Free markets. Real solutions.

R STREET POLICY STUDY NO. 224
March 2021

DONALD TRUMP AND THE FACEBOOK OVERSIGHT BOARD

By Paul Rosenzweig, Chris Riley, Mary Brooks and Tatyana Bolton

CONTENTS

INTRODUCTION

The rise of social media has generated controversies over how these platforms moderate—and fail to moderate—the content posted on their sites. Some observers argue that the social media giants (like Facebook, Twitter and Google) do too little to take down potentially harmful content, while others claim they do too much, criticizing an excess of censorship. Still others argue that whatever these platforms do lacks transparency and is—in any event—too self-serving to be valid or trustworthy.

In response to these concerns, Facebook has been working toward the creation of a new model of social media governance: An Oversight Board (Board), which has the authority to review some (though not all) of Facebook's content moderation decisions.[1] Sometimes colloquially referred to as Facebook's "Supreme Court," the creation of the Board has not been without its own controversy.[2] Some immediately denounced it as ineffective; others thought it a shill for Facebook's self-interest. Still others were willing to suspend judgment pending implementation.

Now, only months after the Board was fully established in late 2020, it is facing its first serious test of legitimacy. In early January 2021, during and immediately following the insurrection at the U.S. Capitol, Facebook removed certain content posted by then-President Donald J. Trump and indefinitely suspended his account.[3] Facebook later referred its decision to deplatform Trump—essentially revoking his access to the social media outlet—to the Board for official review. Thus, rather than having time to work with lower-profile matters to develop its doctrine and procedures, the Board is now faced with a momentous and potentially controversial decision as one of its very first cases. Some have called this the Board's "Marbury moment"—a reference to the seminal American case, *Marbury v. Madison*, in which the U.S. Supreme Court asserted the absolute role of judicial oversight in the American system of government.[4]

1. "Shaping the Board," Oversight Board, 2020. https://oversightboard.com.

2. Kate Klonick, "Inside the Making of Facebook's Supreme Court," *The New Yorker*, Feb. 12, 2021. https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court.

3. Nick Clegg, "Referring Former President Trump's Suspension From Facebook to the Oversight Board," Facebook, Jan. 21, 2021. https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board.

4. Nate Persily (@persily), "I actually think this will be the Oversight Board's Marbury v Madison moment—meaning, even if they uphold the decision to suspend, the way they handle the case, decide on their jurisdiction, and consider the breadth of the issue presented will be important going forward." Jan 21, 2021, 12:35 PM. Tweet. https://twitter.com/persily/status/1352308920031666177.

Whether the Trump deplatforming decision proves to be quite so consequential remains to be seen. But it does appear that the case will provide the Board with the opportunity both to establish its own authority and to develop a doctrine of review that would contribute to a transparent and trustworthy oversight process.

To assist in that development, the R Street Institute recently submitted comments to the Board in response to its request for public comments in the case reviewing Facebook's decision to deplatform Trump.[5]

In those comments, R Street argued there are three principles that must be central components of the Board's decision-making, and which should be formalized into the Board's ongoing review:

1. Context—and therefore case-by-case review—is critical to a valid and appropriate adjudication of the issues presented.
2. A well-structured framework is necessary to particularize the considerations applicable to content moderation decisions.
3. The same framework should be applied to private citizens and political actors, with due regard for the different context in which their expression arises.

While the Board's first round of decisions indicate that it will heavily rely on context in its case evaluation, it has not yet explored the second and third principles above. This paper will explain how the Board should work to close that gap through the consistent use of an explicit, multi-factor framework that is guided by international human rights law but offers more granularity than the high-level principles contained in that body of law. Additionally, the Board should not create separate standards of content moderation for politicians, but should apply the same context-based framework for all users.

Facebook, the Board, and social media platforms more broadly are at an inflection point. The decisions they make now may define the future of digital platform-based communication, with real outcomes for the security and integrity of the internet and those who use it. This white paper suggests how these decisions can be made with recourse to external law and principles, rather than in an *ad hoc* manner.

THE NEED FOR A STANDARDIZED CONTENT MODERATION PROCESS

Currently, there is no universally accepted content moderation framework guiding Facebook's practices or the Board's subsequent analysis. Facebook is a private corporation, not a government, and is free to adopt any legal content moderation standards and practices that suit its community and culture.

It is apparent, however, that Facebook no longer sees some of this freedom and autonomy as desirable. Whether this is a result of backlash from the public, pressure from human rights activists, threats by governments to formally regulate social media, or even out of a desire to be a good actor—or, at least, perceived as one—in the space, Facebook has now explicitly bound itself to an external standard by establishing the Board as a quasi-legal, independent oversight entity. Thus, while Facebook is not obligated to follow a recognized or even formalized set of legal standards for its content moderation, it is choosing to do so on its own terms.

This decision to delegate some of its autonomy is novel. However, the ramifications of this decision in the long term remain unknown, and the extent of Facebook's willingness to abide by external Board decisions has not been strenuously tested, though Facebook has already rejected a recommendation from the Board in at least one case.[6]

One aspect that is immediately clear, however, is that it would be untenable for Facebook to simultaneously retain total freedom to make its own decisions while being bound under this new system of rules. In other words, the decision-making process of the Board must have transparency, consistency and legality. To that end, it is imperative the Oversight Board makes transparent decisions based on consistent rules, values and a systematic framework.

To at least some extent, the Board has realized this necessity and has stated that it intends to govern by case law:

> For each decision, any prior board decisions will have precedential value and should be viewed as highly persuasive when the facts, applicable policies, or other factors are substantially similar.[7]

In making these decisions, the Board's charter also states that it will be guided, first and foremost, by Facebook's own values:

5. Chris Riley and Paul Rosenzweig, "R Street Comments on Trump Ban to Oversight Board: 'Facebook is Justified,'" R Street Institute, Feb. 7, 2021. https://www.rstreet.org/2021/02/07/case-no-2021-001-fb-fbr-facebook-oversight-board.

6. Nick Clegg, "Facebook's Response to the Oversight Board's First Set of Recommendations," Facebook Newsroom, Feb. 25, 2021. https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations.

7. "Oversight Board Charter," Oversight Board, September 2019, p. 5. https://oversightboard.com/governance.

Facebook has a set of values that guide its content policies and decisions. The board will review content enforcement decisions and determine whether they were consistent with Facebook's content policies and values.[8]

However, Facebook and the Board have also made a determination to be guided by principles of international human rights law (IHRL). The Board is chartered to "review content enforcement decisions and determine whether they were consistent with Facebook's content policies and values," and "will pay particular attention to the impact of removing content in light of human rights norms protecting free expression."[9] The Board's rulebook for case review also notes that the Board may commission "research on case context (e.g. cultural, linguistic, political), relevant international standards on freedom of expression and human rights, and Facebook's content policies and values."[10]

In practice, this means that in addition to assessing conformance with Facebook's own values, the Board will look at international concepts such as legality, necessity and proportionality as part of the framework for its analysis.

IHRL is a reasonable source for the Board's decision-making and its review of Facebook's actions. However, most IHRL is stated at too high a level of generality to be of practical use to the Board. It neither provides specific enough guidance to Facebook nor adequate notice to users as to how content decisions will be made. To that end, this paper strives to particularize the high-minded principles of IHRL by identifying more concrete factors that the Board should consider in undertaking its analysis of the cases that come before it.

INTERNATIONAL HUMAN RIGHTS LAW

As set out in the Universal Declaration of Human Rights (UDHR), freedom of expression is a fundamental human right.[11] It is enshrined in international human rights law, most notably under Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which provides that "everyone shall have the right to hold opinions without interference" and that "everyone shall have the right to freedom of expression [across all mediums]."[12]

Notably, the ICCPR qualifies the Article 19 right to free expression, holding that it may be necessary to restrict expression to "respect…the rights or reputations of others" and "[f]or the protection of national security or of public order…or of public health or morals."[13] Article 20 of the ICCPR separately obligates State Parties to prohibit certain types of expression, namely "any propaganda for war" and "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[14]

Clearly, the right to freedom of expression is not unqualified under IHRL. In fact, the United Nations Human Rights Committee has explicitly upheld the need to allow for curbs on free expression:

> In the opinion of the Committee, these required prohibitions [created under ICCPR article 20] *are fully compatible with the right of freedom of expression as contained in article 19*, the exercise of which carries with it special duties and responsibilities [emphasis added].[15]

Nonetheless, the Committee has also warned that restrictions cannot be made overly broad: "It is the interplay between the principle of freedom of expression and…limitations and restrictions which determines the actual scope of the individual's right."[16] While international law generally binds only nation states, the Office of the United Nations High Commissioner for Human Rights believes that corporations have a responsibility to respect "internationally recognized human rights."[17]

The devil, as they say, is in the details. Of course, in some circumstances one may restrict the freedom of expression in service of a superior right (say in a public health emergency). The challenge is identifying when that sort of emergency or circumstance rises to a level that justifies placing restrictions on the exercise of the right of free expression. On its own, IHRL does not, and cannot, provide an operational framework within which actions may be judged. It is too general and indefinite. While IHRL provides a broad guide, greater specificity is required.

8. Ibid.

9. Ibid.

10. "Rulebook for Case Review and Policy Guidance," Oversight Board, November 2020, p. 9. https://oversightboard.com/sr/rulebook-for-case-review-and-policy-guidance.

11. "Universal Declaration of Human Rights," United Nations General Assembly, Dec. 10, 1948. https://www.un.org/en/universal-declaration-human-rights.

12. Office of the United Nations High Commissioner for Human Rights, "International Covenant on Civil and Political Rights," United Nations, Dec. 16, 1966. https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx.

13. Ibid.

14. Ibid.

15. Office of the United Nations High Commissioner for Human Rights, "General Comment No. 11: Prohibition of propaganda for war and inciting national, racial or religious hatred (Article 20)," United Nations, July 29, 1983. https://www.ohchr.org/Documents/Issues/Opinion/CCPRGeneralCommentNo11.pdf.

16. Office of the United Nations High Commissioner for Human Rights, "General Comment No. 10: Freedom of expression (Art. 19)," United Nations, June 29, 1983. https://www.ohchr.org/Documents/Issues/Opinion/CCPRGeneralCommentNo10.pdf.

17. Office of the United Nations High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," United Nations, 2011, p 13. https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

To that end, as noted earlier, the Board has begun to develop a framework for its analysis that identifies neutral principles and factors on which it will rely for all of its decision-making. In the few cases already decided, it has identified issues of legality, necessity and proportionality as those most directly bearing on its decisions.[18] But even that nascent framework of analysis is too indefinite. Necessity and proportionality are, to some degree, in the eye of the beholder.

What is needed next is an articulation of various subsidiary factors that will play a role in the framework analysis. While multifactor tests—such as the proposal outlined below—are, themselves, subject to ambiguity and even manipulation, experience in both domestic and international legal systems over the past 50 years (for example in the Laws of Armed Conflict) has demonstrated the utility of explicitly identifying relevant factors and issues for consideration in rendering judgment.[19] The exposition of factors and their application to various fact-based scenarios will allow the development of, in effect, a common law of content moderation and, over time, provide greater transparency, rigor and clarity in evaluating Facebook's actions and the Board's responses. The following section outlines the factors that should be considered in such a framework and recommendations for how such a framework should be implemented.

FRAMEWORK FACTORS

Four factors are of particular salience to an assessment of the necessity and legality of any decision reviewed by the Board: The truth or falsity of the content at issue; its harmfulness; the imminence of that harm; and the extent to which the content provider intended or incited perceived or actual harm.

Truth or Falsity

The falsity, *vel non*, of a social media post is relevant to the balance between harm and freedom of expression; for example, Holocaust denialism and vaccine disinformation are exemplars of exceptionally damaging falsehoods. In many contexts assessing the truth or falsity of content is difficult, and in some cases, it may be impossible. Of course, opinions are especially unlikely to be classifiable as true or false. However, the difficulty of evaluation should not obscure the reality that some facts are demonstrably true and denying them is demonstrably false. As it was put recently: "The earth is round. Two plus two equals four…These are facts. They are demonstrable and irrefutable."[20] Thus, it is at least somewhat relevant to any analysis of Facebook's decision-making to ask whether or not the content in question has the character of truthfulness.

Of particular concern should be purposeful, coordinated disinformation. This may be the product of artificial information dissemination or, in more difficult cases, a result of individuals acting in concert or as a group led by a single individual. Whatever the source, the single gravest threat to social media is the proliferation of conspiracy theories—such as "Pizzagate," which lead from falsities propagated in the virtual world to real harm in the physical world—and other deliberate attempts to hijack the social conversation.[21] To be sure, discussions of sensitive topics can add positive value, but the purposeful spread of disinformation adds a particularized harm. Again, there will be gray areas, where Facebook and the Board should err on the side of inclusivity, but there are equally clear black and white areas where deliberate disinformation ought to be considered as part of any content moderation analysis.

Harmfulness

Not all false speech is harmful. Similarly, not all truth is harmless. Indeed, the truth sometimes wounds quite deeply, precisely because it is true. Thus, while truth and falsity are relevant to any consideration of content moderation, they are insufficient, by themselves to justify a decision. Consideration must also be given to the harm that might arise from the content in question—its nature, likelihood and degree.

Assessing harm is not an ontological exercise. First, as a categorical matter, not all harms are cognizable and redressable. Within the context of social media, most speech—even some that is quite hurtful—is not a cognizable harm incompatible with the nature of the platform. Facebook recognizes as much when it prioritizes giving content providers a "voice" subject to consideration of cognizable harms related, for example, to "safety" or "privacy."

Even with those later categories, whether or not particular speech raises significant questions of harm requires assessment on a case-by-case basis. Harm analysis is not normative, it is contextual and descriptive in assessing the weight and likelihood of the harm. For that reason, it is questionable whether the Board can, or should, provide general guidelines about speech by politicians, as Facebook has requested.

18. Jacob Schulz, "What Do the Facebook Oversight Board's First Decisions Actually Say?", *Lawfare*, Jan. 29, 2021. https://www.lawfareblog.com/what-do-facebook-oversight-boards-first-decisions-actually-say.

19. Travis Normand and Jessica Poarch, "Law of Armed Conflict," *LOAC Blog*, 2015. https://loacblog.com/loac-basics/4-basic-principles.

20. News Corps and Daily Kos, "The intro to Smartmatic's lawsuit against Fox News reads like a riveting crime novel," AlterNet, Feb. 5, 2021. https://www.alternet.org/2021/02/smartmatic.

21. Matthew Saag and Maya Salam, "Gunman in 'Pizzagate' Shooting Is Sentenced to 4 Years in Prison," *The New York Times*, June 22, 2017. https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html.

The relevant analysis should begin with the real-world context in which the content arises. This part of the assessment could take into account whether real-world physical injury (e.g., riots or child sex trafficking) is reasonably likely to arise from the content in question. Likewise, it would recognize that non-physical harms—digital or psychological—could result. Though all harms are surely the subject of concern, special care must be taken to avoid adverse consequences that directly or indirectly result in acute harm to individuals—systematic or societal harms are significant, but necessarily more diffuse.

However, the converse of this is also true. Though systematic and societal harms may be more diffuse, they may also be far more harmful, precisely because of their broader nature. The size, scale, audience and scope of the content's reach can be a relevant factor—larger megaphones with bigger platforms are often more influential when they speak. As a result, they may well be more dangerous.

Harm therefore arises most prominently at the furthest ends of the spectrum. Content that has narrow, small-scale particularized harm is problematic precisely because of its effect on identified or theoretically identifiable individuals. Content that broadly impacts larger and more foundational norms of behavior is problematic despite the inability to identify a specific victim.

Though different in degree, localized, small-scale, direct harm and diffuse, broad-reaching, large-audience harm are both of significant concern. While assessing harm in this way may, at times, be difficult to do, it is not impossible. To the contrary, within contextual guidelines, it is essential that harm be assessed as it is a critically relevant consideration.

Finally, context is critical. And the origin of the content in question—including the identity of the user who generated it—is part of that context. There is a spectrum from an average citizen to a blue-check influencer, and from Elon Musk to President Jair Bolsonaro. Likewise, recognized political figures come with unique contextual circumstances: They are likely to use social media as a means of communicating with their supporters and also as a means of conducting official business. That duality often creates uncertainty when assessing harm and intent, to which the Board must be sensitive.

Political figures have outsized megaphones and thus rank highly for potential harm. While the dual nature of their speech must be evaluated, it is not categorically any different from any other form of ambiguous speech—some lines will be hard to draw; others will not. In the end, political actors should not have any specialized license to cause harm. The same framework of contextual harm analysis that applies to private citizens should apply to them as well.

Imminence

Another factor to consider is the immediacy of the anticipated harm. Harms that might materialize over a longer time frame are less susceptible to restriction than content directed at imminent events. By contrast, content that exacerbates ongoing events and creates an imminent risk is more likely to be subject to moderation.

This distinction is not new. International law allows greater action when harms are imminent and even permits anticipatory action to forestall imminent harm. The most notable area in which this doctrine arises, the international law of *jus ad bellum*, requires that the anticipatory action be both necessary and proportional.[22] No less should be true in response to the threat of imminent violence related to social media content. As a result, necessary and proportional actions to forestall violence are likely to be viewed as appropriate. This is especially salient and even more likely, in cases where actual harm is ongoing and continued social media content may contribute to or exacerbate the problem.

Incitement

Unlike the harmfulness section, which looks at the context within which the speech occurs, incitement looks at what the content provider actually intended. While motive is often difficult to discern and many actions are taken with a mixed motive, it is almost a truism that a person intends the natural consequences of his or her actions—and thus, motive can often readily be inferred from purposeful acts.

When assessing the incitement potential or qualities of user-generated content, the Board should begin by presuming benign motivation in the content provider. Nevertheless, there will be many instances in which mal-intent can be inferred. Most notable in this category will be circumstances where the content is intended to cause (that is to "incite") adverse, real-world consequences.

While the Board may wish to look at broader international restrictions on violent speech—such as Europe's restrictions on hate speech—at a minimum, the advocacy of illegal, physical violence is a significant factor in assessing the content in question.

To use the extremely narrow standard adopted in the United States, content which is "directed to inciting or producing imminent lawless action" and is "likely to incite or produce such action" should be more readily subject to restriction.[23]

---

22. Anthony Clark Arend, "International law and the preemptive use of military force," *The Washington Quarterly* 26:2 (2003), pp. 89-103. https://www.tandfonline.com/doi/abs/10.1162/01636600360569711.

23. *Clarence Brandenburg, Appellant, v. State of OHIO*. 395 U.S. 444, 89 S.Ct. 1827 (1969). https://www.law.cornell.edu/supremecourt/text/395/444.

Perhaps other, less inciting content may also be subject to review, but even under the most generous reading of IHRL, the advocacy of physical violence is cause for inquiry.

In applying this factor, the Board will ultimately face the challenge of judging content that fairly reads as an incitement of violence against illegitimate, repressive governments. For example, if Facebook was available at the time, content provided by Nelson Mandela designed to end the apartheid regime would be qualitatively different from content calling for the murder of an elected leader of a democratic nation. So, too, content provided by Alexei Navalny today is different from that provided by QAnon; even though both might fairly be read as inciting conflict, the content that originates with Navalny is, by and large truthful, while that from QAnon is, by any measure, false. There is an equivalent American analog as well: the question of what to do when social protest against a legitimate government gains a partial element of violence, as, for example, with the Black Lives Matter protests in 2020.

Again, context matters. Though requiring Facebook and the Board to serve as the arbiters of international legitimacy should be considered with caution, in many ways the necessity of drawing lines makes the role inevitable.

Appropriateness of Sanctions

A fifth and final factor stands apart from the other four, and concerns how to evaluate the appropriateness of a proposed sanction given a determination that sanction of some form is warranted. In the specific context of the Board's review of Facebook's ban of Trump, the proposed sanction was of maximum severity: a permanent future prohibition on posts. For future assessments, the four subfactors detailed below are relevant to weighting the appropriateness of potential sanctions:

1. General Deterrence – One question in determining the effectiveness of a potential sanction is whether or not the sanction in question would stop others from committing similar acts. Given that not all bad actors are caught, our theory of deterrence is that the penalty imposed must be significant enough to deter other actors, even when they discount the punishment to incorporate the possibility of not being caught. If the penalty is too lenient, then others will simply take their chances of similar punishment. In the context of social media, the cost of a total, permanent ban is quite high as it renders the relevant account useless; blocking specific content or limiting distribution of specific content or specific posts would generate much smaller, though non-trivial, costs.
2. Specific Deterrence/Disablement – Distinct from deterrence is the question of whether the particular speaker at issue is effectively disabled from engaging in prohibited acts in the future. This could be implemented through a number of methods, including both prohibitions on certain activity coupled with manual review or other mechanisms for enforcement, as well as total prohibitions on future use. Under high-harm or high-risk circumstances, disablement of some suitable form is sensible as a component of appropriate sanction; under others, it may be unnecessary, where repeated violations are relatively tolerable should they occur.
3. Contrition –The extent to which the actor acknowledges the nature of their prior acts and demonstrates remorse or behavioral adjustment is another clear factor to consider when assessing the severity of a sanction. If no clear recognition of culpability and harm is present, then it is reasonable to believe that the actor will repeat the activity even at risk of additional harm, and deplatforming and severe sanction become more compelling.
4. Availability of Effective Alternative Sanctions – The Board is not positioned to determine a sanction, as that is Facebook's decision to make, but rather to review a single proposed sanction. However, that review is best done by considering other sanctions that could have been imposed, and evaluating whether other, particularly more limited, sanctions could achieve the same practical effect at less cost to freedom of expression.

In the final analysis, taking into account the totality of the circumstances at hand, the chosen sanction should be proportional to the underlying act itself. Above and beyond the other subfactors, the sanction and its impact must not be disproportionately more severe than the original act itself and the consequences of that act.

Articulating these subfactors within a singular framework is useful as a means of increasing objectivity and consistency in analysis of an otherwise highly subjective determination. In particular, the subfactors look to a broader context around the specific incident under review, at similar parallel actions or alternative sanctions as well as what could occur in the future. These subfactors help ground a determination of appropriateness and keep it a meaningful, separate angle of analysis from the other factors such as harmfulness which might otherwise blend into consideration.

THE CASE OF FORMER PRESIDENT TRUMP

The specific question currently posed to the Board is whether former President Trump's indefinite suspension is justified as consistent with Facebook's values. While the question before the Board pertains to prohibiting Trump from future posting, his final posts serve as background for this analysis.

It is important to note that Trump's final posts come against the well-documented backdrop of prior activity that began before the 2020 election and continued well after it concluded. Likewise, these posts came even as Trump's false speech was being denounced by all authoritative decision-makers in U.S. courts and by many State governments. Recounting that history is beyond the scope of this paper; however, Facebook and the Board are entitled to, indeed obliged to, consider the complete context, as well as Trump's posts immediately preceding the suspension that is specifically at issue in the pending case.

In the weeks following the election, Trump and other leading Republicans voiced claims that the 2020 presidential election had been stolen by fraud.[24] These claims were made on both official media channels and via online social media platforms. On Jan. 6, 2021, a crowd of rioters broke into the Capitol with the intent to stop Congress from certifying then President-elect Joseph Biden's victory. In the late afternoon that day—while the rioters were still in the Capitol building, and were in some cases actively fighting with law enforcement—Trump posted a short video to Facebook once again claiming the election had been "stolen" and that it was "a fraudulent election."[25]

Shortly after the Capitol was declared secure by law enforcement, Trump posted the following on Facebook:

> These are the things and events that happen when a sacred landslide election victory is so unceremoniously viciously stripped away from great patriots who have been badly unfairly treated for so long. Go home with love in peace. Remember this day forever![26]

Both posts were promptly removed by Facebook, which stated that "on balance these posts contribute to, rather than diminish, the risk of ongoing violence."[27] Shortly afterwards, Facebook banned Trump from posting to Facebook for 24 hours. The next morning, Facebook founder Mark Zuckerberg announced an indefinite ban on Trump's ability to post on Facebook, saying that Trump's actions:

> involv[e] use of our platform to incite violent insurrection against a democratically elected government… Therefore, we are extending the block…indefinitely and for at least the next two weeks until the peaceful transition of power is complete.[28]

Based on the factors outlined above, within the IHRL framework of necessity/proportionality analysis, Facebook's decision to remove these posts and prohibit former President Trump from posting on Facebook in the future was well justified:

- Truth or falsity: Claims of a "stolen election" and a "fraudulent election" are demonstrably false.[29]
- Harmfulness: Heightened political tension and calls for violence from individuals in Trump's primary and intended audience make the posts especially inflammatory and harmful.
- Imminence: The most recent posts at issue were not only imminent to the commission of violence, but in fact violence was ongoing as the content was posted.[30] In context, these posts could reasonably be read as an incitement to additional, further violence, including specifically the phrases: "I know your pain," "There's never been a time like this, where such a thing happened, where they could take it away from all of us—from me, from you, from our country," and "Remember this day forever!"
- Incitement: Keywords in specific posts by Trump, including the famous "stand back and stand by,"[31] and "fight to the death,"[32] constitute an invitation and incitement to the commission of violence intended to disrupt the lawful transition of power.
- Appropriateness of Sanctions: At no point has Trump demonstrated true contrition, and every signal has indicated his intention to remain undeterred. In

24. See, e.g. "Joint Statement from Senators Cruz, Johnson, Lankford, Daines, Kennedy, Blackburn, Braun, Senators-Elect Lummis, Marshall, Hagerty, Tuberville," Press Office of Senator Ted Cruz, Jan. 2, 2021. https://www.cruz.senate.gov/?p=press_release&id=5541.

25. "Announcing the Oversight Board's next cases," Oversight Board, January 2021. https://oversightboard.com/news/175638774325447-announcing-the-oversight-board-s-next-cases.

26. Ibid.

27. Guy Rosen and Monika Bickert, "Our Response to the Violence in Washington," Facebook Newsroom, Jan. 6. 2021. https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc.

28. Mark Zuckerberg, *Untitled*, Jan. 7, 2021. Facebook Post. https://www.facebook.com/zuck/posts/10112681480907401.

29. Cybersecurity and Infrastructure Security Agency, "Joint Statement from Elections Infrastructure Government Coordinating Council & The Election Infrastructure Sector Coordinating Executive Committees," U.S. Dept. of Homeland Security, Nov. 12, 2020. https://www.cisa.gov/news/2020/11/12/joint-statement-elections-infrastructure-government-coordinating-council-election.

30. "Announcing the Oversight Board's Next Cases." https://oversightboard.com/news/175638774325447-announcing-the-oversight-board-s-next-cases.

31. "Trump: 'Proud Boys, stand back and stand by'," Reuters YouTube channel, Sept. 30, 2020. https://www.youtube.com/watch?v=6MgfX81kWEU.

32. Staff, " Factbox-Trump on Twitter (Dec. 26) - Defense Bill, Election, FBI," *Reuters*, Dec. 26, 2020. https://www.reuters.com/article/usa-trump-tweet/factbox-trump-on-twitter-defense-bill-election-fbi-idUSKBN2900DP.

addition, few high-profile speakers who have echoed similar and supportive messages have faced substantial consequences, thus reinforcing the need for strong sanctions that can effectively disable and deter future behavior. By not previously banning Trump, Facebook could be said to have enabled and encouraged more and more brazen posts by both Trump and other high-profile speakers precisely because there was no sanction. Facebook is reasonable in finding proportionality and a lack of alternatives to permanent suspension in order to mitigate future harm as best as possible.

These are not easy conclusions. And this is ultimately a predictive judgment that future posting would be consistent with past actions given the lack of expressed contrition or regret. However, restoring Trump's posting privileges is defensible under different predictions. In such circumstances, any future restoration should include strict conditions regarding further harmful activity, with permanent suspension as the final step.

Finally, in referring this matter to the Board, Facebook solicited broader, more general guidance as to how it should treat content posted by political actors. Given the necessity for a case-by-case analysis, the Board should decline that invitation. In light of the critical importance of context, no rules and guidelines can be developed that speak generally to the permissible scope of political content on the platform. The factors outlined above can provide some guidance, but the Board should be cautious about providing further, more general, broadly-applicable, direction to Facebook. Such guidance would be, necessarily, ill-defined and likely cause more confusion than benefit.

CONCLUSION

Each suspension decision is unique and based on context-specific considerations. Therefore, it is uncertain whether the Board can provide the generalized guidance to Facebook that has been requested. It is also worth noting that the Board's oversight function will not scale readily and encourage Facebook to adopt a robust implementation strategy to generalize the Board's guidance. The Board can help close that gap through consistent use of an explicit, multi-factor framework, one that offers more granularity beyond general international human rights law principles, to evaluate individual case and context specific matters, such as we have proposed.

Ultimately, the Board's engagement in Facebook's content moderation decisions is a welcome development and its decision to review a high-profile matter—such as the current case involving former President Trump—will heighten the salience of its review. Properly structured, the Board's functionality is a welcome addition that is likely to advance human rights and the development of an international common law of content moderation.

ABOUT THE AUTHORS

Paul Rosenzweig is the resident senior fellow for Cybersecurity and Emerging Threats at the R Street Institute. He works on legal and policy issues related to cybersecurity, homeland security, national security and tech policy, including the intersection of privacy and security.

Chris Riley is the senior fellow of Internet Governance at the R Street Institute. He is leading the Knight Foundation-funded project on content moderation, running convenings of a broad range of stakeholders to develop a framework for platforms managing user-generated content.

Mary Brooks is the senior research associate for Cybersecurity and Emerging Threats at the R Street Institute. Her areas of focus include securing 5G and ICT infrastructure; technology and human rights; and how public audiences understand cyber power and security.

Tatyana Bolton is the policy director for Cybersecurity & Emerging Threats at the R Street Institute. She crafts and oversees the public policy strategy for the department with a focus on secure and competitive markets, data security and data privacy, and diversity in cybersecurity.