interpreted on and off Twitter as endorsements of future violence.[57] However, the framework does not clearly contemplate Twitter having the ability to consider off-platform interpretations, events, or statements in assessing the risk of violence.[58] Yet as many have pointed out, what likely cemented platforms' decisions to suspend Trump "was not the content Trump posted . . . but the events at the Capitol, his speeches through other media, how people responded on other social media platforms . . . and the extent to which this made Trump's [posts] . . . function as dog whistles to his supporters in a volatile environment."[59]

This example demonstrates that Twitter must revise the framework to better assess the risk of real-world violence following framework-eligible leaders' statements. The vast majority of dangerous speech uses language that is inflammatory and harmful, but ambiguous; many Tweets that encourage, endorse, or threaten violence would likely be reported for violating the Twitter Rules and policies against glorification or incitement, but they might not constitute a "clear and direct threat against an individual" that sidesteps public interest analysis entirely or a "declarative call to action that could harm a specific

---

by Trump standards. But the context around them — as well as the possibility that he would use their platforms to incite more violence — was what Twitter and Facebook took into account when making their decision to deplatform Trump."); Evelyn Douek, *Trump Is Banned. Who Is Next?*, ATLANTIC (Jan. 9, 2021), https://www.theatlantic.com/ideas/archive/2021/01/trump-is-banned-who-is-next/617622/ [https://perma.cc/A8AU-Y4Y9].

54. For example, Twitter took no enforcement action on the following President Trump's Tweet on December 19, 2020, which stated: "Big protest in D.C. on January 6th. Be there, will be wild!" Donald J. Trump (@realDonaldTrump), TWITTER (Dec. 19, 2020, 1:42 AM) (text available in Benesch, *supra* note 52). Though the Tweet does not constitute a declarative call to violence or a violent threat, many Trump supporters quickly posted reactions to the Tweet like the following reaction posted on the now-defunct pro-Trump forum TheDonald.win: "We've got marching orders, bois." Benesch, *supra* note 52.

55. *See* Conger & Isaac, *supra* note 27.

56. Twitter, *Permanent Suspension*, *supra* note 4.

57. *Id.*

58. *Compare id.* (showing the text of the two Tweets cited as the basis of the decision), *with Twitter General Guidelines, Glorification of Violence Policy*, *supra* note 23. *See generally* Wright, *supra* note 53.

59. Evelyn Douek, *The Facebook Oversight Board Should Review Trump's Suspension*, LAWFARE (Jan. 11, 2021), https://www.lawfareblog.com/facebook-oversight-board-should-review-trumps-suspension [https://perma.cc/WUS7-G5RJ].

individual or group" that is "more likely" to be removed without applying a notice.[60] Given its focus on face-value language without explicitly addressing the possibility of considering off-platform interpretation and conditions in world leaders' home countries in assessing the risk of harm, the framework will continue to underestimate the extent to which framework-eligible leaders' Tweets cumulatively increase the risk of violence.

Thus, in proposing changes, this Note begins from the premise that Twitter can and should more robustly consider off-platform social and political context, leaders' off-platform statements, popular interpretations of leaders' Tweets, and leaders' history of violence-related rule violations when considering whether to remove a potentially violence-inciting Tweet or suspend a framework-eligible account. In order to improve transparency and more accurately assess the risk of violence, Twitter should (1) formally codify its existing practice of considering off-platform context and popular on-platform interpretations when evaluating world leaders' Tweets under the framework, (2) invest in more robust analytics able to interpret popular interpretations on- and off-platform, and invest in additional staffers with knowledge of local political and social contexts responsible for proactively monitoring all Tweets by certain framework-eligible leaders, and (3) announce that framework-eligible accounts, just like all other accounts, may be permanently suspended if they exhibit a pattern of "repeated[]" or "particularly egregious" violations of violence-related rules,[61] especially if Twitter obtains sufficient evidence that any further violation would carry a significant risk of causing future real-world violence. This approach would mend the current framework's incompatibility with the Trump suspension statement and inability to sufficiently protect against leaders' capacity to inspire violence over time through indirect language.[62]

---

60. *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

61. *See Our Range of Enforcement Options*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/enforcement-options [https://perma.cc/NT2Q-DVEW] [hereinafter *Our Range of Enforcement Options*].

62. In addition, limiting these proposed changes to *framework-eligible* leaders reduces scalability issues in investing additional resources in contextual review. Diving deeper into off-platform context and on-platform interpretations is more feasible, both technologically and financially, when the pool of accounts subject to this review is limited to elected or governmental officials with at least 100,000 followers. *See* The Lawfare Podcast, *Zittrain on the Great Deplatforming*, LAWFARE (Jan. 14, 2021), https://www.lawfareblog.com/lawfare-podcast-jonathan-zittrain-great-deplatforming [https://perma.cc/8EZP-Q63U] (noting that scalability arguments against context-aware moderation are weaker when that form of moderation is limited to, for example, accounts of presidents; platforms would invest more resources into a limited number of accounts, and the limited number makes context easier to evaluate at scale); *cf.* Benesch, *supra* note 52 ("Heads of state are an obvious place to start [in applying her proposed system of software that monitors real-time interpretations of violent speech], and like many other new content moderation policies in the past, this one should be tried first as a small experiment, with a short list of possible inciters."). *See generally*

1. Memorialize and Exercise the Ability to Consider Off-Platform Circumstances and Leader Statements in Initial "Public Interest" Review of Framework-Eligible Leaders' Reported Tweets

First, Twitter should not only consider the off-platform factors discussed above when gauging risk of harm under the framework's balancing test, but also memorialize this ability in the violent threats, glorification, and incitement sections of the general guidelines' public interest exception provision. Currently, the framework does not explicitly refer to this ability, despite Twitter's emphasis on off-platform factors in the Trump suspension statement.[63] The public interest exception provision of the guidelines mentions Twitter's ability to consider recommendations from internal "in-market teams with an understanding of the cultural context in which the Tweet was posted" and explains that Twitter "will especially err on the side of removal in cases where there is evidence the content may be leading to actual or likely offline harm."[64] But it does not explicitly acknowledge the senior leaders of the Trust & Safety Team's ability to consider off-platform local conditions beyond whichever conditions the "in-market teams" mention in their recommendations in making the final decision whether to remove a Tweet, or to evaluate *off-platform* "evidence the content may be leading to actual or likely offline harm" and factor this into final enforcement decisions.[65]

In the interest of providing transparency and notice to leaders and their followers about the types of off-platform context that might impact a decision to remove or apply a notice to a leader's rule-violative Tweet, Twitter should clarify these processes and policies. Otherwise, as with the Trump suspension, Twitter will be confined to rely on pretextual rules-based justifications when takedown-without-notice or suspension decisions were in fact at least partly the result of off-platform circumstances that demonstrated an unacceptably high risk of real-world violence absent that takedown decision.[66]

TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 77 (2018) (outlining the basic "three imperfect solutions to the problem of scale": editorial review, community flagging, and automatic detection).

63. *Compare Twitter General Guidelines, Public-Interest Exception*, *supra* note 1, *with* Twitter, *Permanent Suspension*, *supra* note 4.

64. *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

65. *Id.*

66. *Cf.* Danielle Citron, *The Case for Trump's Permanent Ban from Social Media*, SLATE (Feb. 5, 2021, 12:03 PM), https://slate.com/technology/2021/02/facebook-oversight-board-trump-ban-vote.html [https://perma.cc/RQ65-4KLJ] (making a similar argument in the context of Facebook's initial "indefinite" suspension of Trump).

2. Invest in Analysis of Interpretations of Leaders' Statements on and off Twitter, and Invest in Additional Staff Knowledgeable of Local Contexts Responsible for Monitoring Framework-Eligible Leaders

Further investing in two key areas would likely improve Twitter's capacity to assess on-platform and off-platform context when reviewing framework-eligible accounts for potential violations of the rules against glorifying or threatening violence. First, as alluded to in its Trump suspension statement, Twitter already has some technological capability to analyze the way users interpret leaders' statements on and off Twitter.[67] Twitter should build upon internal software that can monitor the Tweets of all framework-eligible leaders and detect trends in user interpretations in order to protect against the possibility of violence.

What matters, as Susan Benesch points out, in evaluating the likelihood of real-world violence following leaders' posts is not leaders' intent — which is impossible to determine — but the real-time interpretations of their statements.[68] If Twitter's internal analysis software were to detect, for example, a sudden significant shift in followers' comments and engagement, and/or an above-average surge in user reports flagging a framework-eligible leader's Tweet(s), staffers could then manually review the posts to assess whether a sufficient critical mass of users interpreted the leader's statement(s) as endorsing or ordering violence.[69] This could be a joint effort between reviewers with expertise in analytics and reviewers with expertise in specific country conditions. Intentionally coded or ambiguous speech may be very analytically difficult to detect.[70] But whenever a framework-eligible leader Tweets or inspires such speech, combining analytic detection of strong shifts in user responses with subsequent country-context-aware human review would help mitigate this content moderation scalability and detection challenge by limiting the universe of Tweets to framework-eligible accounts, which have a particularly strong ability to incite violence given their large digital megaphones.[71]

Second, in addition to investing in software to detect shifts in patterns of interpretation following framework-eligible leaders' Tweets,

---

67. Twitter, *Permanent Suspension*, *supra* note 4. As Susan Benesch notes: "The company could build software to monitor [large numbers of] accounts and their followers' reactions to them, looking for significant shifts in the sentiment of the followers' comments or posts, and signs that a critical mass of followers understand the political figure to be endorsing or ordering violence." Benesch, *supra* note 52.

68. Benesch, *supra* note 52.

69. *Id.*

70. *Cf.* Jackson & Minow, *supra* note 38 ("Specific contextual knowledge is of special importance in evaluating coded speech, which may be used to communicate messages designed to exclude understanding by outsiders.").

71. *See id.*

and hiring additional reviewers with knowledge of local contexts to assess those shifts and assist framework analysis, Twitter should also adopt the approach David Kaye proposes to more effectively consider local social and political contexts: hiring additional staffers responsible for proactively monitoring the accounts of framework-eligible leaders with a history of condoning violence on and off Twitter, and/or of those who lead countries experiencing conditions ripe for political violence.[72] These staffers would not merely be tasked with offering recommendations on individual pieces of content upon receiving user reports, based on their knowledge of the applicable cultural, social, or political contexts.[73] Rather, this new type of staffer would monitor the Tweets and off-platform statements of certain framework-eligible leaders hailing from the staffer's country or countries of expertise; track the evolving country-specific political and social conditions; and internally flag those leaders' histories of using violent rhetoric or participating in violence on- and off-Twitter, as well as the frequency of violence-related user reports of their Tweets. This would be an expansion upon the size and scope of the task entrusted to Twitter's current "in-market teams"; these new staffers would provide insight into potential temporary or permanent account suspension decisions and advise on individual content decisions under the framework whenever a framework-eligible Tweet that may inspire violence is detected by the aforementioned interpretation analysis or through user reports.[74]

These staffers would supplement Twitter's existing capacity to monitor conversations occurring on other platforms reacting to a leader's Tweets, especially conversations that take place over a longer period of time, when deciding whether a temporary or permanent suspension may be justified. Twitter apparently had some capacity to detect and evaluate posts indicating plans — inspired by Trump's offline and Tweeted statements — to organize future violent protests on platforms other than Twitter.[75] To the extent technologically and financially feasible, Twitter can and should also consider monitoring

72. Phillips et al., *supra* note 6 (quoting David Kaye). Countries "most prone" to political violence might be defined, for example, as those that have experienced a certain number of political violence incidents with a sufficient nexus to incendiary statements by political leaders over the past five years.

73. *Id.*

74. Preparing ongoing dossiers on framework-eligible leaders, especially those determined particularly incendiary or whose countries are particularly ripe for violence, could also save substantial time and effort of the teams who review reported Tweets and apply the public interest exception analysis. These dossiers could be produced partly by web-crawling AI and partly by these expert human reviewers assigned to monitor particular leaders. They could identify context so that when a questionable/troubling threat-like Tweet emerges, much of the context analysis work will already have been completed.

75. Twitter, *Permanent Suspension*, *supra* note 4 ("After close review of recent Tweets from the @realDonaldTrump account and the context around them — specifically how they are being received and interpreted *on and off Twitter* — we have permanently suspended [Trump] . . . .") (emphasis added).

conversations relating to framework-eligible leaders' Tweets taking place on other platforms.

3. Articulate and Apply a Context-Aware Standard for Suspending Framework-Eligible Leaders for Repeated or Egregious Violations of Violence-Related Twitter Rules

As Benesch noted in the context of the insurrection, in general, "rioting crowds must be primed for violence. No one would smash their way into a building on the basis of only one rant, no matter how convincing."[76] Trump successfully "primed" his followers through the combined effect of spreading misinformation and conveying an "us versus them" narrative, in addition to his repeated usage over time of ambiguously violent rhetoric.[77] Given the size of his audience, many users likely reported Trump's Tweets, only for Twitter to determine that the Tweet(s) did not violate any rules or to decide to leave the Tweet(s) online behind a notice under the public interest exception.[78] Yet Twitter declined to permanently suspend Trump until the aftermath of the insurrection.

The fact that Twitter based its decision to suspend Trump on its conclusion that two specific Tweets ostensibly violated the rules against glorifying violence, but also mentioned that the Tweets "must be read in the context of broader events in the country and the ways in which the President's statements can be mobilized by different audiences, including to incite violence, as well as in the context of the pattern of behavior from this account in recent weeks," both illustrates and fails to clarify a significant gap in the framework: it does not address whether, and under what circumstances, a framework-eligible leader may face suspension as an enforcement consequence.[79] By its terms, the framework — as codified in the 2019 World Leaders Policy Statement and the public interest provision of the General Guidelines — discusses Twitter's policies on reviewing and deciding whether to remove individual Tweets from world leaders.[80] But unlike Twitter's general, platform-wide policy on when and how it employs other enforcement actions like permanent suspension, the public interest framework does not address the possibility of temporary or permanent account suspension whatsoever.[81] Twitter's Trump suspension decision and accompanying explanation and the silence of the public interest framework on

76. Benesch, *supra* note 52.

77. *See id.*; *see also* Cineas, *supra* note 53.

78. *See* Twitter, *Permanent Suspension*, *supra* note 4.

79. *Id.*

80. *See Twitter, 2019 World Leaders Policy Statement*, *supra* note 1; *Twitter General Guidelines, Public-Interest Exception*, *supra* note 1.

81. *See Our Range of Enforcement Options*, *supra* note 61.

suspension suggest that Twitter has given itself discretion to similarly suspend a world leader in the future. But the statement justifying the Trump suspension — which cited offline contextual factors as justifying the suspension without referencing any underlying provision of the framework authorizing its ability to do so — leaves it unclear how, exactly, Twitter's suspension decision process for world leaders differs from the standard offline-context-blind suspension decision process outlined in its general guidelines.

Without any clear guidance on whether, and how, Twitter will decide to suspend framework-eligible leaders on the basis of a pattern of violence-related rule violations combined with surrounding offline contexts, leaders will be left guessing whether they might become the next high-profile politician permanently suspended on violence-related grounds. And without delegating itself the explicit ability to consider repeat offenses in light of additional off-platform circumstances indicating a high risk of real-world violence if the account remains undisturbed, Twitter will likely fail to prevent violence that occurs following repeated low-level violations of the rules against glorification or threats. In other words, when each "particular drop of petrol" Tweeted is not "actionable" for removal,[82] the current framework cannot justify suspending leaders who incite violence indirectly over time until a "fire" has already started. Therefore, in order to serve the dual goals of increasing transparency and predictability, while preventing real-world violence that arises after a leader's patterns of violent rhetoric that are not individually sufficient to justify suspension, Twitter should articulate and apply a framework that considers framework-eligible leaders' repeat violations of the rules against glorification or threats and real-world country conditions as a potential basis for suspension.

C. The Case of T. Raja Singh: How More Context-Aware Moderation Would Operate

The example of T. Raja Singh illustrates the shortcomings of Twitter's current approach toward potentially violence-inspiring leaders and can show how the contextual considerations proposed *supra* Sections III.B.1–3 might operate. Singh, whose Twitter account is framework-eligible,[83] has an extensive history of making hateful and violent remarks about specific groups, especially Muslims, sometimes explicitly calling for violence against them in statements off Twitter.[84] But

82. Benesch, *supra* note 52.

83. *See* Raja Singh (@TigerRajaSingh), TWITTER (April 23, 2021), https://twitter.com/TigerRajaSingh?s=20 [https://perma.cc/3U3K-HQYF]; Twitter Safety, *Defining Public Interest on Twitter*, *supra* note 1.

84. *See, e.g.*, *Banned by Facebook Now, T Raja Singh Has 60 Cases Against Him: 7 Times His Hate Speeches Hit Headlines*, NEWS18: POLITICS (Sept. 3, 2020, 8:26 PM),

Singh remains active on Twitter, and many of his controversial Tweets do not clearly constitute direct threats of violence against particular individuals or even groups.

Some of Singh's Tweets and their off-platform contexts illustrate the urgency and likely efficacy of changes like those advocated in Section III.B in more effectively balancing safety against public interest. First, the incoherence of Twitter's justification for its decision to take enforcement action on a three-year-old Tweet by Singh in 2020 highlights its need to refine terms of the framework to acknowledge its nuances and admit the relevance of off-platform context. After journalists flagged and reported on a three-year-old Tweet in which Singh advocated the deportation of Rohingya immigrants "who supported terrorism," Twitter placed the Tweet behind a notice and justified the decision by claiming Twitter has "zero tolerance policies" toward violent threats and hateful conduct.[85] But this response is unconvincing, pretextual, and misstates the framework.[86] Twitter could have instead justified the decision by noting that although the statement arguably violates the hateful conduct rules,[87] under public interest analysis, Twitter decided to keep the post up behind a notice because of the exception's stated leniency toward leaders' "comments on political issues

---

https://www.news18.com/news/politics/banned-by-facebook-now-t-raja-singh-has-60-cases-against-him-7-times-his-hate-speeches-hit-headlines-2845477.html [https://perma.cc/P2JU-D38M]. On Facebook and in offline speeches, in addition his comments on Facebook in 2018 discussed *supra* Part II, Singh has threatened to raze mosques, Purnell & Horwitz, *supra* note 50, stated that those who refuse to worship cows (i.e., non-Hindus) should "have their throats slit with a sword," and described his efforts to form a vigilante army to hunt down "traitors." Lauren Frayer, *Facebook Accused of Violating Its Hate Speech Policy in India*, NPR: ALL THINGS CONSIDERED (Nov. 27, 2020, 3:46 PM), https://www.npr.org/2020/11/27/939532326/facebook-accused-of-violating-its-hate-speech-policy-in-india [https://perma.cc/3ST9-UASK].

In 2020, Facebook permanently banned Singh after domestic political pressure and public outrage snowballed following a Wall Street Journal article reporting that Ankhi Das — Facebook's top policy executive in India, who resigned after the Wall Street Journal controversy — kept Singh's profile up and advocated for leniency in moderating his posts, even though he had been flagged internally for promoting and participating in violence, because she did not want to risk hurting relations with Modi's BJP. Russell Brandom, *Facebook India's Controversial Policy Chief Has Resigned*, VERGE (Oct. 27, 2020), https://www.theverge.com/2020/10/27/21536149/ankhi-das-facebook-india-resigned-quit-bjp-hindu-muslim-conflict [https://perma.cc/85Y8-EUPY].

85. Manish Singh, *Twitter Flags Indian Politician's Years-Old Tweet for Violating its Policy*, TECHCRUNCH (Sept. 15, 2020, 6:30 PM), https://techcrunch.com/2020/09/15/twitter-flags-indian-politicians-years-old-tweet-for-violating-its-policy/ [https://perma.cc/PV93-KD9T].

86. *See Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

87. *See Rules and Policies: Hateful Conduct Policy*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [https://perma.cc/8RHZ-E48M] (prohibiting "incitement against protected categories," such as "to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., 'all [religious group] are terrorists'").

of the day."[88] More importantly, if the framework explicitly allowed consideration of off-platform conditions and on- and off-platform interpretations as Section III.B.2 proposes, Twitter would have been able to remove the Tweet if it had detected that it was being interpreted as condoning using lethal force against Rohingyas following Singh's aforementioned 2018 Facebook video statement, despite the language of the 2017 deportation Tweet not constituting a "clear and direct threat." Twitter also could have potentially justified a decision to suspend his account, invoking a new provision on suspension decisions under the framework as Section III.B.3 proposes, given Singh's history of repeated and/or egregious violence-related rule violations in light of offline conditions in India at the time.

Next, a video Singh Tweeted in July 2017[89] highlights Twitter's need to invest in automated analysis of framework-eligible leaders' Tweets in conjunction with staffers who review auto-detected engagement shifts and proactively monitor certain leaders' accounts in order to fully appreciate the incitement potential of Tweets. In the Tweet, Singh commented on then-ongoing communal riots in West Bengal,[90] which erupted into violence between Hindus and Muslims.[91] In the video, Singh alludes to "what happened in 2002" and encourages Hindus in the region to "respond in the same way."[92] If the revised framework proposed *supra* Sections III.B.1–3 had applied at the time, India-expert and Singh-specific moderators would be able to recognize Singh's Tweet as an invocation of the 2002 anti-Muslim Gujarat riots,[93] constituting a "declarative call to action that could harm a specific individual or group" that is "more likely" to be removed after public interest analysis.[94] It does not directly threaten a specific party, but

88. *Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

89. Raja Singh (@TigerRajaSingh), TWITTER (July 7, 2017, 11:00 AM), https://twitter.com/TigerRajaSingh/status/883340012321984514 [https://perma.cc/9QMB-QV4D]. For a Hindi-to-English translation of key points Singh states in this Tweeted video, and a summary of the context of the communal riots Singh mentions in the video and its accompanying text, see TNM Staff, *Bengal violence: Hyderabad BJP MLA Raja Singh Asks Hindus to Respond like 2002 in Gujarat*, NEWS MINUTE (July 9, 2017, 8:18 PM), https://www.thenewsminute.com/article/bengal-violence-hyderabad-bjp-mla-raja-singh-asks-hindus-respond-2002-gujarat-64895 [https://perma.cc/NMB9-AMKV].

90. *See* TNM Staff, *supra* note 89.

91. *See* Shoaib Daniyal, *A Facebook Post Was All It Took to Undo Decades of Communal Harmony in a Small East Indian Town*, QUARTZ: INDIA (July 17, 2017), https://qz.com/india/1030653/west-bengal-violence-how-a-facebook-post-broke-the-decades-long-communal-peace-of-a-west-bengal-town/ [https://perma.cc/9L5H-VSZF].

92. TNM Staff, *supra* note 89. Singh also addresses his followers directly in the video, stating: "My brothers, you remember what happened in 2002 in Gujarat, when Hindus were killed. The way that the Hindus responded in Gujarat, today, there is a need for Hindus in Bengal, to respond the same way." *Id.*

93. *See generally India: A Decade on, Gujarat Justice Incomplete*, HUM. RTS. WATCH: NEWS (Feb. 12, 2012, 4:22 PM), https://www.hrw.org/news/2012/02/24/india-decade-gujarat-justice-incomplete# [https://perma.cc/F52D-A8PP].

94. *See Twitter, 2019 World Leaders Policy Statement*, *supra* note 1.

probably violates the rule against violent threats. The proposed staff force charged with collecting dossiers on world leaders could advise the Safety Team engaging in the public interest exception decision on the post accordingly.[95] Moreover, these moderators could consider Singh's history of anti-Muslim comments in speeches in the months preceding the Tweet.[96] A deeper understanding of the context of the ongoing Bengal riots at the time, the invocation of the Gujarat riots by Singh's Tweet, and Singh's recent off-platform statements promoting violence against Muslims, would allow context-aware staffers to advise the Safety Team, as proposed in Section III.B.3, that this Tweet merits temporary or permanent suspension of Singh's account at most, or a notice under the exception at least.

Ultimately, this example demonstrates that unless Twitter revises its policies and practices on moderating framework-eligible leaders, it will continue to underestimate their potential to incite real-world violence. Without revision to Twitter's policies, even leaders like Singh with histories of issuing direct threats of or calls to violence off-Twitter will continue to add fuel to the possibility of violence, so long as their Tweets are phrased ambiguously enough to evade scrutiny from algorithmic detection or reviewers without deep contextual knowledge.

IV. CONCLUSION

The changes to Twitter's framework that this Note proposes, particularly the highly context-aware approach to moderating framework-eligible leaders' dangerous but indirect statements, would undoubtedly create new content moderation problems. Such a context-aware approach potentially enables Twitter to over-remove leaders' online speech in a manner that undervalues the benefits of preserving the public's access to leaders' Tweets. Making content moderation judgments that favor safety over speech in certain circumstances is well within Twitter's legal right as a private actor. However, it invites familiar

95. This would apply whether they came across the Tweet in the process of proactively monitoring Singh's account given his history of using violent rhetoric and India's recent history of communal riots, or due to an algorithmically detected surge in user reports for violating the violence-related rules or a surge in interpretations understanding the Tweet to encourage Hindu followers to commit violence against Muslims in the region.

96. For example, in May 2017, Singh called Kashmiri Muslims "traitors" in a speech. *See* TNM Staff, *8 Vile Comments by BJP's Raja Singh in 2017, and He Isn't Stopping*, NEWS MINUTE (Nov. 24, 2017, 4:15 PM), https://www.thenewsminute.com/article/8-vile-comments-bjp-s-raja-singh-2017-and-he-isn-t-stopping-72124 [https://perma.cc/7RSA-9CS9]. In April 2017, he threatened to behead anyone opposed to building a Hindu temple on the site of a destroyed Muslim mosque, which inspired a Muslim advocacy group to file a complaint with authorities based on the country's anti-hate speech laws. *See Ayodhya: India Politician Threatens to Behead Temple Opponent*, BBC NEWS (Apr. 10, 2017), https://www.bbc.com/news/world-asia-india-39552154 [https://perma.cc/R73A-DYTP].

concerns over the vast scope of tech platforms' discretionary power to police global online speech, which critics ranging from Angela Merkel to Singh and Trump's political allies have recently expressed.[97] These concerns often center on the relationship between increased content moderation discretion and the risk that platforms like Twitter will selectively enforce policies more frequently and severely against politicians with whom Twitter does not have a close business relationship or financial ties, rather than engaging in a neutral, nonpartisan, good-faith balancing test between the risk of harm on one hand, and public interest and free speech benefits on the other.[98]

However, the trend of incendiaries like Donald Trump, T. Raja Singh, Rodrigo Duterte, and Jair Bolsonaro using Twitter to promote violence[99] demonstrates that even though a context-aware approach to world leaders involves some risk of over-removal or inconsistent enforcement influenced by business interests and governmental ties, this risk is worth incurring in the context of framework-eligible leaders, particularly in countries with demonstrated recent history of real-world violence inspired by online speech. Moreover, in the process of adopting a revised framework like the one this Note recommends, Twitter can mitigate accountability concerns by articulating the standards and guidelines it will follow as transparently as possible while retaining

---

97. *See, e.g.*, Langvardt, *supra* note 8; Associated Press, *Germany's Merkel: Trump's Twitter Eviction "Problematic"*, AP NEWS (Jan. 11, 2021), https://apnews.com/article/merkel-trump-twitter-problematic-dc9732268493a8ac337e03159f0dc1c9 [https://perma.cc/9YDS-3ZHG]; Lexi Lonas, *Pompeo, Cruz and Other Trump Allies Condemn Twitter's Ban on President*, HILL (Jan. 9, 2021, 1:06 PM), https://thehill.com/policy/technology/533486-pompeo-cruz-and-other-trump-allies-condemn-twitters-ban-on-president [https://perma.cc/G74Q-2KQK]; Frayer, *supra* note 84 (interviewing former Modi staffer Arvind Gupta, cautioning against social media giants' power to "censor" online speech: "Are [social media companies] for free speech? Or are they going to be gatekeepers of what is right or wrong? Today you exclude somebody for an opinion you don't like; tomorrow it would be on [another] basis . . . . Facebook is allowing one or two people to dictate norms.").

98. The Wall Street Journal report that Facebook's India policy chief promoted light-handed moderation of T. Raja Singh and resisted designating him as a "dangerous individual" for fear of damaging Facebook's relationship with the ruling BJP in Facebook's biggest market provides an illustrative example of the danger of discretion when government relations and business interests creep into content moderation decisions. *See generally* Purnell & Horwitz, *supra* note 50; Purnell & Roy, *supra* note 50.

For examples of common concerns in the U.S. over tech platforms' wide power and discretion to moderate online speech and proposed solutions, see, e.g., Langvardt, *supra* note 8; Edward Lee, *Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance*, 70 AM. U. L. REV. 913 (2021) (proposing a "nonpartisan content moderation" framework that platforms should voluntarily adopt in order to improve transparency and avoid messy government entanglement and allegations of partisan bias, presented as a better alternative to reforming Section 230).

99. *See, e.g.*, Phillips et al., *supra* note 6; Douek, *supra* note 59 (discussing Trump's suspension); Elizabeth Dwoskin et al., *Zuckerberg Once Wanted to Sanction Trump. Then Facebook Wrote Rules that Accommodated Him*, WASH. POST (June 28, 2020), https://www.washingtonpost.com/technology/2020/06/28/facebook-zuckerberg-trump-hate/ [https://perma.cc/48DX-EEPM].

flexibility to adapt in response to issues that may well arise in the future.

Finally, if it articulates and enforces grounds for permanently suspending framework-eligible accounts, Twitter should consider the possibility of establishing an independent entity similar to the Facebook Oversight Board endowed with the authority to review any decisions to permanently suspend framework-eligible leaders. Given the strong public interest implications of banning a world leader from a major platform on which they communicate with constituents, such a review system would ensure that such high-stakes unilateral decisions by the platform are not unreviewable or unaccountable.[100] Moreover, a reviewing entity could be empowered to provide Twitter with guidance on how to respond to challenges that arise in considering political and social off-platform context and balancing safety against public interest benefits, as with the Facebook Oversight Board in the context of its decision on suspending Trump.[101]

In sum, Twitter must begin to take steps to improve the shortcomings of its public interest exception; the current framework woefully underestimated the risk of harm that Trump's Tweets posed, and other world leaders will certainly use Twitter to directly or indirectly incite violence in the future. The Trump decision signaled Twitter's apparent capacity to take off-platform context into account, even though it failed to prevent the violence of the insurrection. Moving forward, Twitter must clarify and expand upon these capacities and invest in more context-aware review of a defined class of the most influential government leaders for possible incitement of violence — while swiftly removing clear-cut violence incitement where the risk of real-world harm to named individuals is most imminent.

---

100. *See generally* Douek, *supra* note 59; Evelyn Douek, *Facebook's Oversight Board: Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & TECH. 1 (2019).

101. *See* Douek, *supra* note 100; Heath, *supra* note 33; Douek, *supra* note 33; *Oversight Board Upholds Former President Trump's Suspension, Finds Facebook Failed to Impose Proper Penalty*, OVERSIGHT BD.: NEWS (May 5, 2021), https://oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/ [https://perma.cc/H3A9-D8DM] ("Although Facebook explained that it did not apply its 'newsworthiness' allowance in this case, the Board called on Facebook to address widespread confusion about how decisions relating to influential users are made. . . . Facebook should publicly explain the rules that it uses when it imposes account-level sanctions against influential users."). The Board further proposed that Facebook, "[r]apidly escalate content containing political speech from highly influential users to specialized staff who are familiar with the linguistic and political context [and] insulated from political and economic interference, as well as undue influence." *Id.*