September 4, 2020

Senator Mike Lee
361A Russell Senate Office Building
Washington, D.C. 20510

Dear Senator Lee:

Thank you for your letter of July 30 inquiring about Facebook's content moderation policies.

When we began reviewing content on Facebook over a decade ago, our system relied on people to report things they saw and thought were inappropriate. Our teams then reviewed and removed individual pieces of content if they broke our rules. A lot has changed since then. We've developed more transparent rules, and we have established an Oversight Board that will soon begin to review some of our most difficult content decisions. But the biggest change has been the role of technology in our content moderation efforts. As our Community Standards Enforcement Report shows, our technology to detect violating content is improving and playing a larger role in content review. Our technology helps us with proactive detection, automation, and prioritization.[1] Together, these three applications of technology have transformed our content review process and greatly improved our ability to moderate content at scale.

We take the safety of our community seriously and work hard to prevent abuse on our platform — and we're committed to being transparent about how we do this. That's why we make our Community Standards[2] available to everyone and why we share the changes we make to our policies each month. Our Community Standards Enforcement Reports are designed to show how we're doing at enforcing these policies and to help people hold us accountable and provide feedback.

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

1. What content-moderation standards do you employ when you remove content from your platform, where the content does not violate state or federal laws? Specifically, please explain your standards for removing content related to:

- **a. COVID-19;**
- **b. violent riots, and how you distinguish them from peaceful protests;**
- **c. hate speech;**
- **d. protections for the unborn;**
- **e. misinformation; and**
- **f. terrorist influence**

[1] For more information, see https://about.fb.com/news/2020/08/how-we-review-content/
[2] https://www.facebook.com/communitystandards/

facebook

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

You can find an in depth outline of what is and is not allowed on Facebook and Instagram in our Facebook Community Standards[3] and our Instagram Community Guidelines.[4] These standards must keep pace with changes happening online and in the world, so we are constantly improving the Community Standards over time based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights.

We made some updates to the Community Standards in response to COVID-19 to help protect people from harmful content and new types of abuse related to this unprecedented public health emergency. You can learn more about these updates directly on the Community Standards page.[5] In particular, the following sections may also be helpful based on the topics you raised.

- Violence and Incitement
- Hate Speech
- False News
- Dangerous Individuals and Organizations

You can also find more information about our work on COVID-19 online.[6]

2. How did you formulate standards for the above categories of content? What sources did you look to in forming your content policies?

The goal of our Community Standards is to encourage expression and create a safe environment on Facebook. We base our policies on input from our community and from experts and organizations outside Facebook so we can better understand different perspectives on safety and expression, as well as the impact of our policies on different communities globally. Based on this feedback, as well as changes in social norms and language, our standards evolve over time.

For example, we asked former Senator Jon Kyl to conduct a review of potential anti-conservative bias at Facebook. He and his team at Covington & Burling met with more than 130 conservative politicians and organizations and produced a report outlining the key concerns they heard as well as the changes Facebook has already made to address them.[7] These changes include making our decisions more transparent by providing more information as to why people are seeing specific posts on News Feed; helping Page managers see when enforcement action takes place; launching an appeals process; creating a new Oversight Board for content, made up of people with a diverse range of ideological views; and changes

[3] https://www.facebook.com/communitystandards/
[4] https://help.instagram.com/477434105621119?helpref=page_content
[5] https://www.facebook.com/communitystandards/
[6] https://about.fb.com/news/2020/08/coronavirus/#misinformation-update
[7] See https://newsroom.fb.com/news/2019/08/update-on-potential-anti-conservative-bias/

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

to how we label ads concerning social issues, elections, or politics. The Kyl report also addressed our policy on banning images of patients with medical tubes, which had been applied inconsistently in the past, and was inadvertently impacting pro life advertising. This was an issue you raised with us in a Senate Judiciary Subcommittee hearing last year, as well. We incorporated that feedback and adjusted our policies to try to prevent these unintended consequences. We take the concerns set forth in the report seriously, and we will continue to work with Senator Kyl and his team to examine and, where necessary, adjust our policies and practices going forward.

As part of our policy development process, members of the content policy team run a meeting called the Product Policy Forum to discuss potential changes to our Community Standards. A variety of subject matter experts participate in this meeting, including members of our safety and cybersecurity policy teams, counterterrorism specialists, Community Operations employees, product managers, public policy leads and representatives from our legal, communications and diversity teams. In preparation for these meetings, members of our content policy team reach out to internal and external experts, analyze data, conduct research and study relevant scholarship to inform our policy proposals. This multi-step effort allows us to account for a range of perspectives and opinions across the globe, to ultimately develop stronger policies. When our policies are written or updated, we share those updates on our Community Standards website.[8]

If you're interested in taking a closer look at recent policy development, you can find meeting summaries from the Product Policy Forum online.[9]

3. What are the prerequisites for a content-moderator position at your company? Do you inquire about the political or other beliefs of a candidate before making a hiring decision? Where you use contractors to serve in these roles, how do you ensure that they follow your internal guidelines and standards?

We do not inquire about political beliefs in making employment decisions for content moderators. We work with partners that uphold objective, fair hiring practices. We do not set prerequisites or other guidelines regarding the beliefs of their employees.

We have invested significantly in safety and security and now have over 35,000 people working in this area, about 15,000 of whom review content. The majority of our content reviewers are people who work full-time for our partners and work at sites managed by these partners. We have a global network of partner companies so that we can quickly adjust the focus of our workforce as needed. This approach gives us the ability to, for example, make sure we have the right language or regional expertise. Our partners have a core competency in this type of work and are able to help us adjust as new needs arise or when a situation around the world warrants it.

[8] https://www.facebook.com/communitystandards/recentupdates/
[9] https://about.fb.com/news/2018/11/content-standards-forum-minutes/

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

Our content reviewers undergo extensive training when they join and thereafter are regularly trained and tested with specific examples on how to uphold our Community Standards and take the correct action on a piece of content. This training also occurs when policies are clarified or as they evolve. There are quality control mechanisms as well as management points of contact onsite and available remotely to help or provide guidance to reviewers if needed. When a reviewer is not clear on the action to take based on our Community Standards, they can pass the content decision to another team for review. We also audit the accuracy of reviewer decisions on an ongoing basis, as do our partners.

We are always working with our partners to improve operations and the training and support that are provided to each person that reviews content on behalf of Facebook. Some of these initiatives include improving training materials to include more multimedia to support all learning types; providing additional training resources for well-being, resiliency, and unconscious biases; and providing additional marketized examples for our global network of content reviewers.

Please see Response to Question 4 for additional information.

4. What is the internal process that your content moderators follow to remove content that violates your standards?

Decisions about whether to remove content are based on whether the content violates our Community Standards. Accordingly, content reviewers take action on content based on those Standards. Our Community Standards are global, and all reviewers use the same guidelines when making decisions. We seek to write actionable policies that clearly distinguish between violating and non-violating content; our policies are extremely granular because we want to ensure that the content review process is as objective as possible. Every week, we audit a sample of all reviewer decisions for accuracy and consistency. When we find that a reviewer makes mistakes or misapplies our policies, we follow up with the partner to take appropriate action. We have instituted additional controls and oversight around content review, including robust escalation procedures and updated reviewer training materials. These improvements and safeguards are designed to encourage free expression on our platforms, while keeping our users safe.

Additionally, in April 2018, we announced the launch of content-level appeals, and we make appeals or “disagree with decision” available for certain types of content that is removed from Facebook when we have resources to review the appeals or feedback. We recognize that we make enforcement errors on both sides of the equation—what to allow and what to remove—and that our mistakes may cause a great deal of concern for people, which is why we need to allow the option to request review of the decision when we can and allow users to provide additional context that will help the content review team see the fuller picture as they review the post again. This type of feedback will allow us to continue improving our systems and processes so we can work with our partners and the content reviewers to prevent similar mistakes in the future.

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

In November 2018, Mark Zuckerberg outlined a blueprint for a new system for content governance and enforcement. He said "Facebook should not make so many important decisions about free expression and safety on our own." With our size comes a great deal of responsibility and while we have always taken advice from experts on how to best keep our platforms safe, until now, Facebook through its content reviewers has made the final decisions about what should be allowed on our platforms and what should be removed. And these decisions often are not easy to make – most judgments do not have obvious, or uncontroversial, outcomes and yet many of them have significant implications for free expression. That's why we have created and empowered a new group to exercise independent judgment over some of the most difficult and significant content decisions. In May 2020, we announced the first members of the Oversight Board. These members reflect a wide range of views and experiences. Its long-term success depends on it having members who bring different perspectives and expertise to bear. We expect them to make some decisions that we, at Facebook, will not always agree with – but that's the point: they are truly autonomous in their exercise of independent judgment.

5. How do you ensure that a content-moderation decision is not influenced by the personal beliefs or political views of the moderator?

Our content reviewers undergo extensive training when they join, with over 80 hours of instructor-led, hands-on learning and shadowing of veteran reviewers. They are trained and tested with specific examples on how to uphold the Community Standards and take the correct action on a piece of content. There is also ongoing training when policies are clarified, or as they evolve. We and our partners also audit the accuracy of reviewer decisions on an ongoing basis to coach them and follow up on improving when errors are made. And when we are made aware of incorrect content removals, we review them with our Community Operations team and partners to prevent similar mistakes in the future.

We are always working to improve our operations and the training and support that are provided to each person that reviews content on behalf of Facebook. Some of these initiatives include improving training materials to include more multimedia to support all learning types; providing additional training resources for well-being, resiliency, and unconscious biases; and providing additional marketized examples for our global network of content reviewers.

6. If your content-moderation standards rely on guidance from a government entity, please explain your policy for allowing on your platform speech that disagrees with the government. If CDC guidance is the basis for removing content regarding COVID-19, how is that standard applied consistently? For example, since the CDC says that it is safe for schools to open, do you remove content from your platform that claims it is unsafe to reopen schools?

We remove content that violates our community standards on misinformation and could lead to *imminent physical harm*. Such content includes, for example, false claims about COVID-19 cures and treatments, false claims about prevention, false information about access or the

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

availability of health resources, and false information about the location or severity of the outbreak. We also remove content that discourages people from following certain globally-applicable preventative measures, such as claims that face masks and social distancing are not effective, which health experts like the WHO and CDC have told us could lead to imminent physical harm. We do not categorically remove all claims debating or challenging health guidelines. We would remove a claim that schools are unsafe because masks are ineffective, but not a general claim that schools are safe or unsafe.

7. Where do you clearly articulate your content-moderation standards? How do you convey your moderation standards to consumers? Do you regularly update your users on changes made to your policies?

Our content moderation standards are published online as part of our Community Standards.[10] Our Community Standards Page includes a "Recent Update" tab. Our Community Standards Enforcement Reports provide metrics on how well we enforce our policies. Beginning in August 2020, we are publishing our Community Standards Enforcement Report on a quarterly basis to more effectively track our progress and demonstrate our continued commitment to making our platforms safe.[11]

8. Are your users required to provide knowing consent to the standard before giving you their personal information, data, and content, which gives value to your platform?

Users must click to agree to our Terms of Service and Data Policy when they sign up for an account. Our Terms of Service[12] link to our Data Policy,[13] which explains the information we process, including, for example, the kinds of information we collect and how we use and share this information, as well as user rights with respect to their data.

9. Do you coordinate your content moderation standards with other online platforms or competitors? Have you ever discussed or reached an agreement regarding these standards—or the removal of content generally—with any other online platform or competitor? If so, please list each discussion or agreement, its subject, and the parties thereto.

Where there are implications for safety--such as in the terrorism and child abuse contexts--we coordinate with other platforms, including with our competitors. Additional information about these efforts is below.

Global Internet Forum to Counter Terrorism (GIFCT): Facebook is a founding member of GIFCT, formally established in July 2017 as a group of companies, dedicated to disrupting

[10] https://www.facebook.com/communitystandards/
[11] https://transparency.facebook.com/community-standards-enforcement
[12] https://www.facebook.com/terms.php
[13] https://www.facebook.com/about/privacy/update

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

terrorist abuse of members' digital platforms. The original Forum was led by a rotating chair drawn from the founding four companies—Facebook, Microsoft, Twitter and YouTube—and managed a program of knowledge-sharing, technical collaboration and shared research. The GIFCT mission statement is to prevent terrorists and violent extremists from exploiting digital platforms. In September 2019 during the UN General Assembly the GIFCT announced that it would be reorganized as an independent Non-Governmental Organization. The GIFCT is now an independent 501(c)(3) registered in the United States and has an independent Executive Director and staff.

- While GIFCT does not standardize terms of service or moderation practices, these are topics of discussion within the group. Of most relevance programmatically is the Hash Sharing Consortium of the GIFCT. The consortium shares "hashes" (or "digital fingerprints") of known terrorist images and videos. The image or video is "hashed" in its raw form and is not linked to any source original platform or user data. Hashes appear as a numerical representation of the original content, which means it cannot be easily reverse engineered to create the image and/or video. It is up to each consortium member how they leverage the database, depending on, among other things, their own terms of service, how their platform operates, and how they utilize technical and human capacities.
- Just like governments, intergovernmental institutions, civil society organizations, and academics, companies often have slightly different definitions of "terrorism" and "terrorist content". To find common ground, the original scope of the hash-sharing database was therefore limited to content related to organizations on the United Nations Security Council's consolidated sanctions list. The only hashes that appear in the Hash Sharing database that do not correspond to entities on the UN list were added during a declared Content Incident Protocol, which will be discussed in more detail below.
- More about GIFCT, its core programs and cross-platform as well as multi-sector collaborations can be found online.[14]
- GIFCT's most recent transparency report including a list of GIFCT and Hash Sharing Consortium members, taxonomies and numbers can be found online.[15]

WePROTECT Global Alliance: In March, the governments of the United States, Australia, Canada, New Zealand, and the United Kingdom announced the publication of Voluntary Principles to Counter Online Child Exploitation and Abuse,[16] developed in consultation with

[14] https://gifct.org/about/
[15] https://gifct.org/transparency/
[16] https://www.justice.gov/opa/press-release/file/1256061/download

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

Facebook and several other leading technology companies including Google, Microsoft, Twitter, Snap, and Roblox.

The principles aim to provide a framework to combat online child sexual exploitation and abuse, and are intended to drive collective action. The WePROTECT Global Alliance, which currently comprises 97 governments, 25 technology companies and 30 civil society organisations, will promote and support the adoption of the principles at a global level to drive collective industry action. One of the themes of the principles is "Collaborate & Respond to Evolving Threat" and includes industry commitments to "support opportunities to share relevant expertise, helpful practices, data and tools where appropriate and feasible."

We work across industry to protect kids online. For example, we made our photo and video-matching technologies open source, which allows industry partners, developers and non-profits to more easily identify abusive content, share hashes — or digital fingerprints — of different types of harmful content and allow hash-sharing systems to communicate with each other, making the systems that much more powerful.

We also recently hosted our fifth child safety hackathon, where we brought together engineers, data scientists and designers from across the industry as well as non-profit partners NCMEC, Thorn, SaferNet Brazil, INHOPE, Cybertip.ca, and IWF, to code and prototype more than a dozen projects focused on making the internet a safer place for children.

We have also taken steps across our apps to make the broader internet safer for children. This includes running PhotoDNA on links shared on all our apps from other internet sites and their associated content to detect known child exploitation housed elsewhere on the internet. Not only does this help keep our platforms safer, but it also helps keep the broader internet safer as all violating content is shared with the National Center for Missing and Exploited Children (NCMEC) who work with local law enforcement around the world.

In June, Facebook joined Google, Microsoft and 15 other tech companies to announce the formation of Project Protect: A plan to combat online child sexual abuse – a renewed commitment and investment from the Technology Coalition expanding its scope and impact to protect kids online and guide its work for the next 15 years.[17]

10. Do you coordinate the removal of specific content with other online platforms or competitors? If so, please explain the process and what content has been subject to coordinated removal.

[17] https://www.technologycoalition.org/2020/05/28/a-plan-to-combat-online-child-sexual-abuse/

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

Because online child exploitation is an internet problem, it demands an internet solution, Facebook is committed to a multi-stakeholder, comprehensive global effort to fight child exploitation and thwart the proliferation of CEI. Through NCMEC, the industry shares hashes to ensure we all have the most complete knowledge of known child exploitation images. NCMEC offers a unified solution bringing together non-governmental organizations, law enforcement, technology companies, educators, parents, as well as the voice and perspective of survivors for a global, multi-stakeholder response.

U.S. federal law requires that U.S.-based companies report instances of apparent child pornography that they become aware of on their systems to NCMEC's CyberTipline. In 1998, NCMEC launched the CyberTipline to provide the public and electronic service providers with the ability to report suspected child sexual exploitation including online enticement of children for sexual acts, extra-familial child sexual molestation, child pornography, child sex tourism, child sex trafficking, unsolicited obscene materials sent to children, misleading domain names, and misleading words or digital images on the internet. After NCMEC's review is completed, all information in a CyberTipline report is made available to the appropriate law enforcement agency around the world.

NCMEC works closely with platforms on voluntary initiatives that many companies choose to engage in to deter and prevent the proliferation of online child sexual exploitation images. To date, over 1,400 companies are registered to make reports to NCMEC's CyberTipline and, in addition to making reports, these companies also receive notices from NCMEC about suspected CSAM on their servers.

With survivors speaking increasingly about the long-lasting damage and impact of their images and videos being on the internet, NCMEC is also working with the industry and with children and their families to identify these images and have them tagged for removal from servers.

11. Some of you have removed or threatened to remove the ability of third parties to monetize their content through your advertising platform on the basis of content found in the third party's comments section. What is your policy or standard for such demonetization? Do you believe the same standard should be applied to platforms currently protected by Section 230?

When we remove content for violating our policies, we notify the person who posted it to explain why, with some narrow exceptions to account for things like child exploitation imagery. Beyond enforcing our policies on individual pieces of content, violations of our terms and policies may result in disabling Ad Accounts, Pages, Business Managers and/or individual user accounts, or disabling use of certain products or features of our platform, such as the ability to run ads.

Misinformation is one example of this. If Pages, domains, or Groups repeatedly share misinformation, we'll continue to reduce their overall distribution, and we'll place restrictions on the Pages' ability to advertise and monetize. We don't want people to game

Address: 575 7th Street NW
Suite 700
Washington, DC
20004

the system, so we do not share the specific number of violations that leads to a temporary block or permanent suspension.

We regularly review our policies to make sure they are in the right place.

Thank you, again, for the opportunity to answer your questions. We look forward to working with your office going forward.

Sincerely,

Kevin Martin
Vice President, U.S. Public Policy

facebook

Address: 575 7th Street NW
Suite 700
Washington, DC
20004