WHISTLEBLOWER AID

Report government and corporate lawbreaking.
Without breaking the law.

—ANONYMOUS WHISTLEBLOWER DISCLOSURE—

[redacted]

SEC Office of the Whistleblower
Via Online Portal & Fax

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

Facebook misled investors and the public about "transparency" reports boasting proactive removal of over 90% of identified hate speech when internal records show that "as little as 3-5% of hate" speech is actually removed.

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. **Summary.** Despite reassuring investors and the public that it proactively removes an overwhelming majority of hate speech, this is a material misstatement and/or omission. In fact, internal records show that only a small single-digit minority of such content is identified and removed, causing significant and long-term risks to Facebook and its investors.

3. **Facebook's Definition of "Hate Speech" in its Community Standards States:**

REDACTED FOR CONGRESS

> *"[W]e don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence. We define hate speech as a direct attack against people . . . on the basis of . . . protected characteristics . . ."*[1] (emphasis added)

4. **Facebook's Material Misrepresentations and Omissions on Hate Speech Include Mark Zuckerberg's Sworn Testimony to Congress.**

5. For example, on October 28, 2020, Mark Zuckerberg testified before the Senate:

> *"[I]n general, for each category of harmful content, whether it's terrorist propaganda or incitement of violence and hate speech, we have to build specific systems and specific AI systems. And one of the* ***benefits of I think having transparency and transparency reports*** *into how these companies are doing is that we have to report on a quarterly basis how effectively we're doing at finding those types of contents so you can hold us accountable for how nimble we are. . .* ***what our transparency reports show is that . . . we are proactively identifying, I think it's about 94% of the hate speech that we ended up taking down, and the vast majority of that before people even have to report it to us****."* (emphasis added)[2]

6. In the same hearing, Senator Markey asked about Facebook's response "*when President Trump posted on Facebook that, 'When the looting starts, the shooting starts,' . . . and told a hate group to, quote, 'stand by*.'" Mark Zuckerberg responded:

> "*Incitement of violence is against our policy and there are not exceptions to that, including for politicians.*"

7. Further, in March 2021, Mark Zuckerberg continued to represent:

> *"[T]he prevalence of hateful content people see on our service is less than 0.08 percent. . . Our enforcement effort in Groups demonstrates our commitment to keeping content that violates these policies off the platform.* ***In September, we shared that over the previous year we removed about 1.5 million pieces of content in Groups for violating our policies on organized hate, 91 percent of which we found proactively. We also removed about 12 million pieces of content in Groups for violating our***

[1]https://www.facebook.com/communitystandards/hate_speech.
[2]https://www.rev.com/blog/transcripts/tech-ceos-senate-testimony-transcript-october-28.

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

policies on hate speech, 87 percent of which we found proactively."[3] (emphasis added)

8. **Facebook Has Also Made Misstatements to Shareholders and Investors.**

9. For example, in its Second Quarter 2020 Follow Up Call, Facebook's CFO Dave Wehner reassured investors:

> *"And so we are trying to, I think, hold our values, which are we're a platform for free expression, but also have clear community standards and guidelines that we enforce fairly against. . . And there* ***we tend to be very focused on categories of things like hate speech that would cause imminent harm*** *. . . ."*[4] (emphasis added)

10. Notably, in the 2021 Notice of Annual Meeting and Proxy Statement,[5] shareholders made a proposal to assess the type of actions put in place during the 2020 election cycle to reduce false and divisive information, highlighting:

> *"****The Facebook brand has been diminished in recent years due to the platform's use as a tool for gross disinformation, hate speech, and to incite racial violence****."* (emphasis added)

11. However, in rejecting the referenced shareholder proposal, Facebook asserted:

> *"****We believe that implementing [this] proposal is unnecessary due to our transparency efforts to date, the significant progress we continue to make to address these issues, and our work to help external researchers assess the impact of our efforts in this area*** *as they relate specifically to the U.S. 2020 elections. We work to remove harmful information that violates our policies on our platforms."*[6] (emphasis added)

12. In the same Notice and Proxy Statement, Facebook emphasized:

> *"More broadly,* ***we have taken meaningful action over the years to fight hate on our platforms.*** *We continue to make significant investments in teams and technology to proactively find and remove hate speech and hate organizations. . .* ***In our report for the fourth quarter of 2020, we***

[3]https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Wstate-ZuckerbergM-20210325-U1.pdf.
[4]https://s21.q4cdn.com/399680738/files/doc_financials/2020/q2/Q2'20-FB-Follow-Up-Call-Transcript.pdf.
[5]https://www.sec.gov/Archives/edgar/data/1326801/000132680121000022/facebook2021definitiveprox.htm.
[6]https://www.sec.gov/Archives/edgar/data/1326801/000132680121000022/facebook2021definitiveprox.htm.

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

> ***disclosed that we detected the majority of the content we removed from our platform before anyone reported it to us, and proactively detected about 97% of hate speech on Facebook that we removed before anyone reported it to us, among our progress in other areas.*** *We also disclosed that the prevalence of certain violating content decreased since the report for the third quarter of 2020, including hate speech . . . and violent and graphic content . . .* ***In our February 2021 Community Standards Enforcement Report, we reported that we proactively detected about 97% of hate speech content we removed on both Facebook and Instagram.*** ***We also reported that in the fourth quarter of 2020, the prevalence of hate speech on Facebook, which we define as the percentage of times people see this type of content, was between 0.07% to 0.08%****, or 7 to 8 views of hate speech for every 10,000 views of content. This represented a decrease from our reported prevalence of 0.10% to 0.11% for the third quarter of 2020."*[7] (emphasis added)

13. **Facebook Has Further Made Misstatements in its Public Pages and Reports.**

14. For instance, Facebook's Community Standards "Enforcement Report" for August 2019 (i.e., one of Facebook's quarterly "transparency reports" that it published with enforcement metrics along with a summary of highlights) announced:

> *"Over the last two years,* ***we've invested in proactive detection of hate speech*** *so that we can detect this harmful content before people report it to us and sometimes before anyone sees it . . . With these evolutions in our detection systems,* ***our proactive rate has climbed to 80%*** *. . . and we've increased the volume of content we find and remove for violating our hate speech policy."*[8] (emphasis added)

15. The Enforcement Report for August 2020 went further, stating:

> *"****Despite the impact of COVID-19****, improvements to our technology enabled us to take action on more content in some areas, and increase our proactive detection rate in others.* ***Our proactive detection rate for hate speech on Facebook increased 6 points from 89% to 95%****. In turn, the amount of content we took action on increased from 9.6 million in Q1 to 22.5 million in Q2 . . . On Instagram, our proactive detection rate for hate speech*

[7] https://www.sec.gov/Archives/edgar/data/1326801/000132680121000022/facebook2021definitiveprox.htm.
[8] https://about.fb.com/news/2019/11/community-standards-enforcement-report-nov-2019/.

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

increased 39 points from 45% to 84% and the amount of content we took action on increased from 808,900 in Q1 2020 to 3.3 million in Q2."[9]

16. Likewise, the Enforcement Report for November 2020 continued to emphasize:

> *"Due to our investments in AI, **we have been able to remove more hate speech and find more of it proactively before users report it to us**. . . On Facebook in Q3, **we took action on . . . 22.1 million pieces of hate speech content, about 95% of which was proactively identified** . . . On Instagram in Q3, we took action on . . . 6.5 million pieces of hate speech content . . . about 95% of which was proactively identified . . ."*[10] (emphasis added)

17. Most recently, in August 2021, Facebook's Enforcement Report represented:

> *"**Prevalence of hate speech has decreased for three quarters in a row** since we first began reporting it. **This is due to improvements in proactively detecting hate speech and ranking changes in News Feed**. . . . Hate speech content removal has increased over 15X on Facebook and Instagram since we first began reporting it. . . **Our proactive rate (the percentage of content we took action on that we found before a user reported it to us) is over 90%** for 12 out of 13 policy areas on Facebook and nine out of 11 on Instagram . . . Prevalence of hate speech on Facebook continued to decrease . . . In Q2 [2021], **it was 0.05%, or 5 views per 10,000 views**."*[11] (emphasis added)

18. Regarding these periodic enforcement updates, Facebook has also confirmed that:

> *"[T]ransparency is only helpful if the information we share is useful and accurate. In the context of the Community Standards Enforcement Report, that means the metrics we report are based on sound methodology and accurately reflect what's happening on our platform . . ."*[12]

19. In addition, after the Stop Hate for Profit campaign encouraged global businesses to pause paid advertising on Facebook in mid-2020, Facebook claimed:

[9] https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/.
[10] https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/.
[11] https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/.
[12] https://about.fb.com/news/2020/08/independent-audit-of-enforcement-report-metrics/.

REDACTED FOR CONGRESS

> [In July 2020] "*Through human review and the latest technologies - like advanced AI - **we proactively find nearly 90% of the hate speech content we remove from Facebook before anyone even reports it** . . Facebook Does Not Benefit From Hate . . . Facebook assessed 95.7% of hate speech reports in less than 24 hours . . . we find nearly 90% of the hate speech we remove before someone reports it . . .*" [And in 2021] "*We've made substantial investments to keep hate off our platform.*"[13] (emphasis added)

20. **Facebook's Records Confirm That Facebook's Statements Were False.**

21. For example, Facebook records from March 2021 state:

> "*[W]e estimate that we may action as little as 3-5% of hate and ~0.6% of V&I [violence and inciting content] on Facebook . . . This human reviewer uncertainty likely also contributes to lower precision and recall on classifiers.*"[14] (emphasis added)

22. Further, the documentation shows:

> "*[W]e do not . . . have a model that captures even a majority of integrity harms, particularly in sensitive areas . . . **[W]e only take action against approximately 2% of the hate speech on the platform. Recent estimates suggest that unless there is a major change in strategy, it will be very difficult to improve this beyond 10-20% in the short-medium term***."[15] (emphasis added)

23. There is also evidence that:

> "*[O]ur current approach of grabbing a hundred thousand pieces of content, paying people to label them as Hate or Not Hate, training a classifier, and using it to automatically delete content at 95% precision is just never going to make much of a dent . . . **we're deleting less than 5% of all of the hate speech posted to Facebook. This is actually an optimistic estimate**--previous (and more rigorous) iterations of this estimation exercise have put it closer to 3%, and on V&I we're deleting somewhere around 0.6% . . .we miss 95% of violating hate speech.*"[16] (emphasis added)

---

[13] https://www.facebook.com/business/news/sharing-actions-on-stopping-hate.
[14] [redacted] *Problematic Non-violating Narratives document*, p. 7.
[15] *Demoting on Integrity Signals*, p. 1, 9-10.
[16] [redacted] [redacted] *post,* p. 9-10.

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

24. While Facebook has argued that users are responsible for their content and that it is just a platform under Section 230 and other laws,[17] records demonstrate:

> *"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****."*[18] (emphasis added)

25. In fact, as one report summarizes:

> *"****Actively ranking content in News Feed and promoting content on recommendations . . . makes us responsible for any harm*** *caused by exposure to that content . . . we are responsible for harmful experiences on any surface where we actively present content . . . Currently, there are places where it appears* ***we could be doing more to prevent users from being exposed to harmful content, yet don't*** *[such as "demotions"] . . . There is also* ***more we could be doing to curb the spread of harmful and inauthentic viral content*** *[such as to curb "manufactured" virality through "reshares" i.e., re-posted viral content, which is] often a vector for misinformation and other integrity harms."*[19] (emphasis added)

26. In addition, contrary to representations about a "proactive" approach to removing hate speech, Facebook's records from at least 2019 show that:

> *"[I]n the case of hate speech we need to cut a significant amount of our current capacity in order to fund new initiatives. . .* ***Most of our review costs*** *today [~75%]* ***are still driven by reactive capacity*** *[compared with ~25% proactive capacity], and that's something we know we want to change . . ."*[20] (emphasis added)

27. In particular, one internal report states:

---

[17] *See e.g.,* https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Wstate-ZuckerbergM-20210325-U1.pdf; *In re Facebook, Inc.*, 625 S.W.3d 80 (Tex. 2021).

[18] [redacted] *What is Collateral damage?* p. 35.

[19] [redacted] *Pinfeed and Responsibility*, p. 1, 9-11.

[20] [redacted] *Hate Speech Cost Controls*, p.4-5.

REDACTED FOR CONGRESS

> *"We seem to be having a **small impact** in many language-country pairs **on Hate Speech** and Borderline Hate, **probably ~3%** . . . We are likely having little (if any) impact on violence."*[21] (emphasis added)

28. Likewise, in a study conducted for the Afghanistan market, it was determined that:

> *"98.8 per cent of the total Hate Speech tack [sic] downs are done by human reviewers . . . and **only 0.2 per cent is taken down by automation. While Hate Speech is consistently ranked as one of the top abuse categories** in the Afghanistan market, **the action rate for Hate Speech is worryingly low at 0.23 per cent.**"*[22] (emphasis added)

29. Relating to public interest in hate speech in the summer of 2020, reports show that:

> *"[For] Trump's post (our classifier was almost 90% certain that this post violated Facebook's Violence and Incitement policy . . . [there were] "**extremely high levels of violence and hate speech harm** . . . [and that] we saw a drastic 5X and 3X increase in user reports for violence and hate speech, respectively . . ."*[23] (emphasis added)

30. The issue reached international politics as well, as records relay:

> *"Political parties across Europe claim that **Facebook's algorithm change in 2018 (MSI) has changed the nature of politics. For the worse** . . . they feel that they have been forced to adapt to the change by producing far more negative content than before. . . Many parties. . . worry about the long-term effects on democracy. . . **they are trapped in an inescapable cycle of negative campaigning by the incentive structures of the platform** . . . evidence around how anger reactions, overall, is weaponized by political figures and creating negative incentives on the platform."*[24] (emphasis added)

31. Moreover, internal records outline how:

> *"Multiple offenders for Hate are frequently also multiple offenders for misinformation . . . We may be repeatedly applying authenticity verifications to some or many of these accounts to no effect . . . **99% of these user***

[21] [redacted] *A first look at the minimum integrity holdout*, p. 11, 15.
[22] *Afghanistan Hate Speech analysis*, p. 9.
[23] *Hate Begets hate and violence begets Violence (George Floyd)*, p. 2, 3, 7.
[24] *Political Party response to the '18 Algorithm change*, p. 4, 24, 26.

REDACTED FOR CONGRESS

accounts remain active, and some of them have passed dozens of authenticity checks."[25] (emphasis added)

32. As one example of a 90-day review, out of "279k" "Hate" users:

> *"99%+of user accounts remain activated."*[26]

33. Specifically regarding hate and inciting content in groups, documents outline:

> *"**From the earliest Groups, we saw high levels of Hate**, VNI [violence and inciting], and delegitimization, combined with meteoric growth rates. . . Amplifiers [users who are connected to many others of these vulnerable users] posted 98% more VNI and 40% more hate."*[27] (emphasis added)

34. For example, Facebook records show:

> *"US Civic groups can grow to hundreds of thousands of members real fast (Sometimes in just a few days). **Many of the top civic groups in the US are also full of hate speech**, violence incitement, misinfo, anti-vax content, etc. Many are created or controlled by foreign operations (usually financially motivated). **Our existing integrity systems aren't effectively addressing these issues . . . ~70% of the top 100 most active US civic groups are considered non-recommendable for issues such as hate**, misinfo, bullying and harassment."*[28] (emphasis added)

35. Despite the above, in September 2020, internal records recognized:

> *"Facebook currently has no firewall to insulate content-related decisions from external pressures. It could have one."*[29]

36. Likewise, Facebook documentation confirms:

> *"**Hate speech remains a developing problem area**."*[30] (emphasis added)

37. As far as the impact on higher-risk users:

---

25 [redacted] *Serial misinfo and hate offenders*, p.1.
26 *Serial misinfo and hate offenders*, p.3.
27 *Stop the Steal and Patriot Party*, p.3, 11.
28 *Dangerous Growth of Civic Groups*, p.3, 8.
29 *A firewall for Content Policy*, p.1.
30 *Hate Speech capacity reduction plan*, p. 1.

REDACTED FOR CONGRESS

> *"Respondents in at-risk countries rate hate-speech, dangerous organizations, and violence incitement content as more severe . . ."*[31]

38. **Facebook Has Publicly and Privately Admitted This is a Material Issue.**

39. In particular, in its recent Notice and Proxy Statement, Facebook conceded:

> *"**We agree that the amplification of false, divisive, hateful, and inciting content is harmful to our community**, and we continue to take steps to address this issue."*[32] (emphasis added)

40. In addition, in its 10-Q from 2021, Facebook recognized:

> *"[W]e 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** . . ."*[33] (emphasis added)

41. In its public statement after the Stop Hate for Profit ad boycott, Facebook said:

> *"The people using Facebook don't want to see hate on our services, our advertisers don't want to see it and we have no tolerance for it."*[34]

42. Finally, internal documents confirm, for example:

> *"[R]oughly 50% of Americans [were] worried about how angry, disrespectful, and uncivil political conversations on social media are, as compared to other places where people discuss politics."*[35]

43. **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 hate speech for two reasons. First, to the extent that

[31] [redacted] *User severity for hate speech*, p.1.
[32] https://www.sec.gov/Archives/edgar/data/1326801/000132680121000022/facebook2021definitiveprox.htm.
[33] https://sec.report/Document/0001326801-21-000049/.
[34] https://www.facebook.com/business/news/sharing-actions-on-stopping-hate.
[35] [redacted] *We are Responsible for Viral Content*, p. 2.

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

users and their parents 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 hate speech and then engages in misstatements and omissions on the topic.

---

44. 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.

45. 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.

46. 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.

47. 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
[redacted]

Andrew Bakaj, Attorney at Law
Of Counsel
[redacted]

REDACTED FOR CONGRESS

REDACTED FOR CONGRESS

[redacted]

[redacted]

[redacted]

Enclosures: Internal Facebook documents including —

[redacted] post
[redacted] Problematic Non-violating Narratives is an Archetype in need of Novel Solutions
[redacted] Hate Speech Cost Controls
[redacted] Hate Begets hate and violence begets Violence (George Floyd)
[redacted] Afghanistan Hate Speech Analysis
[redacted] Political Party response to the '18 Algorithm change
[redacted] Serial misinfo and hate offenders
[redacted] Stop the Steal and Patriot Party
[redacted] Dangerous Growth of Civic Groups
[redacted] A firewall for Content Policy
[redacted] 2019 Hate Speech capacity reduction plan
[redacted] User severity for hate speech in At-Risk Countries
[redacted] We are Responsible for Viral Content
[redacted] Adversarial Harmful Networks - India Case study
[redacted] A first look at the minimum integrity holdout
[redacted] Demoting on Integrity Signals is not enough
[redacted] What is Collateral damage?
[redacted] Pinfeed and Responsibility

REDACTED FOR CONGRESS