July 13, 2020

Senator Mark Warner
703 Hart Senate Office Bldg.
Washington, DC 20510

Senator Mazie Hirono
713 Hart Senate Office Bldg.
Washington, DC 20510

Senator Bob Menendez
528 Hart Senate Office Bldg.
Washington, D.C. 20510

Dear Senators,

Thank you for your letter of June 25 regarding hate speech and white supremacy online.

As VP of Global Policy and Communications Nick Clegg noted in a recent op-ed,[1] when society is divided, those divisions play out on social media. More than 100 billion messages are sent on our services every day, and in all of those billions of interactions, a tiny fraction are hateful. When we find hateful posts on Facebook and Instagram, we take a zero-tolerance approach and remove them. We invest billions of dollars each year in people and technology to keep our platform safe. We have tripled — to more than 35,000 — the people working on safety and security and we are a pioneer in artificial intelligence technology to remove hateful content at scale.

Zero tolerance does not mean zero incidences, but we're making real progress. A recent European Commission report found that Facebook assessed 95.7 percent of hate speech reports[2] in less than 24 hours. Last month, we reported that we find nearly 90 percent of the hate speech we remove before someone reports it — up from 24 percent from just over two years ago. We took action against 9.6 million pieces of hate speech content in the first quarter of 2020 — up from 5.7 million in the previous quarter. And 99 percent of the terrorist content we remove is taken down before anyone reports it to us.

Billions of people use Facebook and Instagram because they have good experiences; they don't want to see hateful content, our advertisers don't want to see it, and we don't want to see it. There is no incentive for us to do anything but remove it.

With that context in mind, please find answers to your specific questions below.

[1] https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/

[2] Defined as "illegal hate speech" under national laws transposing the EU Council Framework. See https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf

575 7TH STREET NW
STE 700
WASHINGTON, D.C.
20004

1. Does Facebook affirm its policy against hate speech and will it seriously enforce this policy?

Yes. As noted above, we take a zero-tolerance approach to hate speech we find on our platform. We invest billions of dollars each year in people and technology to keep our platform safe and are leading the way when it comes to deploying artificial intelligence technology to remove hate speech at scale. The European Commission's recent assessment on the Code of Conduct on Countering Illegal Hate Speech Online found that Facebook assessed 95.7 percent of hate speech reports in less than 24 hours; this number jumped to 99.1 percent over a 48 hours.[3]

Relatedly, under our Dangerous Organizations policy, we have designated and banned individuals and organizations because of their ties to terrorism, organized hate, and large-scale criminal activity, including over 250 white supremacist individuals and organizations--including David Duke, American Renaissance, and Richard Spencer.

2. What procedures has Facebook put in place to identify and remove hate speech from its platform? To what degree do these procedures differ with respect to public Facebook pages and private groups?

Specific procedures, applicable both to public pages and groups (including private groups), are summarized below:

- **Hate speech detection:** Over the last three 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, when possible, before anyone sees it. Our detection techniques include text and image matching--identifying images and strings of text that have already been removed as hate speech--and machine-learning classifiers that look at things like language and the reactions and comments to a post to assess how closely they match common phrases, patterns, and attacks that we've seen previously in content that violates our policies against hate.

  Initially, we used these systems to proactively detect potential hate speech violations and send them to our content review teams for human review. Starting late last year, thanks to continued progress in our systems' abilities to correctly detect violations, we began removing some posts automatically, but only when content is either identical or nearly identical to content previously removed. In all other cases when our systems proactively detect potential hate speech, the content is sent to our review teams to make a final determination.

  With the evolution in our detection systems, our proactive rate has climbed to 88.8 percent, and we've increased the volume of content we find and remove for violating our hate speech policy. Although we are pleased with this progress, these technologies are not perfect and we know that mistakes can still happen. That's why we continue to invest in systems that enable us to improve our

[3] See footnote 2.

accuracy in removing content that violates our policies while safeguarding content that discusses or condemns hate speech. Similar to how we review decisions made by our content review team in order to monitor the accuracy of our decisions, our teams routinely review removals by our automated systems to make sure we are enforcing our policies correctly.

- **Organized hate detection:** Three years ago, we started to develop a playbook and a series of automated techniques[4] to detect content related to terrorist organizations such as ISIS, al Qaeda, and their affiliates. We've since expanded these techniques to detect and remove content related to other terrorist groups and organized hate. We're now able to detect text embedded in images and videos in order to understand its full context, and we've built media matching technology to find content that's identical or nearly identical to photos, videos, text and even audio that we've already removed. When we started detecting hate organizations, we focused on groups that posed the greatest threat of violence at that time; we've now expanded to detect more groups tied to different hate-based and violent extremist ideologies and using different languages. In addition to building new tools, we have adapted strategies from our counterterrorism work, such as leveraging off-platform signals to identify dangerous content on Facebook and implementing procedures to audit the accuracy of our AI's decisions over time.

  The team that leads this work is a cross-functional team of 350 people with expertise ranging from law enforcement and national security to counterterrorism intelligence and radicalization. We are not aware of other tech companies with teams of this size or breadth. In the first three months of this year, we removed over 6 million pieces of terrorist content from Facebook--over 99 percent of which was removed before being reported. In the same time period, we removed over 4 million pieces of content tied to organized hate--over 96 percent of which was removed before being reported.

3. Does Facebook affirm its policy against violence and incitement and will it seriously enforce this policy?

Yes. Under our Violence and Incitement policy, we remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety.

With respect to the boogaloo movement, our recent designation of a violent U.S.-based anti-government network affiliated with the broader movement was not the first time we've taken action against violence within the boogaloo movement. We have previously removed boogaloo content when we identified a clear call for violence. As a result, we have removed over 800 of the movement's posts for violating our Violence and

[4] https://about.fb.com/news/2017/06/how-we-counter-terrorism/

Incitement policy[5] over the last two months and limited the distribution of Pages and groups referencing the movement by removing them from the recommendations we show people on Facebook.

4. What procedures has Facebook put in place to identify and remove violence and incitement from its platform? To what degree do these procedures differ with respect to public Facebook pages and private groups?

We have reviewers who are trained to review content against these policies. This type of review— and determining what constitutes a "call to violence" in different contexts—is not conducive to machine learning.

5. Does Facebook affirm its commitment to ban "praise, support and representation of white nationalism and white separatism on Facebook and Instagram" as detailed in the company's May 27, 2019 post and will it seriously enforce this commitment?

Yes. Please see response to Question 1, above. Although we do not separate out white nationalist and white separatist entities from others designated under our dangerous organizations policy, we have designated over 250 white supremacist individuals and organizations and enforced against their presence on the platform--including against symbols they may use to represent their views on white nationalism and white separatism.

6. What steps has Facebook implemented since announcing this policy to remove "praise, support and representation of white nationalism and white separatism on Facebook and Instagram?"

Please see response to Question 2, above, with respect to enforcement of our policies against organized hate. In enforcing our policy against white nationalism and white separatism, we focus on explicit statements, as implicit statements would require us to analyze the content beyond what's visible on its face, which is not always possible given the information available to us and which allows for bias in enforcement. Since adopting a ban on white nationalism and separatism, however, our policy team has continued the work to identify the hate slogans and symbols affiliated with white nationalism and white separatism so we can stay ahead of trends and efforts to subvert our policies.

We continue to think about and study the ways that hate shows up online, including by working with external experts and commissioning independent research studying the symbols and terminology that known hate organizations adopt.

[5] https://www.facebook.com/communitystandards/credible_violence

7. Please provide our offices with any Facebook internal research concerning the platform's amplification of extremist groups.

Our efforts to combat terrorism and organized hate don't end with our policies. In March 2019, we started connecting people who search for terms associated with white supremacy on Facebook Search to resources focused on helping people leave behind hate groups. When people search for terms in this category in the US, they are directed to Life After Hate,[6] an organization founded by former violent extremists that provides crisis intervention, education, support groups, and outreach. We have since expanded this initiative to more communities in other countries - focusing on partnering with local organizations that have expertise in both the issues facing their communities and how to combat them and create off-ramps from violent extremism. For example, in Australia, when people search for terms associated with hate and extremism, they will be directed to EXIT Australia[7] and ruangobrol.id[8] respectively. These are local organizations focused on helping individuals leave the direction of violent extremism and terrorism.

We plan to continue expanding this initiative and we're consulting partners in additional countries, but organizations countering hate need additional support and funding. Although we are honored to work with them and apply Facebook's technology and ability to connect people to support our shared mission of redirecting people away from hate, we are also grateful when other institutions, including government, look to highlight these organizations for their important work.

We are also partnering with Moonshot CVE[9] to measure the impact of these efforts to combat hate and extremism. Being able to measure our impact will allow us to hone our best practices and identify areas for improvement. By using Moonshot CVE's data-driven approach to disrupting violent extremism, we'll be able to develop and refine how we track the progress of these efforts across the world to connect people with information and services to help them leave hate and extremism behind. We will continue to seek out partners in countries around the world where local experts are working to disengage vulnerable audiences from hate organizations.

8. How often are you personally briefed on the status of domestic extremist and white supremacist groups on Facebook and the platform's efforts to address these groups?

Facebook leadership receives regular in-person updates on all our enforcement efforts.

[6] https://www.lifeafterhate.org/

[7] https://www.exit.org.au/

[8] https://www.ruangobrol.id/

[9] http://moonshotcve.com/

9. Who is the senior-most Facebook official responsible for addressing white supremacist groups' activity on Facebook and which Facebook executive does this employee report directly to?

The senior-most official responsible for addressing such content is CEO Mark Zuckerberg.

10. What role did Vice President of Global Public Policy Joel Kaplan play in Facebook's decision to shut down and de-prioritize internal efforts to contain extremist and hyperpolarizing activity on Facebook?

We are committed to understanding polarization and reducing its impact on how people experience our products. In fact, earlier this year, we made a multi-million dollar commitment to help study polarization and misinformation.

11. What role did Mr. Kaplan play in the participation of the Daily Caller, an outlet with longstanding ties to white nationalist groups, in Facebook's fact-checking program?

All of Facebook's third-party fact-checking partners are certified by Poynter's independent International Fact-Checking Network and must subscribe to the IFCN's rigorous Code of Principles.

12. When violent extremist groups actively and openly use a platform's tools to coordinate violence, should federal law continue to protect the platform from civil liability for its role in facilitating that activity?

As Senator Wyden has put it, Section 230 of the Communications Decency Act serves as both a sword and a shield. On the one hand, it has given Facebook and other online service providers the ability to innovate and enhance competition in the marketplace of ideas without expending critical resources on litigation. It has also given these services breathing room and flexibility to remove content they consider harmful or inappropriate for their services. Section 230 does not, and was never intended to, protect us from liability for content we create or for our own involvement in federal crimes.

Again, thank you for the opportunity to address your questions.

Sincerely,

Rachel Lieber

Rachel Lieber
Director and Associate General Counsel

575 7TH STREET NW
STE 700
WASHINGTON, D.C.
20004