United States Senate

WASHINGTON, DC 20510

March 25, 2024

Mr. Mark Zuckerberg
Chief Executive Officer Meta, Inc.
1 Hacker Way
Menlo Park, CA 94025

Dear Mr. Zuckerberg,

We write to follow up on a letter Senator Warren sent you on December 14, 2023, describing the suppression of Palestinian and Palestinian-related content on Instagram and Facebook and urging Meta to provide additional transparency as to content moderation and to address discriminatory algorithmic design.[1] Meta's response, dated January 29, 2024, did not provide any of the requested information necessary to understand Meta's treatment of Arabic language or Palestine-related content versus other forms of content.[2] It is imperative that Meta provide this information so the American people and their elected representatives can understand the impact of Meta's policies on those communities and public debate.

Senator Warren's previous letter outlined allegations that Meta censored and mistranslated Palestinian and Palestinian-related content, applied uneven standards to Palestinian-generated content compared to the rest of the region and world, and disproportionately restricted and censored material and accounts linked to communications regarding Palestine.[3] It highlighted Meta's apparent failure to implement the recommendations of a 2020 Business for Social Responsibility (BSR) report, which found that Meta erroneously removed Arabic content on a higher per-user basis than content in Hebrew or English, hindering Palestinians' ability to share their perspectives and experiences.[4] It also warned of the link between Meta's content moderation practices and the risk of hate-inspired violence.[5]

Recent developments have only deepened our concerns. On December 21, 2023, Human Rights Watch released a comprehensive report documenting Meta's suppression or removal of over 1,000 expressions of peaceful support of Palestinians on Instagram and Facebook over October and November 2023,

---

[1] Letter from Senator Elizabeth Warren to Mark Zuckerberg re. Suppression of Palestinian Content on Instagram, December 13, 2023, https://www.warren.senate.gov/imo/media/doc/Meta%20Letter.pdf.
[2] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 3, on file with the Office of Elizabeth Warren.
[3] Letter from Senator Elizabeth Warren to Mark Zuckerberg re. Suppression of Palestinian Content on Instagram, December 13, 2023, https://www.warren.senate.gov/imo/media/doc/Meta%20Letter.pdf.
[4] BSR, "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, p. 5, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf.
[5] The Intercept, "Facebook Approved an Israeli Ad Calling for Assassination of Pro-Palestine Activist," Sam Biddle, November 21, 2023, https://theintercept.com/2023/11/21/facebook-ad-israel-palestine-violence/; 7amleh, "Meta Should Stop Profiting from Hate," November 21, 2023, https://7amleh.org/2023/11/21/metashould-stop-profitingfrom-hate; Amnesty International, "Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations," September 29, 2022, https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systemspromoted-violence-againstrohingya-meta-owes-reparations-new-report/.

impacting users in 60 countries around the world.[6] In early January 2024, Meta opened an investigation into one of its own employees for circulating a letter demanding transparency regarding Meta's censorship of Palestinian content inside the company and on its platforms.[7] On January 30, 2024, one day after Meta responded to Senator Warren's office, Meta emailed several civil society groups to announce it is revisiting its hate speech policy in relation to the term "Zionist."[8] A Meta representative justified the review by saying that, though "the term 'Zionist' often refers to a person's ideology, which is not a protected characteristic, it can often be used to refer to Jewish or Israeli people."[9] Meta did not explain the contours of the new policy or provide detail on the risk of stifling legitimate expression about political ideology or state policies. These developments underscore the urgent need for improved transparency regarding censorship of information on Instagram and Facebook, two of the world's largest social media giants.[10]

In light of reports that Meta took down posts in Arabic, but not identical English or Hebrew versions, Senator Warren asked Meta to disclose how many posts from the region it had taken down since October 7, 2023 in Arabic, Hebrew, and English respectively.[11] While Meta's response acknowledged the company's heightened censorship of content in the aftermath of Hamas' October 7, 2023 attacks—disclosing that Meta "removed or marked as disturbing more than 2,200,000 pieces of content in Hebrew and Arabic" in the nine days following the attacks– it did not differentiate based on language or region.[12]

Meta admitted to "lower[ing] the confidence level at which [it] automatically take[s] action" on potentially violative content after the October 7 attack, but didn't acknowledge that it had applied a lower threshold to content originating in occupied Palestinian territories, as compared to the rest of the

[6] Human Rights Watch, "Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook," December 21, 2023, https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and; see also Access Now, "It's not a glitch: how Meta systematically censors Palestinian voices," Marwa Fatafta, February 19, 2024, https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/.
[7] Financial Times, "Meta staffer under investigation after claiming 'censorship' of pro-Palestinian views," Hannah Murphy and Cristina Criddle, January 9, 2024, https://www.ft.com/content/a4381d91-2fec-4422-b04b-75a06b643a05.
[8] The Guardian, "Meta's review of hate speech policy sparks concern of further censorship of pro-Palestinian content," Johana Buiyan and Kari Paul, February 9, 2024, https://www.theguardian.com/technology/2024/feb/09/meta-hate-speech-policy-zionist-censorship-pro-palestine-content; The Intercept, "Meta Considering Increased Censorship of the Word 'Zionist,'" February 8, 2024, https://theintercept.com/2024/02/08/facebook-instagram-censor-zionist-israel/.
[9] The Intercept, "Meta Considering Increased Censorship of the Word "Zionist," Sam Biddle, February 8, 2024, https://theintercept.com/2024/02/08/facebook-instagram-censor-zionist-israel/.
[10] Statista, "Most popular social networks worldwide as of January 2024, ranked by number of monthly active users," Stacy Jo Dixon, February 2, 2024, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
[11] TechCrunch, "Meta has a moderation bias problem, not just a 'bug,' that's suppressing Palestinian voices," Morgan Sung, October 19, 2023, https://techcrunch.com/2023/10/19/meta-instagram-palestine-israel-shadowbancensorshipmoderation-bias/; 17 Vox, "Why some Palestinians believe social media companies are suppressing their posts," A.W. Ohlheiser, October 29, 2023, https://www.vox.com/technology/23933846/shadowbanningmeta-israel-hamas-war-palestin; Letter from Senator Elizabeth Warren to Mark Zuckerberg re. Suppression of Palestinian Content on Instagram, December 13, 2023, https://www.warren.senate.gov/imo/media/doc/Meta%20Letter.pdf; Letter from Senator Elizabeth Warren to Mark Zuckerberg re. Suppression of Palestinian Content on Instagram.

[12] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 3, on file with the Office of Elizabeth Warren.

region.[13] Meta also disclosed that it "allow[s] otherwise policy-violating content when its public interest value outweighs the risk of harm," and that the company has "granted very limited exceptions for content related to the Israel-Hamas War," but did not provide the number of or rationale for these exceptions.[14]

Public reports suggest Meta's practices *are* biased. A 2022 BSR report found that Meta erroneously removed more Arabic than Hebrew content on a per user basis during an outbreak of violence in May 2021.[15] Meta generally hides comments designated as hateful only when its content moderation systems are 80% certain that they violate the platform's policies, but it lowered that threshold for users in Palestinian territories to 25% following the October 7 attacks.[16] Meta's letter pointed to its use of machine learning classifiers "to identify harmful content and automatically action content."[17] The letter omitted, however, that Meta had admitted internally in October 2023 that it had not been using its classifier for hostile Hebrew speech on Instagram because the classifier was not functioning adequately.[18]

Content removal has serious implications and is difficult for users to reverse. Meta directs users with removed content to appeal, but makes the process difficult or even inaccessible.[19] Human Rights Watch documented over 300 cases in which users were unable to appeal a restriction on their Instagram or Facebook account, because the buttons to do so did not work, did not lead anywhere when clicked, or were disabled.[20] And users who manage to submit an appeal have no guarantee that Meta's Oversight Board will agree to review it.[21] Users who don't receive such review may be subject to overreaching content removal, as demonstrated by the fact that the Oversight Board overturned both of the content moderation decisions related to the Israel-Gaza conflict that were selected for expedited review.[22]

---

[13] *Id*. p. 6.
[14] *Id*. p. 3.
[15] BSR, "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, p. 5, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf.
[16] Wall Street Journal, "Inside Meta, Debate Over What's Fair in Suppressing Comments in the Palestinian Territories," Sam Schechner et al., October 21, 2023, https://www.wsj.com/tech/inside-meta-debate-over-whatsfairin-suppressing-speech-in-the-palestinian-territories-6212aa58; The Guardian, "Instagram apologises for adding 'terrorist' to some Palestinian user profiles," Josh Taylor, October 19, 2023, https://www.theguardian.com/technology/2023/oct/20/instagram-palestinian-user-profile-bios-terrorist-added-translation-meta-apology.
[17] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 6, on file with the Office of Elizabeth Warren.
[18] "Inside Meta, Debate Over What's Fair in Suppressing Comments in the Palestinian Territories," Sam Schechner, Jeff Horwitz, and Newley Purnell, October 21, 2023, https://www.wsj.com/tech/inside-meta-debate-over-whats-fair-in-suppressing-speech-in-the-palestinian-territories-6212aa58.
[19] Human Rights Watch, "Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook," December 21, 2023, https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and.
[20] *Id*.
[21] Meta, "How to appeal to the Oversight Board," February 22, 2024, https://transparency.fb.com/oversight/appealing-to-oversight-board/; Meta Oversight Board, "Hostages Kidnapped from Israel," https://www.oversightboard.com/decision/FB-M8D2SOGS/; Meta Oversight Board, "Al-Shifa Hospital," https://www.oversightboard.com/decision/IG-WUC3649N/. In both cases, the Board overturned Meta's original decision to remove the content from Instagram. It found that restoring the content, with a "mark as disturbing" warning screen, is consistent with Meta's content policies, values and human-rights responsibilities.
[22] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 4, on file with the Office of Elizabeth Warren.

These concerns are not new. For years, civil society organizations have called on Meta to address systems and practices that result in disproportionate censorship of Palestinians, including unwarranted content removals during Israel's attacks on Gaza in 2021.[23] It is deeply troubling that Meta refuses to acknowledge and address the discriminatory nature of its content removal practices, despite being presented with a plethora of evidence to this effect.

Social media users deserve to know when and why their accounts and posts are restricted, and to receive protection against discrimination and hate speech.[24] Meta's response reveals the company's unwillingness to explain how and why these decisions appear to be having a discriminatory impact.[25] This lack of answers underscores the need for meaningful regulation of the largest tech platforms. Legislative proposals like the bipartisan *Digital Consumer Protection Commission Act* would increase transparency around algorithmic decision making on major online platforms.[26] As the catastrophic situation in Gaza continues to escalate and a ceasefire is further delayed, there is a need now more than ever for such legislation.

Given your failure to provide answers to important questions, we again ask you to provide the following information by April 8, 2024:

1. What steps has Meta taken to implement the recommendations in BSR's September 2022 report titled "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine"?[27] For each recommendation, please specify:
    a. Whether Meta has pursued or is pursuing actions related to the recommendation. If yes, please specify the nature of those actions.
    b. What resources, in terms of staff time and expenses, Meta has allocated toward addressing the recommendation.
    c. What metric(s) Meta is using to define success with regards to the recommendation.
    d. When Meta will complete action related to the recommendation.

2. Please list each instance, in the past five years, in which Meta has changed the content moderation threshold for a particular nation or occupied territory in the manner described in the October 21, 2023 Wall Street Journal article titled "Inside Meta, Debate Over What's Fair in Suppressing Comments in the Palestinian Territories."[28]

---

[23] Human Rights Watch, "Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook," December 21, 2023, https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and.

[24] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, on file with the Office of Elizabeth Warren.

[25] *Id*.

[26] Digital Consumer Protection Commission Act of 2023, S. 2597, https://www.congress.gov/bill/118th-congress/senate-bill/2597/text?s=1&r=4&q=%7B%22search%22%3A%22open+app+market%22%7D#:~:text=Introduced%20in%20Senate%20(07%2F27%2F2023)&text=To%20amend%20the%20Clayton%20Act,%2C%20privacy%2C%20and%20national%20security.

[27] BSR, "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," September 2022, p. 5, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf.

[28] "Inside Meta, Debate Over What's Fair in Suppressing Comments in the Palestinian Territories," Sam Schechner, Jeff Horwitz, and Newley Purnell, October 21, 2023, https://www.wsj.com/tech/inside-meta-debate-over-whats-fair-in-suppressing-speech-in-the-palestinian-territories-6212aa58.

a. For each instance, please list the location, the timespan during which the threshold was changed, the language to which the threshold change applied, the reason for the threshold change, and the level to which the threshold was set.
b. For each instance, please list the number of posts that were flagged or otherwise affected as a result of the new threshold.

3. During the time period spanning October 7, 2023 to the present:
    a. How many Arabic language posts originating from Palestine have been removed?
    b. What percentage of total Arabic language posts originating from Palestine does the above number represent?
    c. What percentage of the removed posts were removed due to automated systems versus human moderation?
    d. How often did Meta limit the reachability of Arabic language posts originating from Palestine without notifying the user?
    e. How often did Meta limit the reachability of Arabic language posts originating from Palestine while notifying the user?

4. During the time period spanning October 7, 2023 to the present:
    a. How many English language posts originating from Palestine have been removed?
    b. What percentage of total English language posts originating from Palestine does the above number represent?
    c. What percentage of the removed posts were removed due to automated systems versus human moderation?
    d. How often did Meta limit the reachability of English language posts originating from Palestine without notifying the user?
    e. How often did Meta limit the reachability of English language posts originating from Palestine while notifying the user?

5. During the time period spanning October 7, 2023, to the present:
    a. How many Hebrew language posts originating from Israel have been removed?
    b. What percentage of total Hebrew language posts originating from Israel does the above number represent?
    c. What percentage of the removed posts were removed due to automated systems versus human moderation?
    d. How often did Meta limit the reachability of Hebrew language posts originating from Israel without notifying the user?
    e. How often did Meta limit the reachability of Hebrew language posts originating from Israel while notifying the user?

6. During the time period spanning October 7, 2023, to the present:
    a. How many English language posts originating from Israel have been removed?
    b. What percentage of total English language posts originating from Israel does the above number represent?
    c. What percentage of the removed posts were removed due to automated systems versus human moderation?

d. How often did Meta limit the reachability of English language posts originating from Israel without notifying the user?
e. How often did Meta limit the reachability of English language posts originating from Israel while notifying the user?

7. During the time period spanning October 7, 2023 to the present:
    a. a. How many appeals did users submit regarding content decisions related to Arabic language posts originating from Palestine?
    b. b. How often were content decisions regarding Arabic language posts originating from Palestine appealed?

8. During the time period spanning October 7, 2023 to the present:
    a. How many appeals did users submit regarding content decisions related to English language posts originating from Palestine?
    b. How often were content decisions regarding English language posts originating from Palestine appealed?

9. During the time period spanning October 7, 2023 to the present:
    a. How many appeals did users submit regarding content decisions related to Hebrew language posts originating from Israel?
    b. How often were content decision regarding Hebrew language posts originating from Israel appealed?

10. During the time period spanning October 7, 2023 to the present:
    a. How many appeals did users submit regarding content decisions related English language posts originating from Israel?
    b. How often were content decisions regarding English language posts originating from Israel appealed?

11. During the time period spanning October 7, 2023 to the present:
    a. How many appeals did users submit globally regarding content decisions?
    b. How often were content decisions appealed globally?

12. What is the average response time a user can typically expect after appealing a content moderation decision from Meta?

13. During the time period spanning October 7, 2023 to the present:
    a. What was the average response time for a user appeal of a content moderation decision for Arabic language posts originating from Palestine?
    b. Based on data from user appeals, what percentage of Arabic language posts originating from Palestine were found to have been wrongfully taken down (i.e. false positives) and then reinstated?

14. During the time period spanning October 7, 2023 to the present:
    a. What was the average response time for a user appeal of a content moderation decision for English language posts originating from Palestine?

b. Based on data from user appeals, what percentage of English language posts originating from Palestine were found to have been wrongfully taken down (i.e. false positives) and then reinstated?

15. During the time period spanning October 7, 2023 to the present:
    a. What was the average response time for a user appeal of a content moderation decision for Hebrew language posts originating from Israel?
    b. Based on data from user appeals, what percentage of Hebrew language posts originating from Israel were found to have been wrongfully taken down (i.e. false positives) and then reinstated?

16. During the time period spanning October 7, 2023 to the present:
    a. What was the average response time for a user appeal of a content moderation decision for Hebrew language posts originating from Israel?
    b. Based on data from user appeals, what percentage of English language posts originating from Israel were found to have been wrongfully taken down (i.e. false positives) and then reinstated?

17. A Meta spokesperson stated that Instagram hid comments containing the Palestinian flag emoji due to Meta's Dangerous Organizations and Individuals policy and community standards.[29] However, the *Intercept* reportedly reviewed several hidden comments containing the Palestinian flag emoji that had no reference to dangerous organizations or individuals and did not otherwise appear to violate Meta's community standards.[30] Is the Intercept report accurate? If not, what is the explanation for the reported hiding of comments?

18. During the time period spanning October 7, 2023 to the present:
    a. How many staffers does Meta employ to review Arabic language content? Of that number, how many staffers possess native language skills?
    b. How many staffers does Meta employ to review English language content? Of that number, how many staffers possess native language skills?
    c. How many staffers does Meta employ to review Hebrew language content? Of that number, how many staffers possess native language skills?

19. What is Meta's current policy for retention of content that may contain evidence of human rights abuses?
    a. What are the criteria for initiating and terminating preservation of content that may contain evidence of human rights abuses?
    b. What is the period of data retention for content that may contain evidence of human rights abuses?
    c. What are Meta's policies for researcher, civil society, and governmental access to content that may contain evidence of human rights abuses?

[29] The Intercept, "Instagram Hid a Comment. It Was Just Three Palestinian Flag Emojis." Sam Biddle, October 28, 2023, https://theintercept.com/2023/10/28/instagram-palestinian-flag-emoji/.
[30] *Id*.

In addition to the aforementioned questions that you previously ignored, please provide the following information regarding your most recent correspondence with Senator Warren's office by April 8, 2024:

1. Your letter references a Meta policy in which Meta makes exceptions for otherwise policy-violating content "when its public interest value outweighs the risk of harm."[31]
    a. Is there any formal process by which these exceptions are made? If so, please describe the process.
    b. In how many instances has content been granted this exception since October 7, 2023?
        i. For each instance, please list the location, the language to which the exception applied, and the reason for the exception.
        ii. For each instance, please list the number of posts that were flagged or otherwise affected as a result of the exception.

2. Your letter cites two expedited appeals.
    a. On what basis does Meta refer a case for expedited review?
    On what basis does the Oversight Board accept a case for expedited review?

3. Your letter states that Meta is "working with third-party fact-checkers in the region."[32] Does Meta directly employ any fact checkers? If yes, how many fact checkers does Meta employ?

Thank you for your attention to this important matter.

Sincerely,

Elizabeth Warren
United States Senator

Bernard Sanders
United States Senator

[31] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 2, on file with the Office of Elizabeth Warren.
[32] Response Letter from Meta to Senator Elizabeth Warren, January 29, 2024, p. 3, on file with the Office of Elizabeth Warren.