WHISTLEBLOWER AID

Report government and corporate lawbreaking.
Without breaking the law.

—ANONYMOUS WHISTLEBLOWER DISCLOSURE—

██████████████

SEC Office of the Whistleblower
Via Online Portal & Fax

Re: Supplemental Disclosure of Securities Law Violations by Facebook, Inc. (NASDAQ: FB), SEC TCR #████████████

Facebook misled investors and the public about its role perpetuating misinformation and violent extremism relating to the 2020 election and January 6th insurrection.

To the SEC Office of the Whistleblower:

1. The instant letter is one of multiple disclosures related to the above-captioned matter. Our anonymous client is disclosing original evidence showing that **Facebook, Inc. (NASDAQ: FB)** has, for years past and ongoing, violated U.S. securities laws by making **material misrepresentations and omissions in statements to investors and prospective investors,** including, *inter alia*, through filings with the SEC, testimony to Congress, online statements and media stories.

2. On August 26, 2021, the **U.S. House of Representatives Select Committee to Investigate the January 6 Attack on the United States Capitol** requested from Facebook:[1]

   All internal or external reviews, studies, reports, data, analyses, and related communications regarding your platform(s) and:

   i. Misinformation, disinformation, and malinformation relating to the 2020 election;

---

[1] https://january6th.house.gov/sites/democrats.january6th.house.gov/files/2021-08-26.BGT%20to%20Facebook.pdf .
Emphasis is added throughout this disclosure in bold/underlined text.

ii. Efforts to overturn, challenge, or otherwise interfere with the 2020 election or the certification of electoral college results;

iii. Domestic violent extremists . . .

iv. Foreign malign influence in the 2020 election . . .

3. Facebook has publicized its work to combat misinformation and violent extremism relating to the 2020 election and insurrection, such as efforts to remove hate groups and inciting content and to employ "fact checkers." In reality, Facebook knew its algorithms and platforms promoted this type of harmful content, and it failed to deploy internally-recommended or lasting counter-measures.

FACEBOOK'S MISSTATEMENTS AND OMISSIONS

4. **Facebook made misstatements and omissions regarding its facilitation of political misinformation, including in testimony before Congress.** For example, in the March 2021 hearing on "Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation," Mark Zuckerberg was asked:[2]

   "Yes or no: ***Do you agree that your company has profited from the spread of disinformation****?"*

5. Mr. Zuckerberg replied:

   "Congressman, ***I don't agree with that.****"*

6. Further, in related prepared testimony, Mark Zuckerberg represented:[3]

   "We did our part to secure the integrity of the election . . . Now, some people say that the problem is that social networks are polarizing us. But that is not at all clear from the evidence or research."

7. Moreover, Mr. Zuckerberg has stated:[4]

   *"****We also work to reduce the incentives for people to share misinformation to begin with****. Since a lot of the misinformation that spreads*

---

[2] https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Transcript-20210325.pdf.
[3] https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Transcript-20210325.pdf.
[4] https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/Witness%20Testimony_Zuckerberg_CAT_CPC_2021.03.25.pdf.

*online is financially motivated spam, we focus on disrupting the business model behind it. **We take action against Pages that repeatedly share or publish content rated false, including reducing their distribution** and, if necessary, removing their ability to monetize. And we've enhanced our recidivism policies to make it harder to evade our enforcement . . . we have built industry-leading policies, teams and systems to keep hate and violence off our platform . . .*

***We remove language that incites or facilitates violence, and we ban Groups that proclaim a hateful and violent mission** from having a presence on our apps. We also remove content that represents, praises, or supports them . . ."*

8. **Facebook also made misrepresentations outside of Congressional testimony, such as in investor calls and public pages.** For instance, in Facebook's First Quarter 2020 Results Conference Call in April 2021, David Wehner, CFO, asserted:

    "[W]e really do more than anyone else in the industry on the safety and security front to prevent things like misinformation."[5]

9. Similarly, in its public page and press release, Facebook stated:[6]

    "[W]e have also taken enforcement action consistent with our policy banning militarized social movements like the Oathkeepers and the violence-inducing conspiracy theory QAnon. We've also continued to enforce our ban on hate groups including the Proud Boys"

10. Further, after Facebook's "independent Oversight Board" upheld the decision to suspend Donald Trump's account, Facebook claimed:[7]

    *"Our Violence and Incitement policy prohibits content calling for or advocating violence, and **we ban organizations and individuals that proclaim a violent mission** under our Dangerous Organizations and Individuals policy. We believe our Dangerous Organizations and Individuals policy has long been the broadest and most aggressive in the industry, and we have used it to ban hate groups. Motivated by a range of indicators that suggested political violence in the United States was possible, in August*

---

[5] https://s21.q4cdn.com/399680738/files/doc_financials/2021/Q1/FB-Q1-2021-Earnings-Call-Transcript.pdf.
[6] https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/
[7] https://about.fb.com/wp-content/uploads/2021/06/Facebook-Responses-to-Oversight-Board-Recommendations-in-Trump-Case.pdf.

WHISTLEBLOWER AID

*Report government and corporate lawbreaking.
Without breaking the law.*

2020, we expanded this policy to address militarized social movements and violence-inducing conspiracy networks, such as QAnon . . .**The responsibility for January 6, 2021 lies with the insurrectionists** . . .”

11. Mark Zuckerberg has stated in Congressional testimony:

    “Platforms . . . should be required to have adequate systems in place to address unlawful content.”[8]

    “[W]hat we should be judging the companies on is how many people see harmful content before the companies act on it.”[9]

    *“One important place to start would be . . . **making sure that companies can't hide behind Section 230 to avoid responsibility** for intentionally facilitating illegal activity on their platforms.”*[10]

12. Moreover, immediately after the January 6 insurrection, Facebook provided its advertising sales teams with talking points for advertisers, assuring them that Facebook was actually removing violent and inciting content:[11]

    “Q: Should I pause my advertising?

    We are monitoring the situation closely and are continuing to enforce our policies which prohibit incitement and calls for violence on our platform. We are actively reviewing and removing any content that breaks these rules . . .

    [New Jan 15] Q: Why didn't you see the Jan 6 violence coming? Why didn't you take more action in advance of January 6?

    While we put emergency measures into place after the Violence in Washington DC, our teams have been actively monitoring and removing Pages, groups and events that violate any of our policies, including calls of violence, since well before January 6.

    *. . . **as we have done since long before January 6, we will continue to remove content, disable accounts, and work with law enforcement when there is a risk of physical harm or direct threats to public safety***.

---

[8] https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Wstate-ZuckerbergM-20210325-U1.pdf.
[9] https://www.rev.com/blog/transcripts/tech-ceos-senate-testimony-transcript-october-28.
[10] https://www.rev.com/blog/transcripts/tech-ceos-senate-testimony-transcript-october-28.
[11] ▮▮▮▮▮▮ *Jan 6 note to advertisers*, p. 1-3, 7.

> *We continue to monitor activity on our platform in real time and remove anything that breaks our rules."*

13. **Facebook has dismissed concerns that it's recommendation systems amplify interests (namely that mainline political interests are redirected to fringe/radical/polarizing variants) by saying "it takes two to tango" - that individuals must take "responsibility" for the "choices" they make which "drives" the content they see in their feeds**

"But ultimately, content ranking is a dynamic partnership between people and algorithms. On Facebook, it takes two to tango.

In a recent speech, the Executive Vice President of the European Commission, Margrethe Vestager, compared social media to the movie The Truman Show. In it, Jim Carrey's Truman has no agency. He is the unwitting star of a reality TV show, where his entire world is fabricated and manipulated by a television production company. But this comparison doesn't do justice to users of social media. You are an active participant in the experience.

The personalized "world" of your News Feed is shaped heavily by your choices and actions. It is made up primarily of content from the friends and family you choose to connect to on the platform, the Pages you choose to follow, and the Groups you choose to join. Ranking is then the process of using algorithms to order that content."[12][13]

ORIGINAL EVIDENCE CONFIRMING MISSTATEMENTS AND OMISSIONS

14. Facebook's records confirm that Facebook knowingly chose to permit political misinformation and violent content/groups and failed to adopt or continue measures to combat these issues, including as related to the 2020 U.S. election and the January 6th insurrection, in order to promote virality and growth on its platforms.

---

[12] Original editorial: Nick Clegg. "You and the Algorithm: It Takes Two to Tango." Nick Clegg. March 31, 2021. https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2
[13] Nick Clegg. "You and the Algorithm: It Takes Two to Tango." Facebook Newsroom Blog. March 31, 2021. https://about.fb.com/news/2021/03/you-and-the-algorithm-it-takes-two-to-tango/

15. Facebook only actions less than 1% of Violence and Inciting to Violence (V&I) content on Facebook - Facebook's strategy of focusing on Content over other solutions lets this content effectively run free

> "Our focus on scaled content-level enforcement in practice means the volume of decisions which need to be made is **impossible for human reviewers to keep up with**, and so we apply classifiers to make content-level decisions at scale. This approach shines where content-level decision can happen without considering context, but is extremely challenging to implement where such context is needed.

> For example, we estimate that we may action as little as 3-5% of hate [speech] and ~0.6% of V&I [n.b. Violent and Inciting Content] on Facebook...[t]his is because the problem of incorporating semantic meaning in classification is extraordinarily challenging, especially when the process needs to be localized across languages and have rudimentary interpretation of intent."[14]

16. Facebook has demonstrated via experiments using brand new test accounts how rapidly Facebook's algorithms can veer people interested in Conservative topics into radical or polarizing ideas and groups/pages, some demonstrating traits of Coordinated Inauthentic Behavior (CIB) akin to what was seen by the Macedonians in 2016:[15]

> "Page and Group Recommendation System Risks

> After a small number of high quality/verified conservative interest follows (Fox News, Donald Trump, Melania Trump — all official pages), within just one day Page recommendations had already devolved towards polarizing content.

> Although the account set out to follow conservative political news and humor content generally, and began by following verified/high quality conservative pages, **Page recommendations began to include conspiracy recommendations after only 2 days** (it took <1 week to get a QAnon recommendation!)

---

[14] ███ *Problematic Non-violating Narratives document, p. 7.*
[15] ███ *- "Carol's Journey to QAnon - A Test user Study of Misino & Polarization Risks Encountered Through Recommendation Systems (Part 1)".*

Group recommendations were slightly slower to follow suit — it took 1 week for in-feed GYSJ recommendations to become fully political/right-leaning, and **just over 1 week to begin receiving conspiracy recommendations**.

Some Pages and Groups about US politics that were recommended to the account [underline/italics in the original] showed signs of low quality or suspicious origins:

- A few of the recommended Groups and Pages that Shared problematic political content were quite new (<2 month old) with foreign admins.
- Several of the Pages and Groups that shared problematic content had other markets of low-quality/biolating ownership activity, such as a high number CO deletes for admins or in Pages or Groups with overlapping sets of admins.
- Admins from these recommended groups and pages often have numerous previous Pages unpublished/taken down for Impersonation and CIB violations in other Pages or Groups they've admin-ed....

Misinformation Risks

A number recommended Groups and Pages contain and share known misinformation

- Despite this, **I have never encountered an MRA treatment for a false item** (Note: Misinformation Related Articles, is our current user-facing Inform treatment for fact-checker rated misinfo).
- Comments on misinformation shared in politically-focused Groups and Pages are nearly always supportive of the misinformation belief/position (disbelief comments are unlikely to be available/helpful signal in these contexts).

Comments are potentially an important source of misinformation, even on seemingly innocuous articles

- Misinfo comments on Pages/Groups content may serve to legitimize support for misinfo beliefs/positions.
- This may be especially true when badged "Top Contributors" are the misinformation sources.

Notifications for posts in polarized and conspiracy Groups and Pages encourage users to view false/hateful content

- *They also encourage users to visit Group/Page surfaces which have less robust misinformation protections (e.g. demotions) than Feed.*

Prominent Page 'Like' buttons on re-shares (which display above the post 'Like'/react button), may encourage users to follow sources of problematic content

- *This may lead to additional future exposure to similar content, when the user may have just been intending to like the post."*

17. **Teams have identified ways to combat misinformation and violence:**

*[June 2020] "[W]e found that a subset of clusters have disproportionately higher in-cluster misinformation circulation on Instagram Home. . . **there are a few potential improvement[s] we can make in our current misinfo detection and prediction process** . . ."[16]*

*[September 2020] "The Segmentation team has **built a segmentation of 'populations at risk' in anticipation of the US 2020 election**, with a particular emphasis on misinformation and voter disenfranchisement. . . .Civic Targeting Risk Scores (CTRS) that will focus on sub-groups of users across four dimensions [low participation, more exposed, higher audience value, and more vulnerable]."[17]*

*"There's a growing set of research showing that some viral channels are used for bad . . . we've also identified **opportunities where reducing virality may significantly reduce prevalence of Integrity problems (10% in some cases)**, across the Family of Apps . . ."[18]*

18. But internal documents show that decisions are based on inherent conflicts of interest:

*"**Facebook's decision-making on content policy is routinely influenced by political considerations**."[19]*

---

[16] ▮▮▮▮ *Misinfo narrowcasting on Instagram Home*, p. 2, 19.
[17] ▮▮▮▮ *Civic Targeted Risk Scores*, p. 1-2.
[18] ▮▮▮▮ *Virality Reduction as Integrity Strategy.pdf*, p. 1.
[19] ▮▮▮▮ *Political Influence Content Policy*, p. 1.

*[September 2020] "Facebook currently has **no firewall to insulate content-related decisions from external pressures**. It could have one."*[20]

*"[Key is] Facebook's role in enforcing our misinformation policies, specifically when we lift certain enforcement consequences. We defer to third-party fact-checkers to rate content on our platforms, but we are responsible for how we enforce on those ratings . . . [in certain circumstances we] lift other penalties." [Commentary] "**Does it present a conflict of interest to have the org that manages relationships with these partners be the one making calls on lifting penalties?** . . . the enforcement decisions need to be independently-reviewed too, otherwise it defeats the whole purpose of hiring outside experts . . . we sometimes decide it would not be appropriate to impose such steep consequences when we feel a fact checker's rating doesn't align with our public ratings definitions."*[21]

19. **Facebook measures success in terms of proactively detecting hate speech before it gets reported but has made it more difficult for users to report hate speech by changing reporting flows:**

[September 2019] "...there are only three levers that the product team has over cost that matter: 1. Reviewing fewer user reports...

In order to review less content we can do one of four things: A. Ignore more benign user reports B. Auto-delete more violating user reports C. Auto-close low-value user reports (ones that are less likely to violate and are getting minimal views) D: Move even earlier up the funnel to before we have the user report in the first place...

D. Improve the top of the reporting funnel

***By tweaking the user reporting process in partnership with the CIX team in London, we can both add thoughtful friction that reduces the number of spurious user reports we receive** and make the reports that do come in more actionable."*[22]

20. **Facebook knows that its products make hate speech and misinformation worse:**

---

[20] *A firewall for content policy,* p. 1.
[21] *Employee concerns on political misinfo policies,* p. 1-3.
[22] *Hate Speech Cost Controls* p. 9-12.

[November 2020] **"Not only do we not do something about combustible election misinformation in comments, we amplify them and give them broader distribution**."[23]

*"We have evidence from a variety of sources that **hate speech, divisive political speech, and misinformation** on Facebook and the family of apps are **affecting societies around the world**. We also have compelling evidence that **our core product mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish** on the platform . . . the net result is that Facebook, taken as a whole, will be actively (if not necessarily consciously) promoting these types of activities. The mechanics of our platform are not neutral."[24]*

*"**Facebook's algorithms have coaxed many Americans into sharing extreme views** on the platform -- **rewarding** them with likes and shares for posts on subjects like **election fraud conspiracies** . . ."[25]*

"What does all of this say about the nature of algorithmic rewards? . . . we were late on data security, misinformation, and foreign interference." [Commentary] *"In changing the face of advertising, we've also (perhaps inadvertently) changed the rules by which narratives are delivered to people via the news and communications media."[26]*

[September 2020] *"[T]here are **opportunities to improve** civic conversations on FB [Facebook] . . . we predicted that comments reacted favorably (e.g., loves) by diverse audiences would be more valuable & higher quality and less likely to contain attacks, ridicule, & toxicity . . . Motifs are patterns of positive or negative reactions and replies users give each other in comment exchanges [e.g., "likes/loves" versus angry face] . . . boosting posts from people who have a history of good motifs . . . found that users perceive content to be more meaningful and respectful [and] . . . increased high-quality comment production in the civic domain . . .*

Echo Chamber trap: In some domains, like civic, politics . . . people tend to form social networks comprised of people who disproportionately share their

---

[23] *Misinfo in comments*, p. 2.
[24] *What is Collateral damage?* p. 35.
[25] *They used to post selfies, now they're trying to reverse the election*, p.1.
[26] *Andrew Bosworth Thoughts for 2020*, p. 4-5, 16.

beliefs and do not fact-check their personal beliefs [ ] In these echo chambers, people polarize, and are more likely to embrace conspiracy theories, hate speech, and false news . . . Therefore, to bring people closer together, we need to ensure we are promoting the content that the most users get the most value from, and not just content liked by some homogeneous group of people in an echo chamber . . . World2vec is a measure of broad diversity and may help enhance the power of motifs to identify high-quality, valuable content . . .

Action Items . . . Integrate that model into comment ranking to boost comments that are more likely to contain civic value and demote comments that more likely to be low in civic value."[27]

21. Facebook relies on expressions of doubt and people reporting misinfo as a critical mechanism for identifying potential misinformation to be Fact Checked by Third Party Fact Checkers, ignoring that homogenous communities are unlikely to challenge the beliefs already held by the community - amplifying the danger of Narrowcast Misinformation (which is 2-3x as a prevalent as viral misinformation):

*[April 2020] **"Narrowcast Content"** is "content in which the majority of vpvs [View Port View or impression], shares, and comments all occur on posts associated with a single sociographic segment . . .One of the concerns with narrowcast misinformation is that **misinformation targeted at a vulnerable population or a population with outsize political influence may be highly impactful without going viral. This means that misinformation in these contexts could be important to identify and remediate** . . ."[28]*

*"**[N]arrowcast misinformation** [which focuses on falsifiable objects circulating among a single/narrow subpopulation] may be **~2-3x as prevalent** as generally viral misinformation."[29]*

"Users are ~3-5x more likely to produce disbelief comments on broadcast content than narrowcast content overall...

Among narrowcast content, users are ~2-3x more likely to produce disbelief comments on out-of-segment shares of narrowcast content than in-segment shares...

[27] ▮▮▮▮▮▮ *Diverse Engagement May Identify Valuable Civic Comments*, p.1, 4-10, 32.
[28] ▮▮▮▮▮▮ *Identifying Narrowcast Misinfo Hard because missing or miscalibrated doubt*, p. 2-3, 7.
[29] ▮▮▮▮▮▮ *Narrowcast Misinfo Prevalence Update*, p. 1-1.

Users are ~2x more likely to produce disbelief comments on cross-cutting shares of civic content…

Across all content, users are ~2-5x more likely to produce disbelief comments on content posted by ideologically cross-cutting pages and users than on content posted by aligned pages and users…

On narrowcast content rated 'false' by 3PFC review, out-of-segment audiences produce disbelief comments at ~3x the rate of in-segment audiences, indicating these discrepancies are not an artifact of differential misinformation prevalence…

Across enqueued [for fact checking] content, in-segment audiences produce disbelief comments at ~10% the rate of out-of-segment audiences and ~1% the rate of broadcast audiences. "[30]

22. Pages that repeat offend for misinformation are permitted to continue to spread misinformation:

"Page Admins who were responsible for 2 or more misinformation posts in pages in the last 60 days are responsible for 59% of misinformation VPVs in the current week. **If we only consider VPVs from US users, page admins who were responsible for 2 or more misinformation posts in the last 60 days are responsible for 78% of misinformation US VPVs in the current week.**

Enforcing on pages moderated by page admins who post 2+ pieces of misinformation in the last 67 days would affect 277,000 pages. Of these pages, 11,000 of them are current RO [n.b. repeat offender] pages and 40,000 would be caught by including matched misinformation as an RO strike. "[31]

Note: A very small number of pieces of content are fact checked each year - getting two 'strikes' is difficult to do.

23. **Facebook has "whitelisted" political users who violate its terms, leading to the spread of misinformation and violence on and off of the platform:**[32]

---

[30] ██████ *Identifying Narrowcast Misinformation May be Uniquely Challenging Due to Missing or Miscalibrated Expressions of Doubt."* p. 1
[31] ██████ *"Sizing the Opportunity for Expanding Misinformation RO Strikes to Page Admins"*
[32] *See also Disclosure re. XCheck (pronounced Cross-Check).*

"Under the political whitelist policy, content posted by whitelisted Pages or Profiles is not subject to our misinformation interventions (i.e., demotions, inform treatments, or enqueuing to 3PFCs [third-party fact checkers]) . . . These results indicate **the current whitelist policy is protecting content that is especially likely to deceive, and hence to harm, people on our platforms**. *[ ] We recommend ending the whitelist, or at least (1) informing users; (2) reducing distribution (e.g. through down-ranking in feed)."*[33]

*"**Facebook routinely makes exceptions for powerful actors when enforcing content policy.** The standard protocol for enforcement and policy involves consulting Public Policy on any significant changes, and their input regularly protects powerful constituencies . . ."*[34]

"Politicians share misinformation on topics with high-risk of societal impact [ ] Widespread expert consensus shows that misinfo shared by politicians has disproportionate impact on users compared to that shared by other sources . . .Users think it's Facebook's responsibility to inform them when their leaders share false information. . . On Sept 24th, we publicly announced our policy on fact-checking political figures [because] . . . We don't currently allow fact checking on political figures."[35]

24. **Facebook knows that "deep reshares"** (where content is reshared multiple times from an original post[36]) **facilitate misinformation and violence but it has only restricted "deep reshares" in very limited circumstances**

"When a user sees a reshare of a reshare (depth > = 2) of a link or a photo, we find they are 4 times more likely to be seeing misinfo compared to when they see links or photos on News Feed in general. This increases to 5-10X at higher reshare depths. . . 30-70% of misinfo viewership occurs on reshares of reshares . . . In the US, that's 65% for photo misinfo, and 35% for link misinfo . . .The reshare depth of a post is defined by how many reshares away it is from the original post. . . **political operatives and publishers tell us that they rely more on negativity and sensationalism for distribution due to recent algorithmic changes that favor reshares."***[37]

---

[33] ███████ *Comparing the effects of misinfo from politicians vs ordinary user sources*, p. 1.
[34] ███████ 2020-08-18 *Political Influence on Content Policy*, p. 4.
[35] ███████ *Effects of Politician Shared Misinformation*, p. 3-4.
[36] ███████ *Reshare ranking exp in India: Indonisia*, p. 1, 4; *see also* ██████ *Reshare depth india and indonesia.*
[37] ███████ *Deep Reshares and Misinformation*, p. 1, 3.

*"A ranking change which reduces ranking based on max reshare depth produces significant wins on a variety of integrity measures . . . Stemming from the initial finding that reshare depth is correlated with misinformation, subsequent research has found that other **integrity harms** also **correlate with reshare depth** . . ."*[38]

*"70% of VPVs [View Port View, Facebook's term of art for a user impression] on known link misinfo come from reshares . . . 50% to 80% of VPVs on known link misinfo come from the top 1K most seen links per day . . . Reshare depth appears to be an especially good signal for targeting link misinfo in India and the Philippines. . . **misinfo prevalence is 15-25X higher among VPVs of depth 2+ reshares** . . ."*[39]

*"**An effective, content-agnostic approach to mitigate the harms** posted by high-harm misinfo (e.g. civic or health) would be to dampen virality within these topics by **hard demoting all deep reshares where the viewer is not a friend or follower of the original poster.**"*[40]

*"Political parties . . . claim that Facebook's algorithm change in 2018 (MSI) has changed the nature of politics. For the worse. They argue that the **emphasis on 'reshareability' systematically rewards provocative, low-quality content**."*[41]

*"[In Myanmar] Misinformation, misrepresentation and account issues are a few known high-risk abuse areas that will be exacerbated by a second COVID outbreak . . . **The Reshare Depth demotion reduces the distribution of highly-viral content** to give more distribution to content produced by friends or connections of friends . . .*

*[N.B.: Measures enacted to demote reshare depth to reduce the distribution of highly-viral content in fact **reduced Viral Inflammatory Prevalence by 25.1% and Photo Misinformation by 48.5% in Myanmar, Facebook said:**] We plan to roll back this intervention after the Myanmar election in November."*[42]

---

[38] ███████ *Max Reshare Depth experiment, p. 1-2.*
[39] ███████ *Further reading from Reshare Depth Article, p. 2.*
[40] ███████ *Fighting high harm misinfo with deep reshare damping, p. 1; see also* ███████ *Reshare depth by country.*
[41] ███████ *Political Party response to the '18 Algorithm change, p. 1.*
[42] ███████ *Reshare depth exp in Myanmar, p. 1.*

25. Facebook's failure to limit reshares is incredibly risky because under one percent of users drive half of all the user impressions [called "VPVs"] on reshare posts:

> "Considering the 28 days from February 22 - March 21, 2021
>
> - 2.82 B users were active on Facebook.
> - 978 M users created at least one reshare post (34.6%).
> - 37 M users (1.3% created half of all reshare posts.
> - 21 M users (0.7%) created reshapes accounting for half of all VPVs on reshare posts (within three days)."[43]

26. To avoid criticism resulting from inevitable false positives when removing harmful content, Facebook chooses to "demote" it instead, which it knows to be an ineffective response:

> "Due to the unboundedness of ranking scores, neither global all demotions nor viewed term demotions can effectively re-rank some content...[I]f we have some signal that a piece of content is low-quality or will cause a bad experience and we want to counter the engagement score, we have to apply a pretty big demotion . . . How big would the integrity demotion on global all have to be for a piece of content to be ranked X places lower? . . . **some content would not be moved even if you had a global all demotion of 90%+** . . . .Let me make this last point clearer. **If you demoted the top-ranked VPV by** 10%, 74% of these VPVs would maintain their top spot. If you demoted the top-ranked VPV by 50%, 27% would be unmoved. **4.5% of VPs are impervious to even a 90% demotion.**"[44]

27. **Facebook has avoided or rolled back interventions for "groups" and "narrow subpopulations" that it knew reduced misinformation, violence and incitement, because those interventions reduced the platform's growth:**

> "Harmful communities have been seen to grow via using our platform affordances . . . **the QAnon community relied on minimally-connected bulk group invites . . . One member sent over 377,000 group invites in less than 5 months!** . . . The civic integrity team explored a variety of changes to the way civic content was ranked during US 2020 and identified several features which promoted a more healthy ecosystem . . ."[45]

---

[43] ████ *"Quantifying the Concentration of Reshares and their VPVs Among Users"* p. 1.
[44] ████ *Demotions in practice*, p. 5, 9-10.
[45] ████ *Harmful Non-Violating Narratives*, p. 8, 10, 15, 19, 20.

WHISTLEBLOWER AID

Report government and corporate lawbreaking.
Without breaking the law.

[However:] "**We have rolled back the pre-election rate limit of 100 Group invites/day due to it having significant regression on Group growth** . . . *[which was implemented in October 2020 to reduce] joins to sensitive groups or groups that became sensitive.*"[46]

"*We've seen that integrity* **harms tend to be concentrated** *in a small number of User Interest Subpopulations,* **with 50% or more of a given harm often in 1% or less of Subpopulations** . . . *More than 50% of civic misinformation on Facebook is associated with a handful of highly-civic communities. . . Misinformation on Instagram is also disproportionally concentrated among a few Subpopulations. . . Over 50% of QAnon & MSM (Militarized Social Movements) Groups are associated with a single Subpopulation . .*"[47]

[March 2020] "*[T]here are likely meaningful opportunities for impact via integrating Segmentation with either Voter Suppression or Hate Speech [ ] The majority of US Groups posts containing Voter Suppression over the past 2 months occurred in Groups associated with just 2 sociographic segments [ ] The Groups in segments with the highest prevalence experience 100x as many borderline hate speech posts as the overall average.*"[48]

[April 2020] "*Disbelief comments are produced at disproportionate rates on civic content posted by extreme partisans. This suggests content posted by extreme partisans might be more likely to be false . .*"[49]

Facebook's own tools are used to direct narrowcast misinformation:

"**Up to 64% of IRA [Internet Research Agency, a Russian troll farm] ads may have been targeted using Facebook's automated targeting suggestions** . . . "[50]

[From August 2020] "*Many Users Are Repeat or Serial Offenders For Both Misinformation and Hate . . . surfaced a connection between misinformation and hate within the context of highly polarized communities along with patterns of serial offending and potential recidivism . . . There may be substantial intersections between hate and misinformation, which could be*

[46] *Killswitch Plan for all Group Recommendation Surfaces*, p. 3.
[47] *Subpopulations: Segmentations Wiki*, p. 1, 19-21.
[48] *Sociographic Segments may be impactful for hate speech and voter suppression*, p. 1.
[49] *Identifying Narrowcast Misinfo Hard because missing or miscalibrated doubt*, p. 3.
[50] *Measuring Human Perception to Defend Democracy*, p. 24.

*an area for intervention . . . 59M users have posted or shared at least 3 pieces of predicted **misinformation** in the past 90d. (!!) . . . **99% of these user accounts remain active, and some of them have passed dozens of authenticity checks.**"[51]*

2020 Election & Capitol Insurrection

28. In April 2020, removing "Downstream MSI" from Facebook's Meaningful Social Interactions algorithm (the News Feed's goaling mechanism) was suggested as a "soft action" that would significantly decrease the amount of hateful, polarization, or violence inciting content from the News Feed without doing *any* content based censorship. CEO Mark Zuckerberg rejected this intervention that could have reduced the risk of violence in the 2020 election:

> *"Downstream model depreciation:* **Mark doesn't think we could go broad . . . We wouldn't launch [the proposed intervention to reduce misinformation] if there was a material tradeoff with MSI impact** *["meaningful social interactions"] ."[52]*

> *Note: The "Meaningful" part of Meaningful Social Interactions is more a flourish than substance - for example, all comments were considered "meaningful", even if they were hate speech or bullying (for six months after this memo). All likes were considered "meaningful" along with all reshares.[53]*

29. Facebook's records state:

> *[August 2020] "We should avoid making significant changes to the IFR [in feed recommendations] publisher ecosystem prior to the US 2020 election, if those changes will have a large impact on entity-level distribution for political figures, commentators, or news organizations in the US . . . FINAL DECISION . . . **we have decided to not make any changes until the election is over.** We will revisit the issue after the election."[54]*

> *[August 2020]* **"*I've seen promising interventions from integrity product teams, with strong research and data support, be prematurely stifled or severely constrained by key decision makers--often based on fears of public and policy stakeholder responses*.** *. . . . For example, we've known*

---

[51] ███████ *Serial misinfo and hate offenders, p. 2,3.*
[52] ███████ *Mark Feedback on Soft Action Proposal + Deck presented to Mark, p. 1.*
[53] ███████ "Filtering out engagement-bait, bullying and excessive comments from MSI Deltoid metric 11/1 onwards"
[54] ███████ *Update On Political Publisher Issues, p. 1-2.*

- REDACTED FOR CONGRESS -

WHISTLEBLOWER
AID
Report government and corporate lawbreaking.
Without breaking the law.

for over a year now that our recommendation systems can very quickly lead users down the path to conspiracy theories and groups. [ ] While the Recommendations Integrity team has made impressive strides in cleaning up our recs, **FB [Facebook] has been hesitant to outright ban/filter conspiracy groups like QAnon until just last week.** *[ ] In the meantime, this fringe group/set of beliefs has grown to national prominence with QAnon congressional candidates and QAnon hashtags and groups trending in the mainstream.* **We were willing to act only *after* things had spiraled into a dire state**.*"*[55]

30. Thus, Facebook allowed misinformation and harmful content to persist.

"Through most of 2020, we saw non-violating content promoting QAnon spreading through our platforms. Belief in the QAnon conspiracy took hold in multiple communities, and we saw multiple cases in which such belief motivated people to kill or conspire to kill perceived enemies . . . Policies don't fully cover harms [ ] We implement policies for many of these areas that limit our ability to act . . . high-profile entities were able to serially spread . . . claims without crossing our falsifiable misinformation based lines for enforcement. . . we've often taken minimal action initially due to a combination [of] policy and product limitations making it extremely challenging to design, get approval for, and roll out new interventions quickly. Afterward, we've often been prompted by societal outcry at the resulting harms to implement entity takedowns. For instance, we've enacted this pattern with both QAnon and delegitimization, wherein we **initially took limited or no action**, *and later decided to take down Groups, Pages, and even Users supporting these movements . . . Bringing problematic content prevalence [conspiracies about COVID and vaccine discouragement] in the top 2% of communities in line with average communities could reduce the size of the overall problem by up to ~80%. Based on our experiences with US 2020 election delegitimization, QAnon, and Dangerous Orgs, there are at least three conceptual roles we need to address . . ."*[56]

31. Only after and in response to the Capitol protest and public pressure did Facebook take certain steps related to the above items on a temporary basis, such as:

---

[55] ▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮ *Badge Post - DS Misinfo*, p. 2-3.
[56] ▮▮▮▮▮▮▮▮▮▮▮▮ *Harmful Non-Violating Narratives*, p. 8, 10, 15, 19, 20.

*"**Immediate response to V&I** [violence and inciting] risk in DC [ ] US2020 Levers, **Previously Rolled Back** . . . Increased [ ] ability to action more . . . Applied demotion to content deemed likely to violate our Community Standards in the areas of hate speech, graph violence, and violence & incitement . . Made users less likely to see content we expect them to report . . . Require Admins review and approve posts in civic groups that accumulate four strikes . . ."*

"Reduced the strength of a LIVE video boost that was contributing to ultra-rapid virality for low-quality videos . . . removed 10 product boosts for civic and health content . . . Capped the strength which videos could be boosted . . . Explicitly include delegitimizing terms (ex. stop the steal) into the civic classifier . . . Lower[ed] threshold to demote hateful comments . . . Dropped the Groups invite rate limits from 100 to 30 to address evidence of concerning groups building growth through mass invitations . . ."[57]

32. Internal analysis also noted:

*"Early January 2021 had three events on the political calendar that deserved attention . . . **we anticipated new, narrow-topic misinfo stories to appear and spread rapidly** . . ."[58]*

*"IG [Instagram] 1/6 Capitol Hill Violence Data Update . . . **Many of the top reported posts for V&I [violence and incitement] were from Donald Trump** or were videos of Donald Trump. Several of the top [reported] posts called for [violence], suggested the overthrow of the government would be [desirable] . . Top Reported User . . . President Donald J. Trump [with 121 reports, compared with the second highest at 46 reports]."[59]*

*"[W]e should have adapted already long ago . . . There were **dozens of Stop the Steal groups active up until yesterday** . . . With the unprecedented resources we have, we should do better."*

*"We've been **fueling this fire for a long time** and we shouldn't be surprised it's now out of control."*

*"[D]o you genuinely think 24 hours is a meaningful ban? You mention the list of things we've changed in the past few years **but how are we expected to**

[57]   ▇▇▇▇ *Capitol Riots Breaks the Glass*, pp 1 - 4.
[58]   ▇▇▇▇ *Response to events in early January 2021 - lessons for algorithmic interventions*, p. 3-4.
[59]   ▇▇▇▇ *IG Jan 6 Capitol Riot Violence Update* p. 1-2.

ignore when leadership overrides research based policy decisions to better serve people like the groups inciting violence today.. Rank and file workers have done their part to identify changes to improve our platform but have been actively held back."

"*A 24-hour ban is nothing for Trump. His proxies on social media can easily keep the fires going during that time, and he will return . . .*"

"*How about refusing to take another dime of Trump's advertising money? I do acknowledge that a 24-hour ban is a pretty big deal, but that's only because up until now, our response has been completely tepid.*"[60]

33. Facebook acknowledged that how they resolved tradeoffs was dangerous, which is why they chose more conservative solutions for the US 2020 election. In contrast to their claims to their advertisers that they done all they could do to prevent the insurrection, they reverted back to these safer defaults only after the Insurrection flared up:

"Action: CAP LVEQ (Live Video EQ) boost linearly for civic videos at max 83.5x instead of 850x - Return to original US2020 threshold Description: Reduced the strength of a LIVE video boost that was contributing to ultra-rapid vitality for low-quality videos."[61]

"US2020 [election] Levers, Previously Rolled Back...

ID	Name	Description
RE45	**V&I (Violence and Inciting):** Cluster V&I ignores and drive specialist (PResc) Review at the cluster tlevel	Increased our ability to action more similar content in bulk, based on SME (subject matter expert) review
PE1	Use classifiers to demote content starting with a 5% demotion for content that has a 5% chance of violating. - English hate speech - Spanish hate speech - Graphic violence - Violence & Incitement	Applied demotion to content deemed likely to violate our Community Standards in areas of hate speech, graphic violence, and violence & incitement; the demotion's strength is in proportion to how confident we are that content is violating.

[60] ▮▮▮ *Comments on Zuckerberg's response to capitol riots*, p. 4, 5, 8.
[61] ▮▮▮ *Capitol Riots Breaks the Glass*, pp 4.

RE49	V&I : Increase demotion strength cap from 50% to 80%	Increased the max strength of linear probable violating V&I demotion (PE1), allowing us to further downrank higher confidence (50-80% confidence) V&I content.
RE48	V&I: Increase strength of p(report) to reduce visibility of V&I posts likely to be reported	Made users less likely to see content we expect them to report to Facebook, addressing subjectively problematic content.
PE20	**Turned on mandatory post approval** (MPAA) for Civic Groups with 8 violations	Require Admins review and approve posts in civic groups that accumulate four strikes (strikes require admins to have created or approved the violating content) as well as 57 fast-moving new groups, helping ensure that Admins are properly policing and being held accountable for serious violations during the election.
PE17	**Freeze Commenting on posts in groups that start to have a high rate of hate speech and violence and incitement** comments.	Prevents commenting on Groups posts that show a high rate of hate speech and violence incitement in associated comments, addressing concern that group Admins and members are "baiting" violating material and coordination of harm via otherwise non-violating posts.
RE58	V&I: Add reduced criteria to trigger comment auto-disable in group comment threads (extension of PE17)	Updated pre-emptive lever to disable comments in group threads (PE17) that were starting to have too much hate speech or violence & incitement to better catch threads early on.
RE54	Prevent groups from changing their names to delegitimizing terms	Prevents groups from changing names to terms associated with recent attempts to delegitimize the US election results (Stop the Steal), helping to slow down abusive audience building.
RE53	Preventing pages from changing their names to delegitimizing terms	Prevent Abusive Audience Building by re-naming and re-purposing/growth existing Pages associated with recent attempts to delegitimize the US election results (Stop the Steal)
RE57	Filter delegitimizing entities from recommendations	Filter entities with names matching "stop the steal" regex from recommendation surfaces across FB inc.

RE46	SEt up a special re-review queues for high risk contents from our user reports with Risk & Response	Addressed challenges with enforcing existing scaled policy and a spike in violations of escalations-only subsets of the V&I policy by re-reviewing and deleting more user reports.
RE44	V&I: reduce reactive auto-delete threshold from P90 to P70	Increased the volume of user reported content on the platform we were auto-deleting for violating our V&I policy, allowing us to manage the influx of reports.

"[62]

34. **Invoking immunity under Section 230:** Facebook has attempted to claim "immunity" for content under Section 230 of the Communications Decency Act, as it did in a recent court case, arguing:

> *"Section 230 requires dismissal of lawsuits like plaintiffs' [victims of sex trafficking on Facebook and Instagram] that seek to impose liability on interactive computer services for harmful content posted or communicated on their platforms by third parties . . . Section 230's plain text makes clear that it immunizes interactive service provides [i.e., Facebook] from suit, not just from liability."[63]*

KNOWLEDGE OF MATERIALITY

35. **Facebook has admitted investor risks with these issues.** For instance, Facebook's recent 10-K[64] acknowledged:

> *"**Our brands may also be negatively affected by the actions of users that are deemed to be hostile or inappropriate to other users**, by the actions of users acting under false or inauthentic identities, by the use of our products or services to disseminate information that is deemed to be misleading (or intended to manipulate opinions) . . ."*

36. Likewise, Facebook's 10-Q[65] stated:

---

[62] ▮▮▮▮ *Capitol Riots Breaks the Glass,* pg. 1-3.
[63] Facebook's Petition for Writ of Mandamus, *In re Facebook, Inc. and Facebook, Inc. d/b/a Instagram,* Cause Nos. 2018-69816 & 2018-82214 et seq., (Supreme Court of Texas, Filed May 29, 2020).
[64]https://sec.report/Document/0001326801-21-000014/.
[65] https://sec.report/Document/0001326801-21-000049/.

*"In addition, we have been, and may in the future be, subject to **negative publicity in connection with our handling of misinformation and other illicit or objectionable use of our products or services**, including in connection with the COVID-19 pandemic and elections in the United States and around the world. **Any such negative publicity could have an adverse effect on the size, engagement, and loyalty of our user base and marketer demand for advertising on our products, which could result in decreased revenue and adversely affect our business and financial results**, and we have experienced such adverse effects to varying degrees."*

*"For example, we have been the subject of significant media coverage involving concerns around **our handling of political speech** and advertising, hate speech, and other content, and we continue to receive negative publicity related to these topics. . . . **Any such negative publicity could have an adverse effect** on the size, engagement, and loyalty of our user base and marketer demand for advertising on our products, **which could result in decreased revenue and adversely affect our business and financial results**, and we have experienced such adverse effects to varying degrees from time to time."*

37. **Role for the SEC.** The SEC is charged with enforcing the laws that protect investors in public companies like Facebook. Facebook's investors care about misrepresentations and omissions by Mark Zuckerberg and other Facebook executives on the topic of misinformation relating to the insurrection for two reasons. First, to the extent that users become aware of the dangers that Facebook platforms present, they are likely to use the platforms less, leading to lower advertising revenue and lower profits. Second, some investors simply will not want to invest in a company that facilitates misinformation and violence on and off of the internet and then engages in misstatements and omissions on the topic.

38. Whistleblower Aid is a non-profit legal organization that helps workers report their concerns about violations of the law safely, lawfully, and responsibly. We respectfully request the SEC's assistance ensuring that our client never faces retaliation.

WHISTLEBLOWER AID

Report government and corporate lawbreaking.
Without breaking the law.

39. On information and belief, none of the documents enclosed here constitute attorney-client communications, were obtained during a meeting with an attorney, or otherwise indicate that they are in any way privileged.

40. We plan to continue supplementing this disclosure with additional information and evidence. Our client would be happy to meet with investigators at your convenience. Please feel free to contact us using the information below.

41. We are representing an anonymous whistleblower who is making the above disclosures solely for reporting the suspected violation of laws as outlined.
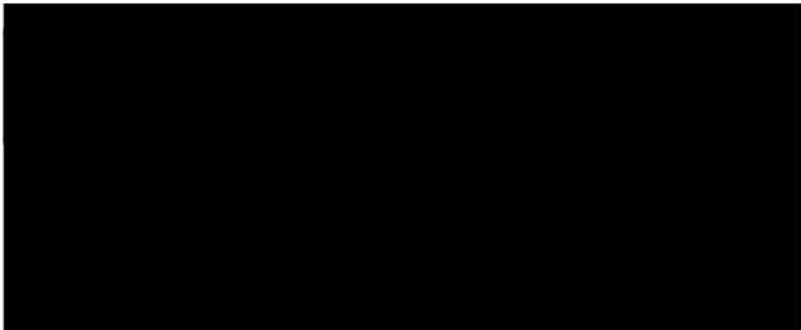
Sincerely,

John N. Tye, Attorney at Law
Chief Disclosure Officer

Andrew Bakaj, Attorney at Law
Of Counsel

WHISTLEBLOWER AID

Report government and corporate lawbreaking.
Without breaking the law.

██████████████████████████

Enclosures: ███████████ Facebook documents including —

███████ Reshare depth exp in Myanmar
███████ Capitol Protest Break The Glass
███████ At Risk Countries Planning
███████ Comments on Zuck's response to capitol riots
███████ Capitol Riots Break the Glass
███████ IG Jan 6 Capitol Riot Violence Update

████████████████████

███████ Harmful Topic Communities
███████ Coordinated Social Harm
███████ Misinfo got worse in 2019
███████ Harmful Conspiracies
███████ Lotus Mahal
███████ Adversarial Harmful Networks - India Case study.
███████ What is Collateral damage?
███████ Andrew Bosworth Thoughts for 2020
███████ Measuring Human Perception to Defend
Democracy
███████ Diverse Engagement May Identify Valuable Civic
Comments
███████ Misinfo in comments
███████ They used to post selfies, now they're trying to
reverse the election
███████ Comparing the effects of misinfo from politicians vs
ordinary user sources

██████ 2020-08-18 Political Influence on Content Policy

██████ Serial misinfo and hate offenders

██████ Effects of Politician Shared Misinformation

██████ Reshare ranking exp in India: Indonisia

██████ Reshare depth india and indonesia.

██████ Deep Reshares and Misinformation

██████ Max Reshare Depth experiment

██████ Further reading from Reshare Depth Article

██████ Fighting high harm misinfo with deep reshare damping

██████ Reshare depth by country.

██████ Harmful Non-Violating Narratives

██████ Killswitch Plan for all Group Recommendation Surfaces

██████ Subpopulations: Segmentations Wiki

██████ Sociographic Segments may be impactful for hate speech and voter suppression

██████ Narrowcast Misinfo Prevalence Update

██████ Identifying Narrowcast Misinfo Hard because missing or miscalibrated doubt

██████ Political Party response to the '18 Algorithm change

██████ Demotions in practice

██████ Misinfo narrowcasting on Instagram Home

██████ Civic Targeted Risk Scores

██████ Virality Reduction as Integrity Strategy.pdf

████████████████ Badge Post - DS Misinfo

██████ Political Influence Content Policy

██████ Update On Political Publisher Issues

██████ Employee concerns on political misinfo policies

██████ RTB Opex Review

██████ Response to events in early January 2021 - lessons for algorithmic interventions

██████ Mark Feedback on Soft Action Proposal + Deck presented to Mark

██████ CFO Lookback and Forward

██████ Random ██████

█████████ Polarization H2 Retro

█████████ Stop the Steal and Patriot Party

█████████ Civic Summit Q1 2020

█████████ Crisis Detection Pillar Lookback

█████████ False Choices - Take down or leave up

█████████ Mary Beth Hunzaker Badge Post - DS Misinfo

█████████ Hate Speech Cost Controls

█████████ Super Consumers of Misinfo

█████████ Users that repeatedly share misinformation

█████████ Deamplify content from those who repeatedly post misinformation

█████████ Misinfo User Repeat Offender

█████████ Sizing the opportunity of Repeat Offender Strikes on Page Admins

█████████ Reduce and Inform - Surfacing Repeatedly Fact-checked Hoaexs with CIRD and Regular Expressions

█████████ Problematic Non-violating Narratives document, p. 7.

█████████ Carol's Journey to QAnon

█████████ Karen & the Echo Chamber of Reshares - A test User Study of Misinfo and polarization risks encountered through recommendations - p2 (liberal)

█████████ Quantifying the concentration of reshares - and their VPVs among users

█████████ "Filtering out engagement-bait, bullying and excessive comments from MSI Deltoid metric 11/1 onwards