EXAMINING SOCIAL MEDIA COMPANIES' EFFORTS TO COUNTER ON-LINE TERROR CONTENT AND MISINFORMATION

HEARING

BEFORE THE

COMMITTEE ON HOMELAND SECURITY
HOUSE OF REPRESENTATIVES

ONE HUNDRED SIXTEENTH CONGRESS

FIRST SESSION

JUNE 26, 2019

Serial No. 116–30

Printed for the use of the Committee on Homeland Security

Available via the World Wide Web: http://www.govinfo.gov

U.S. GOVERNMENT PUBLISHING OFFICE
38–783 PDF WASHINGTON : 2020

COMMITTEE ON HOMELAND SECURITY

Bennie G. Thompson, Mississippi, *Chairman*

Sheila Jackson Lee, Texas
James R. Langevin, Rhode Island
Cedric L. Richmond, Louisiana
Donald M. Payne, Jr., New Jersey
Kathleen M. Rice, New York
J. Luis Correa, California
Xochitl Torres Small, New Mexico
Max Rose, New York
Lauren Underwood, Illinois
Elissa Slotkin, Michigan
Emanuel Cleaver, Missouri
Al Green, Texas
Yvette D. Clarke, New York
Dina Titus, Nevada
Bonnie Watson Coleman, New Jersey
Nanette Diaz Barragán, California
Val Butler Demings, Florida

Mike Rogers, Alabama
Peter T. King, New York
Michael T. McCaul, Texas
John Katko, New York
John Ratcliffe, Texas
Mark Walker, North Carolina
Clay Higgins, Louisiana
Debbie Lesko, Arizona
Mark Green, Tennessee
Van Taylor, Texas
John Joyce, Pennsylvania
Dan Crenshaw, Texas
Michael Guest, Mississippi

Hope Goins, *Staff Director*
Chris Vieson, *Minority Staff Director*

CONTENTS

Page

STATEMENTS

The Honorable Bennie G. Thompson, a Representative in Congress From the State of Mississippi, and Chairman, Committee on Homeland Security:
Oral Statement 1
Prepared Statement 2
The Honorable Mike Rogers, a Representative in Congress From the State of North Carolina, and Ranking Member, Committee on Homeland Security:
Oral Statement 3
Prepared Statement 4

WITNESSES

Ms. Monika Bickert, Head of Global Policy Management, Facebook:
Oral Statement 5
Prepared Statement 7
Mr. Nick Pickles, Global Senior Strategist for Public Policy, Twitter:
Oral Statement 11
Prepared Statement 12
Mr. Derek Slater, Global Director of Information Policy, Google:
Oral Statement 17
Prepared Statement 19
Ms. Nadine Strossen, John Marshall Harlan II, Professor of Law, New York Law School:
Oral Statement 22
Prepared Statement 23

FOR THE RECORD

The Honorable Bennie G. Thompson, a Representative in Congress From the State of Mississippi, and Chairman, Committee on Homeland Security:
Letter, October 30, 2017 74
Letter, February 22, 2018 76
Letter, December 18, 2018 78
Statement of the ADL (Anti-Defamation League) 80

APPENDIX

Questions From Chairman Bennie G. Thompson for Monika Bickert 91
Questions From Honorable Lauren Underwood for Monika Bickert 95
Questions From Ranking Member Mike Rogers for Monika Bickert 96
Questions From Chairman Bennie G. Thompson for Nick Pickles 98
Questions From Ranking Member Mike Rogers for Nick Pickles 101
Questions From Chairman Bennie G. Thompson for Derek Slater 104
Questions From Honorable Lauren Underwood for Derek Slater 104
Questions From Ranking Member Mike Rogers for Derek Slater 104
Question From Ranking Member Mike Rogers for Nadine Strossen 105

EXAMINING SOCIAL MEDIA COMPANIES' EFFORTS TO COUNTER ON-LINE TERROR CONTENT AND MISINFORMATION

Wednesday, June 26, 2019

U.S. HOUSE OF REPRESENTATIVES,
COMMITTEE ON HOMELAND SECURITY,
Washington, DC.

The committee met, pursuant to notice, at 10 a.m., in room 310, Cannon House Office Building, Hon. Bennie G. Thompson (Chairman of the committee) presiding.

Present: Representatives Thompson, Jackson Lee, Langevin, Correa, Torres Small, Rose, Underwood, Slotkin, Cleaver, Green of Texas, Clarke, Titus, Watson Coleman, Barragán, Demings, Rogers, King, Katko, Walker, Higgins, Lesko, Green of Tennessee, Taylor, Joyce, Crenshaw, and Guest.

Chairman THOMPSON. The Committee on Homeland Security will come to order.

The committee is meeting today to receive testimony on "Examining Social Media Companies' Efforts to Counter On-line Terror Content and Misinformation."

In March, a white supremacist terrorist killed 51 people and wounded 49 more at two mosques in Christchurch, New Zealand. Our thoughts and prayers continue to be with the victims and their families.

The motive behind the attack is not in question. The terrorist had written an extensive manifesto outlining his white supremacist, white nationalist, anti-immigrant, anti-Muslim, and fascist beliefs. His act was horrifying beyond words, and it shook the conscience.

Shockingly, the terrorist was able to live-stream the attack on Facebook, where the video and its gruesome content went undetected initially. Instead, law enforcement officials in New Zealand had to contact the company and ask that it be removed.

When New Zealand authorities called on all social media companies to remove these videos immediately, they were unable to comply. Human moderators could not keep up with the volume of videos being reposted, and their automated systems were unable to recognize minor changes in the video. So the video spread on-line and spread around the world.

The fact that this happened nearly 2 years after Facebook, Twitter, Google, Microsoft, and other major tech companies established the Global Internet Forum to Counter Terrorism, pronounced "GIF-C-T," is troubling to say the least. The GIFCT was created for tech companies to share technology and best practices to combat the spread of on-line terrorist content.

Back in July 2017, representatives of the GIFCT briefed this committee on this new initiative. At the time, I was optimistic about its intentions and goals and acknowledged that its members demonstrated initiative and willingness to engage on this issue while others have not. But after a white supremacist terrorist was able to exploit social media platforms in this way, we all have reason to doubt the effectiveness of the GIFCT and the companies' efforts more broadly.

On March 27 of this year, representatives of GIFCT briefed this committee after the Christchurch massacre. Since then, myself and other Members of this committee have asked important questions about the organization and your companies and have yet to receive satisfactory answers.

Today, I hope to get answers regarding your actual efforts to keep terrorist content off your platforms. I want to know how you will prevent content like the New Zealand attack video from spreading on your platforms again.

This committee will continue to engage social media companies about the challenges they face in addressing terror content on their platforms. In addition to terror content, I want to hear from our panel about how they are working to keep hate speech and harmful misinformation off their platforms.

I want to be very clear: Democrats respect the free-speech rights enshrined in the First Amendment. But much of the content I am referring to is either not protected speech or violates the social media companies' own terms of service.

We have seen time and time again that social media platforms are vulnerable to being exploited by bad actors, including those working at the behest of foreign governments, who seek to sow discord by spreading misinformation. This problem will only become more acute as we approach the 2020 elections. We want to understand how companies can strengthen their efforts to deal with this persistent problem.

At a fundamental level, today's hearing is about transparency. We want to get an understanding of whether and to what extent social media companies are incorporating questions of National security, public safety, and integrity of our domestic institutions into their business models. I look forward to having that conversation with the witnesses here today and to our on-going dialog on behalf of the American people.

I thank the witnesses for joining us and the Members for their participation.

With that, I now recognize the Ranking Member of the full committee, the gentleman from Alabama, Mr. Rogers, for 5 minutes for the purpose of an opening statement.

[The statement of Chairman Thompson follows:]

Statement of Chairman Bennie G. Thompson

June 26, 2019

In March, a white supremacist terrorist killed 51 people and wounded 49 more at 2 mosques in Christchurch, New Zealand. Our thoughts and prayers continue to be with the victims and their families. The motive behind the attack is not in question—the terrorist had written an extensive manifesto outlining his white supremacist, white nationalist, anti-immigrant, anti-Muslim, and fascist beliefs. His act was horrifying beyond words, and it shook the conscience. Shockingly, the terrorist was able to live-stream the attack on Facebook, where the video and its gruesome content went undetected initially. Instead, law enforcement officials in New Zealand had to contact the company and ask that it be removed. When New Zealand authorities called on all social media companies to remove these videos immediately, they were unable to comply. Human moderators could not keep up with the volume of videos being reposted, and their automated systems were unable to recognize minor changes to the video. So, the video spread on-line spread around the world.

The fact that this happened nearly 2 years after Facebook, Twitter, Google, Microsoft, and other major tech companies established the Global Internet Forum to Counter Terrorism, or GIFCT, is troubling to say the least. The GIFCT was created for tech companies to share technology and best practices to combat the spread of on-line terrorist content. Back in July 2017, representatives from GIFCT briefed this committee on this new initiative. At the time, I was optimistic about its intentions and goals, and acknowledged that its members demonstrated initiative and willingness to engage on this issue while others have not. But after a white supremacist terrorist was able to exploit social media platforms in this way, we all have reason to doubt the effectiveness of the GIFCT and the companies' efforts more broadly. On March 27 of this year, representatives of GIFCT briefed this committee after the Christchurch massacre. Since then, myself and other Members of this committee have asked important questions about the organization and your companies, and we have yet to receive satisfactory answers.

Today, I hope to get answers regarding your actual efforts to keep terrorist content off your platforms. I want to know how you will prevent content like the New Zealand attack video from spreading on your platforms again. This committee will continue to engage social media companies about the challenges they face in addressing terror content on their platforms. In addition to terror content, I want to hear from our panel about how they are working to keep hate speech and harmful misinformation off their platforms. I want to be very clear—Democrats respect the free speech rights enshrined in the First Amendment, but much of the content I am referring to is either not protected speech or violates the social media companies' own terms of service.

We have seen time and time again that social media platforms are vulnerable to being exploited by bad actors, including those working at the behest of foreign governments, who seek to sow discord by spreading misinformation. This problem will only become more acute as we approach the 2020 elections. We want to understand how companies can strengthen their efforts to deal with this persistent problem. At a fundamental level, today's hearing is about transparency. We want to get an understanding of whether—and to what extent—social media companies are incorporating questions of National security, public safety, and the integrity of our democratic institutions into their business models.

Mr. ROGERS. Thank you, Mr. Chairman.

Concerns about violent and terror-related on-line content has existed since the creation of the internet. This issue has peaked over the last decade, with the growing sophistication in which foreign terrorists and their global supporters have exploited the openness of on-line platforms to radicalize, mobilize, and promote their violent messages. These tactics prove successful, so much so that we are seeing domestic extremists mimic many of the same techniques to gather followers and spread hateful, violent propaganda.

Public pressure has grown steadily on the social media companies to modify their terms of service to limit posts linked to terrorism, violence, criminal activity, and, most recently, the hateful rhetoric of misinformation. The large and mainstream companies have responded to this pressure in a number of ways, including the creation of the Global Internet Forum to Counter Terrorism, or GIFCT. They are also updating their terms of service and hiring more human content moderators.

Today's hearing is also an important opportunity to examine the Constitutional limits placed on the Government to regulate or restrict free speech. Advocating violent acts and recruiting terrorists on-line is illegal, but expressing one's political views, however repugnant they may be, is protected under the First Amendment.

I was deeply concerned to hear the recent news reports about Google's policy regarding President Trump and conservative news media. Google's head of responsible innovation, Jen Gennai, recently said, "We all got screwed over in 2016. The people got screwed over, the news media got screwed over, everybody got screwed over. So we've rapidly been like, what happened there? How do we prevent this from happening again?".

Then, Ms. Gennai again on video remarked, "Elizabeth Warren is saying that we should break up Google. That will not make it better. It will make it worse, because all these smaller companies that don't have the same resources that we do will be charged with preventing the next Trump situation."

Now, Ms. Gennai is entitled to her opinion, but we are in trouble if her opinions are Google's policy.

That same report details alarming claims about Google's deliberate attempt to alter search results to reflect the reality Google wants to promote rather than objective facts. This report and others like it are a stark reminder of why the Founders created the First Amendment.

In fact, the video I just quoted from has been removed from YouTube. That platform is owned by Google, who is joining us here today. I have serious questions about Google's ability to be fair and balanced when it appears they have colluded with YouTube to silence negative press coverage.

Regulating speech quickly becomes a subjective exercise for Government or the private sector. Noble intentions often give way to bias and political agendas. The solution to this problem is complex. It will involve enhanced cooperation between the Government, industry, individuals, while protecting the Constitutional rights of all Americans.

I appreciate our witnesses' participation here today. I hope that today's hearing can be helpful in providing greater transparency and understanding of this complex challenge.

With that, I yield back, Mr. Chairman.

[The statement of Ranking Member Rogers follows:]

Statement of Ranking Member Mike Rogers

June 26, 2019

Concerns about violent and terror-related on-line content have existed since the creation of the internet. The issue has peaked over the past decade with the growing sophistication in which foreign terrorists and their global supporters have exploited the openness of on-line platforms to radicalize, mobilize, and promote their violent messages.

These tactics proved successful—so much so that we are seeing domestic extremists mimic many of the same techniques to gather followers and spread hateful and violent propaganda.

Public pressure has grown steadily on the social media companies to modify their terms of service to limit posts linked to terrorism, violence, criminal activity, and, most recently, to hateful rhetoric and misinformation.

The large, mainstream companies have responded to this pressure in a number of ways, including the creation of the Global Internet Forum to Counter Terrorism, or GIFCT. They are also updating their terms of service and hiring more human content moderators.

Today's hearing is also an important opportunity to examine the Constitutional limits placed on the Government to regulate or restrict free speech.

Advocating violent acts and recruiting terrorists on-line is illegal. But expressing one's political views, however repugnant they may be, is protected under the First Amendment.

I was deeply concerned to hear news reports about Google's policies regarding President Trump and conservative news media.

Google's "Head of Responsible Innovation" Jen Gennai said recently, "Well all got screwed over in 2016 . . . the people got screwed over, the news media got screwed over . . . everybody got screwed over so we've rapidly been like what happened there and how do we prevent it from happening again?"

Ms. Gennai then remarked: "Elizabeth Warren is saying that we should break up Google . . . That will not make it better it will make it worse because all these smaller companies that don't have the same resources that we do will be charged with preventing the next Trump situation."

Ms. Gennai is entitled to her opinion but we are in trouble if her opinions are Google's policy.

That same report details alarming claims about Google's deliberate attempt to alter search results to reflect the reality Google wants to promote rather than objective facts.

This report, and others like it, are a stark reminder of why the Founders created the First Amendment.

In fact, the video that I just quoted from has been removed from YouTube. That platform is owned by Google who is joining us here today.

I have serious questions about Google's ability to be fair and balanced when it appears to have colluded with YouTube to silence this negative press coverage.

Regulating speech quickly becomes a subjective exercise for Government or the private sector.

Noble intentions often give way to bias and political agendas.

The solution to this problem is complex. It will involve enhanced cooperation between Government, industry, and individuals, while protecting the Constitutional rights of all Americans.

I appreciate our witness' participation here today. I hope that today's hearing can be helpful in providing greater transparency and understanding of this complex challenge.

Chairman THOMPSON. Thank you very much.

Other Members of the committee are reminded that, under the committee rules, opening statements may be submitted for the record.

I welcome our panel of witnesses.

Our first witness, Ms. Monika Bickert, is the vice president of global policy management at Facebook. Next, we are joined by Mr. Nick Pickles, who currently serves as the global senior strategist for public policy at Twitter. Our third witness is Mr. Derek Slater, the global director of information policy at Google. Finally, we welcome Ms. Nadine Strossen, who serves as the John Marshall Harlan II professor of law at New York Law School.

Without objection, the witnesses' full statements will be inserted in the record.

I now ask each witness to summarize his or her statement for 5 minutes, beginning with Ms. Bickert.

STATEMENT OF MONIKA BICKERT, HEAD OF GLOBAL POLICY MANAGEMENT, FACEBOOK

Ms. BICKERT. Thank you, Chairman Thompson, Ranking Member Rogers, and Members of the committee, and thank you for the opportunity to appear before you today.

I am Monika Bickert, Facebook's vice president for global policy management, and I am in charge of our product policy and counterterrorism efforts. Before I joined Facebook, I prosecuted Federal crimes for 11 years at the Department of Justice.

On behalf of our company, I want to thank you for your leadership combating extremism, terrorism, and other threats to our homeland and National security.

I would also like to start by saying that all of us at Facebook stand with the victims, their families, and everyone affected by the recent terror attacks, including the horrific violence in Sri Lanka and New Zealand. In the aftermath of these acts, it is even more important to stand together against hate and violence. We make this a priority in everything that we do at Facebook.

On terrorist content, our view is simple: There is absolutely no place on Facebook for terrorists. They are not allowed to use our services under any circumstances. We remove their accounts as soon as we find them. We also remove any content that praises or supports terrorists or their actions. If we find evidence of imminent harm, we promptly inform authorities.

There are three primary ways that we are implementing this approach: First, with our products that help stop terrorists and their propaganda at the gates; second, through our people, who help us review terrorist content and implement our policies; and, third, through our partnerships outside the company, which help us stay ahead of the threat.

So, first, our products. Facebook has invested significantly in technology to help identify terrorist content, including through the use of artificial intelligence but also using other automation and technology. For instance, we can now identify violating textual posts in 19 different languages.

With the help of these improvements, we have taken action on more than 25 million pieces of terrorist content since the beginning of 2018. Of the content that we have removed from Facebook for violating our terrorism policies, more than 99 percent of that is content that we found ourselves, using our own technical tools, before anybody has reported it to us.

Second are people. We now have more than 30,000 people who are working on safety and security across Facebook across the world, and that is 3 times as many people as we had dedicated to those efforts in 2017.

We also have more than 300 highly-trained professionals exclusively or primarily focused on combating terrorist use of our services. Our team includes counterterrorism experts, former prosecutors like myself, former law enforcement officials, former intelligence officials. Together, they speak more than 50 languages, and they are able to provide 24-hour coverage.

Finally, our partnerships. In addition to working with third-party intelligence providers to more quickly identify terrorist material on the internet, we also regularly work with academics who are studying terrorism and the latest trends and Government officials.

Following the tragic attacks in New Zealand, Facebook was proud to be a signatory to the Christchurch Call to Action, which is a 9-point plan for the industry to better combat terrorist attempts to use our services.

We also partner across industry. As Mr. Chairman and the Ranking Member mentioned, in 2017 we launched the Global Internet Forum to Counter Terrorism, or GIFCT, with YouTube, Microsoft, and Twitter. GIFCT, the point of that is we bring companies together from across industry to share information and also to share technology and research to better combat these threats.

Through GIFCT, we have expanded an industry database for companies to share what we call hashes, which are basically digital fingerprints of terrorist content, so that we can all remove it more quickly and help smaller companies do that too. We have also trained over 110 companies from around the globe in best practices for countering terrorist use of the internet.

Now, Facebook took over as the chair of GIFCT in 2019, and, along with our fellow members, we have this year worked to expand our capabilities, including making new audio and text hashing techniques available to other member companies, especially these smaller companies, and we have also improved our crisis protocols.

In the wake of the horrific Christchurch attacks, we communicated in real-time across our companies and were able to stop hundreds of versions of the video of the attack despite the fact that bad actors were actively trying to edit the video to upload it to circumvent our systems.

We know there are adversaries who are always evolving their tactics, and we have to improve if we want to stay ahead. Though we will never be perfect, we have made real progress, and we are committed to tirelessly combating extremism on our platform.

I appreciate the opportunity to be here today, and I look forward to answering your questions. Thank you.

[The prepared statement of Ms. Bickert follows:]

Prepared Statement of Monika Bickert

June 26, 2019

I. INTRODUCTION

Chairman Thompson, Ranking Member Rogers, and distinguished Members of the committee, thank you for the opportunity to appear before you today. My name is Monika Bickert, and I am the vice president of global policy management at Facebook. In that role, I lead our efforts related to Product Policy and Counterterrorism. Prior to assuming my current role, I served as lead security counsel for Facebook, working on issues ranging from children's safety to cybersecurity. And before that, I was a criminal prosecutor with the Department of Justice for 11 years in Chicago and Washington, DC, where I prosecuted Federal crimes including public corruption and gang violence. On behalf of Facebook, I want to thank you for your leadership in combating extremism, terrorism, and other threats to our National security.

I want to start by saying that all of us at Facebook stand with the victims, their families, and everyone affected by recent terrorist attacks, including the horrific violence in Sri Lanka and New Zealand. In the aftermath of such heinous acts, it is more important than ever to stand against hate and violence. We will continue to make that a priority in everything we do at Facebook. Facebook's mission is to give people the power to build community and bring the world closer together. We are proud that more than 2 billion people around the world come to Facebook every month to share with friends and family, to learn about new products and services, to volunteer or donate to organizations they care about, or to help in a crisis. But people need to feel safe in order to build this community. And that is why we are committed to fighting any efforts by terrorist groups to use Facebook. That is also why Facebook has rules against inciting violence, bullying, harassing, and threatening others. Our goal is to ensure that Facebook is a place where both expression and personal safety are protected and respected.

II. FACEBOOK'S EFFORTS TO COMBAT TERRORISM

On terrorist content, our view is simple: There is absolutely no place on Facebook for terrorism. Our long-standing Dangerous Individuals and Organizations policy bans any organization or individual that proclaims a violent mission or has engaged in acts of violence, including terrorist activity and organized hate. Regardless of whether or not these individuals or groups post content that would violate our policies, we remove their accounts as soon as we find them. They simply are not allowed to use our services under any circumstances. Furthermore, we remove any content that praises or supports terrorists or their actions whenever we become aware of it, and when we uncover evidence of imminent harm, we promptly inform authorities.

We recognize the challenges associated with fighting on-line extremism, and we are committed to being part of the solution. We are working to address these threats in three ways: Through products that help us stop terrorists at the gate, people who help us implement our policies, and partnerships outside the company which can help us stay ahead of the threat.

A. Products

One of the challenges we face is identifying the small fraction of terrorist content—less than 0.03 percent—posted to a platform used by more than 2 billion people every month. Facebook has invested significantly in technology to help meet this challenge and to identify proactively terrorist content, including through the use of artificial intelligence (AI) and other automation. These technologies have become increasingly central to keeping hateful or violent content off of Facebook.

Importantly, we do not wait for ISIS or al-Qaeda to upload content to Facebook before placing it into our internal detection systems. Instead, we proactively go after it. We contract with groups like SITE Intelligence and the University of Alabama at Birmingham to find propaganda released by these groups before it ever hits our site. We put this content, and other content we are able to identify from elsewhere on the internet, into our matching systems. And once we are aware of a piece of terrorist content, we remove it. We know that terrorists adapt as technology evolves, and that is why we constantly update our technical solutions. We use these solutions, as well as human expertise, so we can stay ahead of terrorist activity on our platform. We have provided information on our enforcement techniques in the past, and I would like to describe in broad terms some new tactics and methods that are proving effective.

1. Machine Learning Tools

We use machine learning to assess Facebook posts that may signal support for ISIS or al-Qaeda. Our machine learning tools produce a score indicating the likelihood that the post violates our counterterrorism policies, which, in turn, helps our team of reviewers prioritize reviewing posts with the highest scores. The system ensures that our reviewers are able to focus on the most important content first. And when the tool is sufficiently confident that a post contains support for terrorism, we automatically and immediately remove that post. We have seen real gains as a result of our efforts; for example, prioritization powered by our new machine learning tools has been critical to reducing significantly the amount of time terrorist content reported by our users stays on the platform.

2. Changes To Facebook Live

Facebook has also made changes to Facebook Live in response to the tragic events in Christchurch. We now restrict users from using Facebook Live if they have violated certain rules—including our Dangerous Organizations and Individuals policy. We apply a "one strike" policy to Live: Anyone who violates our most serious policies will be restricted from using Live for set periods of time—for example, 30 days—starting on their first offense. And we are working on extending these restrictions in the weeks to come, beginning with preventing those same people from creating ads on Facebook.

3. Improvements To Existing Tools And Partnerships

We have improved several of our existing proactive techniques and are now able to detect more effectively terrorist content. For example, our tools to algorithmically identify violating text posts (what we refer to as "language understanding") now work across 19 languages. Similarly, though we have long used image- and video-hashing—which converts a file into a unique string of digits that serves as a "fingerprint" of that file—we now also use audio- and text-hashing techniques for detecting terrorist content.

These improvements in our technical tools and partnerships have allowed for continued and sustained progress in finding and removing terrorist content from Facebook. Since the beginning of 2018, we have taken action on more than 25 million pieces of terrorist content, and we found over 99 percent of that content before any user reported it.

B. People

We know that we cannot rely on AI alone to identify terrorist content. Context often matters. To understand more nuanced cases, we need human expertise. One of our greatest human resources is our community of users. Our users help us by reporting accounts or content that may violate our policies—including the small fraction that may be related to terrorism. To review those reports, and to prioritize the safety of our users and our platform more generally, including with respect to counterterrorism, we have more than 30,000 people working on safety and security across the company and around the world. That is three times as many people as we had dedicated to such efforts in 2017. Our safety and security professionals review reported content in more than 50 languages, 24 hours a day. Within our safety and security team, we have also significantly grown our team of dedicated counterterrorism specialists. Distinct from our content review teams, we have more than 300 highly-trained professionals who are exclusively or primarily focused on preventing terrorist content from ever appearing on our platform and quickly identifying and removing it if it does. This team includes counterterrorism experts, former prosecutors, and law enforcement personnel. Together, they speak over 30 languages and are working 24 hours a day around the world to detect and remove terrorist content.

Because our reviewers are human, our performance is not always perfect. We make mistakes. And sometimes we are slower to act than we want to be. But keeping our platform and our users safe is one of Facebook's highest priorities, and we are always working to improve.

C. Partnerships

We are proud of the work we have done to make Facebook a hostile place for terrorists. We understand, however, that simply working to keep terrorism off Facebook is not an adequate solution to the problem of on-line extremism, particularly because terrorists are able to leverage a variety of platforms. We believe our partnerships with others—including other companies, civil society, researchers, and governments—are crucial to combating this threat.

In 2017, Facebook co-launched the Global Internet Forum to Counter Terrorism (GIFCT) with YouTube, Microsoft, and Twitter. The GIFCT shares information between the participants and has trained 110 companies from around the globe. Just last week, we held an event in Jordan that brought together more than 100 people from Government, industry, and civil society to share best practices.

Through GIFCT we expanded a database—which now contains hashes for more than 200,000 visually distinct images or videos—in which 15 companies share "hashes," or digital fingerprints, to better enable companies to identify noxious terrorist content.

Facebook took over as the chair of the GIFCT in 2019 and we have worked to expand its capabilities, including increasing hash sharing. In fact, we are freely providing our hashing technology to companies participating in the consortium.

Our efforts to work with others in the industry to tackle the on-line terrorist threat go further still. On May 15, 2019, in the wake of the tragic New Zealand attacks, Facebook and other tech companies, including Google, Twitter, Microsoft, and Amazon, signed the Christchurch Call to Action. The Christchurch Call expands on the GIFCT and builds on our other initiatives with Government and civil society to prevent the dissemination of terrorist and violent extremist content.

Facebook joined with others in the industry to commit to a 9-point plan that sets out concrete steps the industry will take to address the spread of terrorist content. Those steps are:

(1) *Terms of Use.*—We commit to updating our terms of use, community standards, codes of conduct, and acceptable use policies to expressly prohibit the distribution of terrorist and violent extremist content.

(2) *User Reporting of Terrorist and Violent Extremist Content.*—We commit to establishing one or more methods within our on-line platforms and services for users to report or flag inappropriate content, including terrorist and violent extremist content. We will ensure that the reporting mechanisms are clear, conspicuous, and easy to use, and provide enough categorical granularity to allow the company to prioritize and act promptly upon notification of terrorist or violent extremist content.

(3) *Enhancing Technology*.—We commit to continuing to invest in technology that improves our capability to detect and remove terrorist and violent extremist content on-line, including the extension or development of digital fingerprinting and AI-based technology solutions.

(4) *Livestreaming*.—We commit to identifying appropriate checks on livestreaming, aimed at reducing the risk of disseminating terrorist and violent extremist content on-line. These may include enhanced vetting measures and moderation where appropriate. Checks on livestreaming necessarily will be tailored to the context of specific livestreaming services, including the type of audience, the nature or character of the livestreaming service, and the likelihood of exploitation.

(5) *Transparency Reports*.—We commit to publishing on a regular basis transparency reports regarding detection and removal of terrorist or violent extremist content on our on-line platforms and services and ensuring that the data is supported by a reasonable and explainable methodology.

(6) *Shared Technology Development*.—We commit to working collaboratively across industry, governments, educational institutions, and NGO's to develop a shared understanding of the contexts in which terrorist and violent extremist content is published and to improve technology to detect and remove terrorist and violent extremist content including by creating robust data sets to improve AI, developing open-source or other shared tools to detect and remove content, and by enabling all companies to contribute to the effort.

(7) *Crisis Protocols*.—We commit to working collaboratively to respond to emerging or active events on an urgent basis, so relevant information can be quickly and efficiently shared, processed, and acted upon by all stakeholders with minimal delay. This includes the establishment of incident management teams that coordinate actions and broadly distribute information that is in the public interest.

(8) *Education*.—We commit to working collaboratively to help understand and educate the public about terrorist and extremist violent content on-line. This includes educating and reminding users about how to report or otherwise not contribute to the spread of this content on-line.

(9) *Combating Hate and Bigotry*.—We commit to working collaboratively across industry to attack the root causes of extremism and hate on-line. This includes providing greater support for relevant research—with an emphasis on the impact of on-line hate on off-line discrimination and violence—and supporting the capacity and capability of NGO's working to challenge hate and promote pluralism and respect on-line.

Our work to combat terrorism is not done. Terrorists come in many ideological stripes—and the most dangerous among them are deeply resilient. At Facebook, we recognize our responsibility to counter this threat and remain committed to it. But we should not view this as a problem that can be "solved" and set aside, even in the most optimistic scenarios. We can reduce the presence of terrorism on mainstream social platforms, but eliminating it completely requires addressing the people and organizations that generate this material in the real world.

III. FIGHTING OTHER HARMFUL CONTENT

Facebook recognizes that terrorist content is not the only threat to our users' safety and well-being. There will always be people who try to use our platforms to spread hate. And we have seen foreign actors trying to interfere with elections by sowing division and spreading false information. We are also working to address new tools of distortion, including manipulated media. We are developing technologies to identify manipulated content, dramatically reduce its distribution, and provide additional context to inform our community about its falsity. And we have partnered with outside fact checkers, researchers, and our colleagues across the industry to help with these efforts. We know that people want to see accurate information on Facebook, so we will continue to make fighting misinformation a priority.

We take all of these problems very seriously. Hate of any kind has no place on Facebook. Any organization or individual that espouses violence or hatred violates our standards. A few months ago, we updated our policies to make it clear that all praise, support, and representation of white nationalism or white separatism, in addition to white supremacy, violates our rules. Any such content is removed from our platform under our Dangerous Organizations and Individuals policy. And Facebook does not tolerate attempts to undermine the integrity of an election or suppress voter turnout. These issues are difficult, but we will continue to work to craft policies that protect people; to apply those policies consistently and without bias; and to give voice to a community that transcends regions, cultures, and languages.

IV. CONCLUSION

Security is an arms race and our adversaries are always evolving their tactics. We constantly have to improve to stay ahead. Though we will never be perfect, we have made progress. And we are committed to tirelessly combating extremism on our platform by regularly reviewing our policies, adopting technical solutions, and strengthening our partnerships with external stakeholders. I appreciate the opportunity to be here today, and I look forward to your questions.

Chairman THOMPSON. I thank you.

I now recognize the gentleman, Mr. Pickles, for 5 minutes.

STATEMENT OF NICK PICKLES, GLOBAL SENIOR STRATEGIST FOR PUBLIC POLICY, TWITTER

Mr. PICKLES. Chairman Thompson, Ranking Member Rogers, Members of the committee, thank you for the opportunity to appear here today to discuss these important issues of combating terrorist content on-line and manipulation for the public conversation.

We keep the victims, their families, and the affected communities of the attack in Christchurch and around the world in our minds as we undertake this important work.

We have made the health of Twitter our top priority and measure our efforts by how successfully we encourage healthy debates, conversations, and critical thinking on the platform. Conversely, hateful conduct, terrorist content, and deceptive practices detract from the health of the platform.

I would like to begin by outlining three key policies.

First, Twitter takes a zero-tolerance approach to terrorist content on our platform. Individuals may not promote terrorism, engage in terrorist recruitment or terrorist acts.

Since 2015, we have suspended more than 1.5 million accounts for violations of our rules related to the promotion of terrorism and continue to see more than 90 percent of these accounts suspended through proactive measures. In the majority of cases, we take action at the account-creation stage before an account has even tweeted. The remaining 10 percent is identified through a combination of user reports and partnerships.

Second, we prohibit the use of Twitter by violent extremist groups. These are defined in our rules as groups who, whether by their statements on or off the platform, promote violence against civilians or use violence against civilians to further their cause, whatever their ideology. Since the introduction of this policy in 2017, we have taken action on 184 groups globally and permanently suspended more than 2,000 unique accounts.

Third, Twitter does not allow hateful conduct on its service. An individual on Twitter is not permitted to promote violence or directly attack or threaten people based on protected characteristics. Where any of these rules are broken, we will take action to remove the content and will permanently remove those who promote terrorism or violent extremist groups on Twitter.

As you have heard, Twitter is a member of the Global Internet Forum to Counter Terrorism, a partnership between YouTube, Twitter, Facebook, and Microsoft that facilitates information sharing and technical information across industry, as well as providing essential export for smaller companies.

We learned a number of lessons from the Christchurch attacks. The distribution of media was manifestly different from how IS or other terror organizations worked. This reflects a change in the wider threat environment that requires a renewed approach and a focus on crisis response.

After Christchurch, an array of individuals on-line sought to continuously re-upload the content created by the attacker, both the video and the manifesto. The broader internet ecosystem presented then and still presents a challenge we cannot avoid. A range of third-party services were used to share content, including some forums and websites that have long hosted some of the most egregious content available on-line.

Our analysis found that 70 percent of the views of the video posted by the Christchurch attacker came from verified accounts on Twitter, including news organizations and individuals posting the video to condemn the attack. We are committed to learning and improving, but every entity has a part to play.

We should also take some heart from the social examples we have seen on Twitter around the world as users come together to challenge hate and challenge division. Hashtags like #PrayForOrlando, #JeSuisCharlie, or, after the Christchurch attack, #HelloBrother reject terrorist narratives and offer a better future for us all.

In the months since the attack, governments, industry, and civil society have united behind our mutual commitments to a safe, secure, open, and global internet. In fulfilling our commitment to the Christchurch call, we will take a wide range of actions, including to continue investing in technology so we can respond as quickly as possible to a future instance.

Let me now turn to our approach to dealing with attempts to manipulate the public conversation.

As a uniquely open service, Twitter enables the clarification of falsehoods in real-time. We proactively enforce policies and use technology to halt the spread of content propagated through manipulated tactics. Our rules clearly prohibit coordinated account manipulation, malicious automation, and fake accounts.

We continue to explore how we may take further action, through both policy and products, on these types of issues in the future. We continue to critically examine additional safeguards we can implement to protect the health of the conversation occurring on Twitter. We look forward to working with the committee on these important issues.

Thank you.

[The prepared statement of Mr. Pickles follows:]

Prepared Statement of Nick Pickles

June 26, 2019

Chairman Thompson, Ranking Member Rogers, and Members of the committee: Twitter's purpose is to serve the public conversation. Twitter is a place where people from around the world come together in an open and free exchange of ideas. My statement today will provide information and deeper context on: (I) Twitter's work to protect the health of the public conversation, including combating terrorism, violent extremist groups, hateful conduct, and platform manipulation, and (II) our partnerships and societal engagement.

I. TWITTER'S WORK TO PROTECT THE HEALTH OF THE PUBLIC CONVERSATION

All individuals accessing or using Twitter's services must adhere to the policies set forth in the Twitter Rules. Accounts under investigation or which have been detected as sharing content in violation with the Twitter Rules may be required to remove content, or in serious cases, will see their account permanently suspended. Our policies and enforcement options evolve continuously to address emerging behaviors on-line.

A. Policy on Terrorism

Individuals are prohibited from making specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism.

We have now suspended more than 1.5 million accounts for violations related to the promotion of terrorism between August 1, 2015, and December 31, 2018. In 2018, a total of 371,669 accounts were suspended for violations related to promotion of terrorism. We continue to see more than 90 percent of these accounts suspended through proactive measures.

The trend we are observing year-over-year is a steady decrease in terrorist organizations attempting to use our service. This is due to zero-tolerance policy enforcement that has allowed us to take swift action on ban evaders and other identified forms of behavior used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage—before the account even tweets.

Government reports constituted less than 0.1 percent of all suspensions in the last reporting period. Continuing the trend we have seen for some time, the number of reports we received from governments of terrorist content from the second half of last year decreased by 77 percent compared to the previous reporting period (January–June 2018).

We are reassured by the progress we have made, including recognition by independent experts. For example, Dublin City University Professor Maura Conway found in a detailed study that "ISIS's previously strong and vibrant Twitter community is now . . . virtually non-existent."

In tandem with removing content, our wider efforts on countering violent extremism going back to 2015 have focused on bolstering the voices of non-Governmental organizations and credible outside groups to use our uniquely open service to spread positive and affirmative campaigns that seek to offer an alternative to narratives of hate.

We have partnered with organizations delivering counter and alternative narrative initiatives across the globe and we encourage the committee to consider the role of Government in supporting the work of credible messengers in this space at home and abroad.

B. Policy on Violent Extremist Groups

In December 2017, we broadened our rules to encompass accounts affiliated with violent extremist groups. Our prohibition on the use of Twitter's services by violent extremist groups—i.e., identified groups subscribing to the use of violence as a means to advance their cause—applies irrespective of the cause of the group.

Our policy states:

Violent extremist groups are those that meet all of the below criteria:

- identify through their stated purpose, publications, or actions as an extremist group;
- have engaged in, or currently engage in, violence and/or the promotion of violence as a means to further their cause; and
- target civilians in their acts and/or promotion of violence.

An individual on Twitter may not affiliate with such an organization—whether by their own statements or activity both on and off the service—and we will permanently suspend those who do so.

We know that the challenges we face are not static, nor are bad actors homogenous from one country to the next in how they behave. Our approach combines flexibility with a clear, consistent policy philosophy, enabling us to move quickly while establishing clear norms of unacceptable behavior.

Since the introduction of our policy on violent extremist groups, we have taken action on 184 groups under this policy and permanently suspended 2,182 unique accounts. Ninety-three of these groups advocate violence against civilians alongside some form of extremist white supremacist ideology.

C. Policy on Hateful Conduct

People on Twitter are not permitted to promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm toward others on the basis of these categories.

We do not allow individuals to use hateful images or symbols in their profile image or profile header. Individuals on the platform are not allowed to use the user name, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate toward a person, group, or protected category.

Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, or referring to someone by their full name.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behavior, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

D. Manipulation of the Public Conversation

Our policies regarding terrorism, violent extremist groups, and hateful conduct are strictly enforced, as are all our policies. We take additional steps to safeguard the public conversation from manipulation.

As a uniquely open, public service, the clarification of falsehoods could happen in seconds on Twitter. We proactively enforce policies and use technology to halt the spread of content propagated through manipulative tactics, such as automation or attempting to deliberately game trending topics.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. In partnership with teams across the company, we employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it. We also partner closely with governments, law enforcement, academics, researchers, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

For example, we typically challenge 8 to 10 million accounts per week for these behaviors, requesting additional details, like email addresses and phone numbers in order to authenticate the account. We also recently acquired a new business to augment our efforts in this regard. This strategic investment will be a key driver as we work to protect the public conversation and help all individuals on our service see relevant information.

Attempts to execute misinformation campaigns rely on tactics like coordinated account manipulation or malicious automation—all of which are against Twitter's Rules. We are continuing to explore ways at how we may take action—through both policy and product—on these types of issues in the future. We continue to critically examine additional safeguards we can implement to protect the conversation occurring on Twitter.

In October 2018, we published the first comprehensive archive of tweets and media associated with known state-backed information operations on Twitter and since then we have provided two further updates covering a range of actors. Thousands of researchers from across the globe have now made use of these datasets, which contain more than 30 million tweets and more than 1 terabyte of media, using our archive to conduct their own investigations and to share their insights and independent analysis with the world.

By making this data open and accessible, we seek to empower researchers, journalists, governments, and members of the public to deepen their understanding of critical issues impacting the integrity of public conversation on-line, particularly around elections. This transparency is core to our mission.

E. Investing in Tech: Behavior vs. Content

Twitter's philosophy is to take a behavior-led approach, utilizing a combination of machine learning and human review to prioritize reports and improve the health of the public conversation. That is to say, we increasingly look at how accounts behave before we look at the content they are posting. This is how we seek to scale our efforts globally and leverage technology even where the language used is highly context-specific. Twitter employs extensive content detection technology to identify potentially abusive content on the service, along with allowing users to report content to us either as an individual or a bystander.

We have made the health of Twitter our top priority, and our efforts will be measured by how we help encourage more healthy debate, conversations, and critical thinking on the platform. Conversely, abuse, automation, hateful conduct, terrorism, and manipulation will detract from the health of our platform.

For abuse, this two-pronged strategy has allowed us to take 3 times the amount of enforcement of action on abuse within 24 hours than this time last year. We now proactively surface nearly 40 percent of abusive content we remove compared to 20 percent a year ago to reduce the burden on the individual. Since we started using machine learning 3 years ago to reduce the visibility on abusive content:

- 80 percent of all replies that are removed were already less visible;
- Abuse reports have been reduced by 7.6 percent;
- The most visible replies receive 45 percent less abuse reports;
- 100,000 accounts were suspended for creating new accounts after a suspension during January through March 2019—a 45 percent increase from the same time last year;
- 60 percent faster response to appeals requests with our new in-app appeal process;
- 3 times more abusive accounts suspended within 24 hours after a report compared to the same time last year; and
- 2.5 times more private information removed with a new, easier reporting process.

II. PARTNERSHIPS AND SOCIETAL ENGAGEMENT

We work closely with the Federal Bureau of Investigation, along with law enforcement and numerous public safety around the world. As our partnerships deepen, we are able to better respond to the changing threats we all face, sharing valuable information and promptly responding to valid legal requests for information.

A. Industry Collaboration

Collaboration with our industry peers and civil society is also critically important to addressing common threats from terrorism globally. In June 2017, we launched the Global Internet Forum to Counter Terrorism (the "GIFCT"), a partnership among Twitter, YouTube, Facebook, and Microsoft.

The GIFCT facilitates, among other things, information sharing; technical cooperation; and research collaboration, including with academic institutions. In September 2017, the members of the GIFCT announced a significant financial commitment to support research on terrorist abuse of the internet and how governments, tech companies, and civil society can respond effectively. Our goal is to establish a network of experts that can develop platform-agnostic research questions and analysis that consider a range of geopolitical contexts.

Technological collaboration is a key part of GIFCT's work. In the first 2 years of GIFCT, two projects have provided technical resources to support the work of members and smaller companies to remove terrorist content.

First, the shared industry data base of "hashes"—unique digital "fingerprints"—for violent terrorist propaganda now spans more than 100,000 hashes. The database allows a company that discovers terrorist content on one of its sites to create a digital fingerprint and share it with the other companies in the forum, who can then use those hashes to identify such content on their services or platforms, review against their respective policies and individual rules, and remove matching content as appropriate or block extremist content before it is posted.

Second, a year ago, Twitter began working with a small group of companies to test a new collaborative system. Because Twitter does not allow files other than photos or short videos to be uploaded, one of the behaviors we saw from those seeking to promote terrorism was to post links to other services where people could access files, longer videos, PDFs, and other materials. Our pilot system allows us to alert other companies when we removed an account or Tweet that linked to material that promoted terrorism hosted on their service. This information sharing ensures the hosting companies can monitor and track similar behavior, taking enforcement action pursuant with their individual policies. This is not a high-tech approach, but it is simple and effective, recognizing the resource constraints of smaller companies.

Based on positive feedback, the partnership has now expanded to 12 companies and we have shared more than 14,000 unique URLs with these services. Every time a piece of content is removed at source, it means any link to that source—wherever it is posted—will no longer be operational.

We are eager to partner with additional companies to expand this project, and we look forward to building on our existing partnerships in the future.

B. The Christchurch Attack

The Christchurch attack was unprecedented both in the way it exploited the online environment but also the disparate range of on-line communities that were involved in sharing the Christchurch video and the hateful manifesto of the attacker.

We saw a wide range of individuals on the service continue to upload excerpts of the attacker's video even after it had been removed. This included those who sought to condemn the attack, including those who combined video of their condemnation and prayers with video of the attack, and others who saw baseless conspiracies and wanted to provide evidence to refute such claims. There were those who believed to remove the content was censorship and those who wanted to amplify the hatred the video embodied. Our analysis found 70 percent of the views of footage of the attack in Christchurch on Twitter were from content posted by verified accounts, including media outlets and those seeking to condemn the violence. In all of these circumstances we removed the relevant content.

As a uniquely open service, we see regular examples around the world of our users, communities, and groups challenging hate and division, particularly following violent acts. As the world began to comprehend the horror of what took place in Christchurch, some may have sought to promote hate, but there was another conversation taking place, one that reached many more people. The hashtag #HelloBrother saw people around the world recognizing the brave act of one victim and rejecting the terrorist's narrative, while hundreds of thousands of tweets expressed similar sentiments in their own way. This is the potential of open public conversation and what it can empower—a global platform for the best of society to challenge violence and hatred.

C. THE CHRISTCHURCH CALL TO ACTION

In the months since the attack, New Zealand Prime Minister Jacinda Ardern has led the international policy debate, and that work has culminated in the Christchurch Call. Twitter's Chief Executive Officer Jack Dorsey attended the launch of the Christchurch Call in Paris, meeting with the Prime Minister to express our support and partnership with the New Zealand Government.

Because terrorism cannot be solved by the tech industry alone, the Christchurch Call is a landmark moment and an opportunity to convene governments, industry, and civil society to unite behind our mutual commitment to a safe, secure open, global internet. It is also a moment to recognize that however or wherever evil manifests itself, it affects us all.

In fulfilling our commitments in the Call, we will take a wide range of actions. We continue to invest in technology to prioritize signals, including user reports, to ensure we can respond as quickly as possible to a potential incident, building on the work we have done to harness proprietary technology to detect and disrupt bad actors proactively.

As part of our commitment to educate users about our rules and to further prohibit the promotion of terrorism or violent extremist groups, we have updated our rules and associated materials to be clearer on where these policies apply. This is accompanied by further data being provided in our transparency report, allowing public consideration of the actions we are taking under our rules, as well as how much content is detected by our proactive efforts.

Twitter will take concrete steps to reduce the risk of livestreaming being abused by terrorists, while recognizing that during a crisis these tools are also used by news organizations, citizens, and governments. We are investing in technology and tools to ensure we can act even faster to remove video content and stop it spreading.

Finally we are committed to continuing our partnership with industry peers, expanding on our URL-sharing efforts along with wider mentoring efforts, strengthening our new crisis protocol arrangements, and supporting the expansion of GIFCT membership.

D. A WHOLE-OF-SOCIETY RESPONSE

The challenges we face as a society are complex, varied, and constantly evolving. These challenges are reflected and often magnified by technology. The push and pull factors influencing individuals vary widely and there is no one solution to prevent an individual turning to violence. This is a long-term problem requiring a long-term response, not just the removal of content.

While we strictly enforce our policies, removing all discussion of particular viewpoints, no matter how uncomfortable our customers may find them, does not eliminate the ideology underpinning them. Quite often, it moves these views into darker corners of the internet where they cannot be challenged and held to account. As our peer companies improve in their efforts, this content continues to migrate to less-governed platforms and services. We are committed to learning and improving, but every part of the on-line ecosystem has a part to play.

We have a critical role. Tech companies and content removal on-line cannot alone, however, solve these issues. They are systemic and societal and so they require an whole-of-society approach. We welcome the opportunity to continue to work with our industry peers, Government, academics, and civil society to find the right solutions.

Our goal is to protect the health of the public conversation and to take immediate action on those who seek to spread messages of terror and violent extremism. However, no solution is perfect, and no technology is capable of detecting every potential threat.

Twitter's efforts around the globe to support civil society voices and promote positive messages have seen Twitter employees train groups on 5 continents and we have provided pro-bono advertising to groups to enable their messages to reach millions of people. When we at Twitter talk about the health of the public conversation, we see the principles of civility, empathy, and mutual respect as foundational to our work. We will not solve problems by removing content alone. We should not underestimate the power of open conversation to change minds, perspectives, and behaviors.

We stand ready to assist the committee in its important work regarding the issue of the tools that internet companies can employ to stop the spread of terrorist content and misinformation on our services.

Chairman THOMPSON. Thank you for your testimony.

I now recognize Mr. Slater to summarize his testimony for 5 minutes.

STATEMENT OF DEREK SLATER, GLOBAL DIRECTOR OF INFORMATION POLICY, GOOGLE

Mr. SLATER. Chairman Thompson, Ranking Member Rogers, and distinguished Members of the committee, thank you for the opportunity to appear before you today. I appreciate your leadership on the important issues of radicalization and misinformation on-line and welcome the opportunity to discuss Google's work in these areas.

My name is Derek Slater, and I am the global director of information policy at Google. In my role, I lead a team that advises the company on public policy frameworks for on-line content.

At Google, we believe that the internet has been a force for creativity, learning, and access to information. Supporting the free flow of ideas is core to our mission: To organize and make the world's information universally accessible and useful.

Yet there have always been legitimate limits, even where laws strongly protect free expression. This is true both on-line and off, especially when it comes to issues of terrorism, hate speech, and misinformation. We take these issues seriously and want to be a part of the solution.

In my testimony today, I will focus on two areas where we are making progress to help protect our users: First, on the enforcement of our policies around terrorism and hate speech; and, second, in combating misinformation broadly.

On YouTube, we have rigorous policies and programs to defend against the use of our platform to spread hate or incite violence. Over the past 2 years, we have invested heavily in machines and people to quickly identify and remove content that violates our policies.

First, YouTube's enforcement system starts from the point at which a user uploads a video. If it is somewhat similar to videos that already violate our policies, it is sent for humans to review. If they determine that it violates our policies, they remove it, and the system makes a digital fingerprint so it can't be uploaded again.

In the first quarter of 2019, over 75 percent of the more than 8 million videos removed were first flagged by a machine, the majority of which were removed before a single view was received.

Second, we also rely on experts to find videos that the algorithm might be missing. Some of these experts sit at our intel desk, which proactively looks for new trends in content that might violate our policies. We also allow expert NGO's and governments to notify us of bad content in bulk through our Trusted Flagger program.

Finally, we go beyond enforcing our policies by creating programs to promote counter-speech. Examples of this work include our Creators for Change program, which supports YouTube creators that are acting as positive role models. In addition, Alphabet's Jigsaw group has deployed the redirect method, which uses targeted ads and videos to disrupt on-line radicalization.

This broad and cross-sectional work has led to tangible results. In the first quarter of 2019, YouTube manually reviewed over 1 million suspected terrorist videos and found that only fewer than 10 percent, about 90,000, violated our terrorism policy. As a comparison point, we typically remove between 7 million and 9 million videos per quarter, which is a tiny fraction of a percent of YouTube's total views during this time period.

Our efforts do not stop there. We are constantly taking input and reacting to new situations. For example, YouTube recently further updated its hate speech policy. The updated policy specifically prohibits videos alleging that a group is superior in order to justify discrimination, segregation, or exclusion based on qualities like age, gender, race, caste, religion, sexual orientation, or veteran status.

Similarly, the recent tragic events in Christchurch presented some unprecedented challenges. In response, we took more drastic measures, such as automatically rejecting new uploads of videos without waiting for human review to check if it was news content. We are now reexamining our crisis protocols and have also signed the Christchurch Call to Action.

Finally, we are deeply committed to working with Government, the tech industry, and experts from civil society and academia to protect our services from being exploited by bad actors, including during Google's chairmanship of the GIFCT over the last year and a half.

On the topic of combating misinformation, we have a natural long-term incentive to prevent anyone from interfering with the integrity of our products. We also recognize that it is critically important to combat misinformation in the context of democratic elections, when our users seek accurate, trusted information that will help them make critical decisions.

We have worked hard to curb misinformation in our products, and our efforts include designing better ranking algorithms, implementing tougher policies against monetization of misrepresentative content, and deploying multiple teams that identify and take action against malicious actors.

At the same time, we have to be mindful that our platforms reflect a broad array of sources and information, and there are important free-speech considerations. There is no silver bullet, but we will continue to work to get it right.

In conclusion, we want to do everything we can to ensure users are not exposed to harmful content. We understand these are difficult issues of serious interest to the committee. We take them seriously and want to be responsible actors who do our part.

Thank you for your time, and I look forward to taking your questions.

[The prepared statement of Mr. Slater follows:]

Prepared Statement of Derek Slater

June 26, 2019

Chairman Thompson, Ranking Member Rogers, and distinguished Members of the committee: Thank you for the opportunity to appear before you today. I appreciate your leadership on the important issues of radicalization and misinformation online, and welcome the opportunity to discuss Google's work in these areas.

My name is Derek Slater, and I am the global director of information policy at Google. In my role, I lead a team that advises the company on public policy frameworks for on-line content—including hate speech, terrorism, and misinformation. Prior to my role at Google, I worked on internet policy at the Electronic Frontier Foundation and at the Berkman Center for Internet and Society.

At Google, we believe that the internet has been a force for creativity, learning, and access to information. Supporting this free flow of ideas is core to our mission to organize and make the world's information universally accessible and useful. We build tools that empower users to access, create, and share information like never before giving them more choice, opportunity, and exposure to a diversity of opinions. Products like YouTube, for example, have expanded economic opportunity for small businesses to market and sell their goods; have given artists, creators, and journalists a platform to share their work, connect with an audience, and enrich civic discourse; and have enabled billions to benefit from a bigger, broader understanding of the world.

While the free flow of information and ideas has important social, cultural, and economic benefits, there have always been legitimate limits, even where laws strongly protect free expression. This is true both on-line and off, especially when it comes to issues of terrorism, hate speech, and misinfomation. We are deeply troubled by the increase in hate and violence in the world, particularly by the acts of terrorism and violent extremism in New Zealand. We take these issues seriously and want to be a part of the solution.

This is why, in addition to being guided by local law, we have Community Guidelines our users have to follow. We also work closely with Government, industry, and civil society to address these challenges in partnership within the United States and around the world. In my testimony today, I will focus on two key areas where we are making progress to help protect our users: (i) On the enforcement of our policies around terrorism and hate speech; and (ii) in combatting misinformation broadly.

ENFORCEMENT ON YOUTUBE FOR TERRORISM AND HATE SPEECH

We have rigorous policies and programs to defend against the use of our platform to spread hate or incite violence. This includes: Terrorist recruitment, violent extremism, incitement to violence, glorification of violence, and videos that teach people how to commit terrorist attacks. We apply these policies to violent extremism of all kinds, whether inciting violence on the basis of race or religion or as part of an organized terrorist group.

Tough policies have to be coupled with tough enforcement. Over the past 2 years, we have invested heavily in machines and people to quickly identify and remove content that violates our policies against incitement to violence and hate speech:

(1) YouTube's enforcement system starts from the point at which a user uploads a video. If it is somewhat similar to videos that already violate our policies, it is sent for humans to review. If they determine that it violates our policies, they remove it and the system makes a "digital fingerprint" or hash of the video so it can't be uploaded again. In the first quarter of 2019, over 75 percent of the

more than 8 million videos removed were first flagged by a machine, the majority of which were removed before a single view was received.

(2) Machine learning technology is what helps us find this content and enforce our policies at scale. But hate and violent extremism are nuanced and constantly evolving, which is why we also rely on experts to find videos the algorithm might be missing. Some of these experts sit at our intel desk, which proactively looks for new trends in content that might violate our policies. We also allow expert NGO's and governments to notify us of bad content in bulk through our Trusted Flagger program. We reserve the final decision on whether to remove videos they flag, but we benefit immensely from their expertise.

(3) Finally, we go beyond enforcing our polices by creating programs to promote counterspeech on our platforms to present narratives and elevate the voices that are most credible in speaking out against hate, violence, and terrorism.

(a) For example, our Creators for Change program supports creators who are tackling tough issues, including extremism and hate, by building empathy and acting as positive role models. There have been 59 million views of 2018 Creators for Change videos so far; the creators involved have over 60 million subscribers and more than 8.5 billion lifetime views of their channels; and through 'Local chapters' of Creators for Change, creators tackle challenges specific to different markets.

(b) Alphabet's Jigsaw group, an incubator to tackle some of the toughest global security challenges, has deployed the Redirect Method, which uses Adwords targeting tools and curated YouTube playlists to disrupt on-line radicalization. The method is open to anyone to use, and we know that NGO's have sponsored campaigns against a wide spectrum of ideologically-motivated terrorists.

This broad and cross-sectional work has led to tangible results. In Q1 2019, YouTube manually reviewed over 1 million suspected terrorist videos and found that only fewer than 10 percent (90K videos) violated our terrorism policy. Even though the amount of content we remove for terrorism is low compared to the overall amount our users and algorithms flag, we invest in reviewing all of it out of an abundance of caution. As comparison point, we typically remove between 7 and 9 million videos per quarter—a fraction of a percent of YouTube's total views during this time period. Most of these videos were first flagged for review by our automated systems. Over 90 percent of violent extremist videos that were uploaded and removed in the past 6 months (Q4 2018 & Q1 2019) were removed before receiving a single human flag, and of those, 88 percent had fewer than 10 views.

Our efforts do not end there. We are constantly taking input and reacting to new situations. For example, YouTube recently further updated its Hate Speech policy. The updated policy specifically prohibits videos alleging that a group is superior in order to justify discrimination, segregation, or exclusion based on qualities like age, gender, race, caste, religion, sexual orientation, or veteran status. This would include, for example, videos that promote or glorify Nazi ideology, which is inherently discriminatory. It also prohibits content denying that well-documented violent events, like the Holocaust or the shooting at Sandy Hook Elementary, took place. We began enforcing the updated policy the day it launched; however, it will take time for our systems to fully ramp up and we'll be gradually expanding coverage over the next several months.

Similarly, the recent tragic events in Christchurch presented some unprecedented challenges and we had to take some unprecedented steps to address the unprecedented volume of new videos related to the events—tens of thousands, exponentially larger than we had ever seen before, at times coming in as fast as one per second. In response, we took more drastic measures, such as automatically rejecting new uploads of clips of the video without waiting for human review to check if it was news content. We are now reexamining our crisis protocols, and we've been giving a lot of thought to what additional steps we can take to further protect our platforms against misuse. Google and YouTube also signed the Christchurch Call to Action, a series of commitments to quickly and responsibly address terrorist content on-line. The effort was spearheaded by New Zealand's prime minister to ensure another misuse of on-line platforms like this cannot happen again.

Finally, we are deeply committed to working with Government, the tech industry, and experts from civil society and academia to protect our services from being exploited by bad actors. During Google's chairmanship of the Global Internet Forum to Counter Terrorism over the last year-and-a-half, the Forum sought to expand its membership and to reach out to a wide variety of stakeholders to ensure we are responsibly addressing terrorist content on-line. For example, we hosted a summit in Sunnyvale so G7 security ministers could hear the concerns of smaller platforms We have also convened workshops with activists and civil society organizations to find ways to support their on-line counter-extremism campaigns, and sponsored workshops around the world to share good practices with other tech companies and platforms.

Combating Misinformation

We have a natural, long-term incentive to prevent anyone from interfering with the integrity of our products. We also recognize that it is critically important to combat misinformation in the context of democratic elections, when our users seek accurate, trusted information that will help them make critical decisions. We have worked hard to curb misinformation in our products. Our efforts include designing better-ranking algorithms, implementing tougher policies against monetization of misrepresentative content, and deploying multiple teams that identify and take action against malicious actors. At the same time, we have to be mindful that our platforms reflect a broad array of sources and information and there are important free-speech considerations. There is no silver bullet, but we will continue to work to get it right, and we rely on a diverse set of tools, strategies, and transparency efforts to achieve our goals.

We make quality count in our ranking systems in order to deliver quality information, especially in contexts that are prone to rumors and the propagation of false information (such as breaking news events). The ranking algorithms we develop to that end are geared toward ensuring the usefulness of our services, as measured by user testing. The systems are not designed to rank content based on its political perspective.

Since the early days of Google and YouTube, some content creators have tried to deceive our ranking systems in order to increase their visibility, a set of practices we view as a form of spam. To prevent spam and other improper activity during elections, we have multiple internal teams that identify malicious actors wherever they originate, disable their accounts, and share threat information with other companies and law enforcement officials. We will continue to invest resources to address this issue and to work with law enforcement, Congress, and other companies.

In addition to tackling spam, we invest in trust and safety efforts and automated tools to tackle a broad set of malicious behaviors. Our policies across Google Search, Google News, YouTube, and our advertising products clearly outline behaviors that are prohibited, such as misrepresentation of one's ownership or primary purpose on Google News and our advertising products, or impersonation of other channels or individuals on YouTube. We make these rules of the road clear to users and content creators, while being mindful not to disclose so much information about our systems and policies as to make it easier for malicious actors to circumvent our defenses.

Finally, we strive to provide users with easy access to context and a diverse set of perspectives, which are key to providing users with the information they need to form their own views. Our products and services expose users to numerous links or videos from different sources in response to their searches, which maximizes exposure to diverse perspectives or viewpoints before deciding what to explore in depth. In addition, we develop many tools and features to provide additional information to users about their searches, such as knowledge or information panels in Google Search and YouTube.

CONCLUSION

We want to do everything we can to ensure users are not exposed to content that promotes or glorifies acts of terrorism. Similarly, we also recognize that it is critically important to combat misinformation in the context of democratic elections, when our users seek accurate, trusted information that will help them make critical decisions. Efforts to undermine the free-flow of information is antithetical to our mission. We understand these are difficult issues of serious interest to the committee. We take them seriously and want to be responsible actors who are a part of the solution.

We know that our users will value our services only so long as they continue to trust them to work well and provide them with the most relevant and useful information. We believe we have developed a responsible approach to address the evolving and complex issues that manifest on our platform.

We look forward to continued collaboration with the committee as it examines these issues. Thank you for your time. I look forward to taking your questions.

Chairman THOMPSON. Thank you for your testimony.

I now recognize Ms. Strossen to summarize her statement for 5 minutes.

STATEMENT OF NADINE STROSSEN, JOHN MARSHALL HARLAN II PROFESSOR OF LAW, NEW YORK LAW SCHOOL

Ms. STROSSEN. Thank you so much, Chairman Thompson and Ranking Member Rogers and other Members of the committee.

My name is Nadine Strossen. I am a professor of law at New York Law School and the immediate past president of the American Civil Liberties Union.

Of great pertinence, last year, I wrote a book which is directly pertinent to the topic of this hearing called "Hate: Why We Should Resist It with Free Speech, Not Censorship."

I note, Mr. Chairman, that you referred to hate speech as problematic content in addition with terror content and misinformation. All of these kinds of speech, while potentially harmful, present enormous dangers when we empower either government or private companies to censor and suppress the speech for this reason: The concepts of hate speech, terrorist content, and misinformation are all irreducibly vague and broad, therefore having to be enforced according to the subjective discretion of the enforcing authorities. The discretion has been enforced in ways that both under-suppress speech that does pose a serious danger, as the Chairman pointed out and the Ranking Member pointed out, but also do suppress very important speech, as also has been pointed out, speech that actually counters terrorism and other dangers.

What is worse is that, in addition to violating free speech and democracy norms, these measures are not ineffective in dealing with the underlying problems. I thought that was something that was pointed out by comments by my co-panelists. In particular, Nick Pickles' testimony, written testimony, talked about the fact that, if somebody is driven off one of these platforms, they will then take refuge in darker corners of the web, where it is much harder to engage with them, to use them as sources of information for law enforcement and counterterrorism investigations.

So we should emphasize other approaches that are consistent with free speech and democracy but have been lauded as at least as effective and perhaps even more so than suppression. I was very heartened that the written statements of my co-panelists all emphasize these other approaches. Monika Bickert's testimony talked about how essential it is to go after the root causes of terrorism. The testimony of Nick Pickles and Derek Slater also emphasize the importance of counter-speech, counter-narratives, and redirection.

Now, I recognize that every single one of us in this room is completely committed to free speech and democracy, just as every single one of us is committed to countering terrorism and disinformation. After all, the reason we oppose terrorism and disinformation is precisely because of the harm that they do to democracy and liberty.

Before I say anything further, I do have to stress something that I know everybody here knows but many members of the public do not, that these social media companies are not bound by the First Amendment free-speech guarantee. So none of us has a free-speech right to air any content on their platforms at all. Conversely, they have their own free-speech rights to choose what will be and what will not be on their platforms. So it would be unconstitutional, of course, for Congress to purport to tell them what they must put up and what they must take down, to the extent that the takedowns would go beyond First Amendment-unprotected speech.

Chairman Thompson, you did completely accurately, of course, note that much of the content that is targeted as terrorist is unprotected, but much of it is protected under the Constitution, and much of it is very valuable, including human rights advocacy that has been suppressed under these necessarily overbroad and subjective standards.

Although the social media companies do not have a Constitutional obligation to honor freedom of speech, given their enormous power, it is incredibly important that they be encouraged to do so.

In closing, I am going to quote a statement from the written testimony of Nick Pickles which I could not agree with more, when he said that "we will not solve the problems by removing content alone. We should not underestimate the power of open conversation to change minds, perspectives, and behaviors."

Thank you very much.

[The prepared statement of Ms. Strossen follows:]

Prepared Statement of Nadine Strossen [1]

June 26, 2019

INTRODUCTION

Chairman Thompson, Ranking Member Rogers, and distinguished Members of this committee: I am honored to join my esteemed co-panelists in addressing this hearing's important topics. I appreciate the committee Members' and panelists' commitments to counter the potential * serious adverse impacts of the pertinent on-line expression: Expression that could promote terrorism; and misinformation, which could defraud individuals and distort elections and other democratic processes and institutions.

I thank the committee for exercising its important oversight functions to examine the content moderation policies of the powerful social media companies that are represented today.** Even though any direct regulation of these policies would raise serious concerns about abridging the companies' First Amendment rights, it is essential to consider how the companies should exercise those rights in ways that promote the free speech and other rights of the rest of us—and in ways that promote democracy, security, and other important concerns.

I understand that I was invited to complement this panel of social media leaders, despite my relative lack of specific experience with social media in particular, because of my longstanding scholarship and advocacy about freedom of speech for potentially harmful speech in general (including speech with terror content and misinformation) in many contexts, including on-line media.[2] For example, I was deeply involved in the developments leading to the historic 1997 Supreme Court case that first considered—and upheld—First Amendment free speech rights on-line: *Reno* v.

[1] Nadine Strossen is the John Marshall Harlan II professor of law at New York Law School and the immediate past national president of the American Civil Liberties Union (1991–2008). She gratefully acknowledges the following NYLS students for providing valuable assistance with this testimony, including the preparation of end notes: Managing research assistant Marc D. Walkow, and research assistants Aaron Hansen and Serene Qandil.

*I deliberately refer to the "potential" adverse impacts of expression with terrorist content and misinformation because many experts have concluded that such expression will not necessarily contribute to the feared potential harms, and that non-censorial strategies such as the ones I discuss can significantly reduce that potential danger.

**I am confining my focus to the dominant large companies, including the 3 that are represented at this hearing (Facebook, Google, and Twitter). They exercise outsize influence, thus as a practical matter requiring many people to use their services. Concerning smaller social media companies, potential users retain real choices about whether or not to participate. Accordingly, such smaller companies should (as a normative matter) have more latitude to choose content and define communities (again, as a legal matter, all of these companies have such latitude).

[2] See, e.g., Nadine Strossen, *HATE: Why We Should Resist It with Free Speech, Not Censorship* (New York: Oxford University Press, 2018).

ACLU.[3] I was the national ACLU president throughout all the pertinent developments, including lobbying and litigating against Congress's first on-line censorship law (enacted in 1996), which the high Court struck down, essentially unanimously. The Court celebrated the internet as "a unique . . . medium of worldwide human communication," whose "content . . . is as diverse as human thought."[4]

Today's discussion can gain much from the teachings of this landmark case, and from other past efforts to restrict various media expression feared to potentially cause harm. Not only did the Court strike down Congress's first internet censorship law in *Reno* v. *ACLU,* but it also struck down Congress's revised version of that law in subsequent rulings.[5] Likewise, in its most recent decision about on-line expression, 2 years ago, the Court again unanimously struck down a law restricting such expression (in that case, a State law).[6] Moreover, the Court again hailed the unique importance of on-line communications, declaring:

> "While in the past there may have been difficulty in identifying the most important places . . . for the exchange of views, today the answer is clear. It is cyber space—the 'vast democratic forums of the internet' in general, . . . and social media in particular."[7]

As the preceding outline of relevant Supreme Court rulings indicates, my support for on-line free expression is largely paralleled by the Court's speech-protective rulings, and those in turn reflect the views of Justices across the ideological spectrum. Despite all the polarization in our political system and society, these particular issues about on-line expression should garner broad consensus in the other branches of Government, as they have on the Court. Notwithstanding how divided our views might be on contested public policy issues, we all have the same stake in preserving the most robust freedom of speech for all such views—no matter how extreme, controversial, or generally feared such views might be. In fact, those of us who are engaged in public policy debates have the greatest stake in strong freedom of speech. As the Court consistently has held, speech on public policy issues is the most important expression in our political system, essential not only for individual freedom and equality, but also for our very democracy itself. In its words: "Speech concerning public affairs is more than self-expression; it is the essence of self-government."[8]

The speech at issue in today's hearings—speech with terrorist[***] content and misinformation—certainly concerns public affairs; indeed, that is precisely why it is potentially so harmful, as well as undeniably important. As the Court likewise has explained, such speech deserves the strongest protection not despite its potential serious harmful impact, but rather precisely because of such powerful potential. Let me quote a 2011 decision upholding freedom for extremely controversial, provocative speech (this decision was nearly unanimous, with only one dissenting vote):

> "Speech is powerful. It can stir people to action, move them to tears of both joy and sorrow, and—as it did here—inflict great pain. [W]e cannot react . . . by punishing the speaker. As a Nation we have chosen a different course—to protect . . . speech on public issues to ensure that we do not stifle public debate."[9]

OVERVIEW

I will first set out my three major conclusions about the important, challenging issues raised by today's hearing. I will then lay out some more specific points that reinforce these conclusions.

[3] *Reno* v. *American Civil Liberties Union,* 521 U.S. 844 (1997). Justice O'Connor authored a partial dissent, in which Justice Rehnquist joined, but this concerned only a narrow particular application of the statute (as applied to an on-line communication involving only one adult and one or more minors, such as when an adult knowingly sends an email to a minor); both of these Justices agreed with the majority's broad holdings about the law's general unconstitutionality, making the decision essentially unanimous. Reno, 521 U.S. at 886 (O'Connor, J., concurring in part and dissenting in part).

[4] Reno, 521 U.S. at 850, 852.

[5] *Ashcroft* v. *American Civil Liberties Union,* 535 U.S. 564 (2002); *Ashcroft* v. *American Civil Liberties Union,* 542 U.S. 656 (2004), cert. denied *Mukasey v. American Civil Liberties Union,* 2009 U.S. LEXIS 598 (2009).

[6] *Packingham* v. *North Carolina,* 137 S. Ct. 1730 (2017).

[7] Packingham, 137 S. Ct. at 1735.

[8] *Garrison* v. *Louisiana,* 379 U.S. 64, 74–75 (1964).

[***] The committee's designated topic for this hearing uses the phrase "terror content"; in this testimony, I also use the phrase "terrorist content" interchangeably.

[9] *Snyder* v. *Phelps,* 562 U.S. 443, 460–61 (2011).

THREE MAJOR CONCLUSIONS

FIRST.—Any effort to restrict on-line terror content and misinformation will be at best ineffective and at worst counterproductive in achieving the important goal of such efforts: To counter the expression's potential adverse impacts.

As the Electronic Frontier Foundation ["EFF"] recently concluded:

> "[C]ontent moderation was never meant to operate at the scale of billions of users . . . [A]s pressure from lawmakers and the public to restrict various types of speech—from terrorism to fake news—grows, companies are desperately looking for ways to moderate content at scale. They won't succeed—at least if they care about protecting on-line expression."[10]

EFF and others who have monitored content moderation efforts for years have consistently reached the same conclusion. For example, in a 2017 report, EFF stated:

> "Over the years, we've found that companies' efforts to moderate on-line content almost always result in overbroad content takedowns or account deactivations. We therefore are justifiably skeptical [about] the latest efforts . . . to combat pro-terrorism content."[11]

Concepts such as "terror content" and "misinformation" are inherently, inescapably vague and broad. Therefore, anyone who decides whether particular social media posts should be so classified, and hence restricted, inevitably exercises enormous discretion. Enforcers of any such concepts will necessarily exercise this discretion in accordance with subjective values—their own, or those of their social media employers, or those of powerful political and other established interests. As the old saying observes, "One person's terrorist is another's freedom fighter." Likewise, one person's "misinformation" or "fake news" is someone else's cherished truth.

The definitions of prohibited terrorist or extremist content that Twitter and Facebook have enforced were cited as examples of these inevitable definitional problems of vagueness and overbreadth, in an important 2018 Report by the United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye ["UN Special Rapporteur's Report"].[12]

The unavoidable indeterminacy of these elastic concepts means that their enforcement will be arbitrary at best, discriminatory at worst. Predictably, these concepts will be disproportionately enforced against marginalized, unpopular, dissident individuals and groups, those that lack political power.

I will now cite a few illustrations of the foregoing general, inevitable problems specifically concerning the particular expression at issue: Social media speech with terrorist content and misinformation.

When social media target speech with terrorist (or "extremist") content, they inevitably suppress much valuable speech, including human rights advocacy

These problems were detailed, for example, in a May 30, 2019 joint report by the Electronic Frontier Foundation, Syrian Archive, and Witness.[13] Syrian Archive engages in "documentation related to human rights violations committed by all sides involved in the conflict in Syria," and Witness promotes effective video advocacy for human rights. Noting that social media "companies have come under increasing pressure" to restrict extremist or terrorist expression, the report explained that both algorithmic and human content moderation techniques have "caught in the net" "not only content deemed extremist, but also . . . useful content like human rights documentation," with "mistakes at scale that are decimating human rights content." As the report elaborated:

> "[I]t is difficult for human reviewers—and impossible for machines—to consistently differentiate activism, counter-speech, and satire about extremism from extremism

[10] Jillian C. York and Corynne McSherry, "Content Moderation is Broken. Let Us Count the Ways," *Electronic Frontier Foundation,* Apr. 29, 2019, *https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways.*

[11] Sophia Cope, Jillian C. York and Jeremy Gillula, "Industry Efforts to Censor Pro-Terrorism Online Content Pose Risks to Free Speech," *Electronic Frontier Foundation,* July 12, 2017, *https://www.eff.org/deeplinks/2017/07/industry-efforts-censor-pro-terrorism-online-content-pose-risks-free-speech.*

[12] David Kaye, *Report of the Special Rapporteur to the Human Rights Council on online content regulation,* ¶26, U.N. Doc. A/HRC/38/35 (April 6, 2018), *https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf.*

[13] Jillian C. York, "Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content," *Electronic Frontier Foundation,* May 30, 2019, *https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content.*

itself. Blunt content moderation systems at scale inevitably make mistakes, and marginalized users are the ones who pay for those mistakes.

The report documented multiple examples of such counterproductively suppressed marginalized speakers, including groups advocating for the independence of the Chechen Republic of Iskeria, groups advocating for an independent Kurdistan, satirical commentary, and conflict documentation by journalists and human rights defenders in Syria, Yemen, and Ukraine.

In the same vein, a 2017 *New York Times* story described how YouTube's "effort to purge extremist propaganda from its platform" had led it to "inadvertently remove[] thousands of videos that could be used to document atrocities in Syria, potentially jeopardizing future war crimes prosecutions."[14] Given the breakdown of independent media since the start of the Syrian conflict, individuals and civil society organizations have subsequently used YouTube to document the war, including atrocities and human rights violations. Since some of the "disappeared" videos cannot be restored, we are losing "the history of this terrible war," and "the richest source of information about human rights violations in closed societies," according to experts whom the *Times* quoted.

This persistent problem, inherent in the irreducibly vague, overbroad concepts of terrorist or extremist content (as well as misinformation), had previously been documented in a 2017 EFF report, which cited further illustrations, including that "Facebook . . . deactivated the personal accounts of Palestinian journalists" on the ground that "they were involved in 'terrorist activity,'" and temporarily banned a journalist from the United Arab Emirates "for posting a photograph of Hezbollah leader Hassan Nasrallah with an LGBTQ pride flag overlaid on it—a clear case of parody counter-speech that Facebook's content moderators failed to grasp."[15]

Suppressing speech with terrorist content may well not promote counter-terrorism efforts and could even undermine them

The Electronic Frontier Foundation has assembled expert testimony about the strategic downsides of suppressing this expression, even beyond the adverse impact that such suppression has on free speech:

> "[T]he question is not whether terrorists are using the internet to recruit new operatives—the question is whether taking down pro-terrorism content and accounts will meaningfully contribute to the fight against global terrorism. Governments have not sufficiently demonstrated this to be the case. And some experts believe this absolutely not to be the case."[16]

Let me quote just a few of the many experts who have reached this negative conclusion, for the multiple reasons indicated.

Censorship of terrorist content doesn't promote National security

Michael German, a former FBI agent with counter-terrorism experience, who is now a fellow at the Brennan Center for Justice, stated: "Censorship has never been an effective method of achieving security, and . . . suppressing on-line content will be as unhelpful as smashing printing presses."[17]

Keeping terrorist content on-line may provide opportunities for constructive engagement that could avert terrorist acts

For example, a Kenyan government official opposed shutting down an al-Shabaab Twitter account, because "al-Shabaab needs to be engaged positively and [T]witter is the only avenue."[18]

More generally, this conclusion was reached by a United Nations report on "The Use of the Internet for Terrorist Purposes":

> "On-line discussions provide an opportunity to present opposing viewpoints or to engage in constructive debate, which may have the effect of discouraging potential supporters. Counter-narratives with a strong factual foundation may be conveyed

[14] Malachy Browne, "YouTube Removes Videos Showing Atrocities in Syria," *The New York Times,* Aug. 22, 2017, *https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.*

[15] Cope, York, and Gillula, "Industry Efforts."

[16] Cope, York, and Gillula, "Industry Efforts."

[17] Jenna McLaughlin, "The White House Asked Social Media Companies to Look for Terrorists. Here's Why They'd #Fail.," *The Intercept,* Jan. 20, 2016, *https://theintercept.com/2016/01/20/the-white-house-asked-social-media-companies-to-look-for-terrorists-heres-why-theyd-fail.*

[18] Jillian C. York and Trevor Timm, "U.S. Government Threatens Free Speech With Calls for Twitter Censorship," *Electronic Frontier Foundation,* Jan. 6, 2012, *https://www.eff.org/deeplinks/2012/01/us-government-calls-censor-twitter-threaten-free-speech.*

through on-line discussion forums, images, and videos. Successful messages may also demonstrate empathy with the underlying issues that contribute to radicalization, such as political and social conditions, and highlight alternatives to violent means of achieving the desired outcomes."[19]

A powerful specific example of the effective use of social media platforms to counter on-line terrorist propaganda comes from the U.S. Center for Strategic Counterterrorism Communications. Noting that the Center uses Facebook and YouTube for such purposes, the U.N. report cited one illustration of the touted strategy of "reducing radicalization and extremist violence by identifying in a timely manner extremist propaganda . . . on the internet and responding swiftly with targeted counter-narratives":

"For instance, in May 2012, the Center . . . responded, within 48 hours, to banner advertisements promoting extremist violence posted on various websites by al-Qaeda in the Arabian Peninsula, with counter-advertisements on the same websites featuring an altered version of that same message that was intended to convey that the victims of the terrorist organization's activities were Yemeni nationals."[20]

Keeping terrorist content on-line facilitates intelligence gathering and counter-terrorism efforts

Let me again quote the above-cited U.N. report:

"While terrorists have developed many ways to use the internet in furtherance of illicit purposes, their use of the internet also provides opportunities for the gathering of intelligence and other activities to prevent and counter acts of terrorism, as well as for the gathering of evidence for the prosecution of such acts. A significant amount of knowledge about the functioning, activities and sometimes the targets of terrorist organizations is derived from . . . internet communications. Further, increased internet use for terrorist purposes provides a corresponding increase in the availability of electronic data which may be compiled and analysed for counter-terrorism purposes. Law enforcement, intelligence and other authorities are developing increasingly sophisticated tools to proactively prevent, detect and deter terrorist activity involving use of the internet."[21]

Social media companies' restrictions on misinformation likewise have suppressed much valuable information, and also have reinforced misinformation

Efforts to clearly, consistently define and enforce prohibited "misinformation" are at least as futile as those to define prohibited "terror content." The U.N. Special Rapporteur's Report stressed the inevitable vagueness and overbreadth of restrictions on "disinformation," warning that some such "measures, particularly those that . . . restrict[] . . . news content, may threaten independent and alternative news sources or satirical content."[22] Likewise, EFF's May 1, 2019 report concluded that "when tech companies ban an entire category of content" such as "disinformation," "they have a history of overcorrecting and censoring accurate, useful speech—or, even worse, reinforcing misinformation."[23]

One especially ironic illustration of the latter problem is a 2018 incident in which Facebook's training materials used an egregious example of disinformation that was incendiary to boot. It was a photograph that depicted dozens of Buddhist monks surrounded by piles of dead, barely-clothed bodies, which was captioned as "The Bod[ies] of Muslims slaught[er]ed by Buddhist[s]." Facebook's training materials described this image as "a newsworthy exception" to Facebook's general ban on nudity (another inherently vague, overbroad concept of restricted speech) because it depicted "the victims of violence in Burma [Myanmar]." In fact, though, this image actually depicted the aftermath of an earthquake in another country years earlier.[24]

SECOND.—Social media companies' most effective strategies for countering the potential adverse impact of terrorist content and misinformation are non-censorial, including: Altering the algorithmic curation that amplifies some potentially dan-

[19] United Nations Office on Drugs and Crime, *The use of the Internet for terrorist purposes* (Vienna: United Nations, 2012), 12.

[20] UNODC, *Use of the internet,* 13.

[21] UNODC, *Use of the internet,* 12.

[22] Kaye, *Report of the Special Rapporteur,* ¶31.

[23] Jillian C. York, David Greene, and Gennie Gebhart, "Censorship Can't be the Only Answer to Disinformation Online," *Electronic Frontier Foundation* (May 1, 2019), *https://www.eff.org/deeplinks/2019/05/censorship-cant-be-only-answer-disinformation-online.*

[24] Joseph Cox, "Facebook's Own Training Materials Fell for Fake News," *Motherboard/Tech by Vice* (Sep. 5, 2018), *https://www.vice.com/en_us/article/j5ny5d/facebook-training-manuals-documents-fell-fake-news.*

gerous content; and empowering users with more individualized tools to understand and control the content they see, and to assess its credibility.

As stated by Vera Eidelman, a staff attorney with the ACLU's Speech, Privacy, and Technology Project: "Rather than focus[ing] their resources and innovation on how best to censor, [social media] companies should invest in user controls and enabling third party innovation re[garding] user controls and content moderation."[25]

Likewise, in a May 1, 2019 report critiquing social media restrictions on "disinformation," the EFF endorsed two interrelated technological approaches that social media should pursue to empower all of us to make our own voluntary, informed choices about what on-line material to view, and what not to view, consistent with our own interests and values: "addressing the algorithmic 'megaphone' at the heart of the problem and giving users control over their own feeds."[26] Although this particular report focused on disinformation, its conclusions apply fully to other potentially problematic on-line content, including terrorist material:

> "Algorithms like Facebook's Newsfeed or Twitter's timeline make decisions about which . . . content to promote and which to hide. That kind of curation can play an amplifying role for some types of incendiary content, despite the efforts of platforms like Facebook to tweak their algorithms to 'disincentivize' or 'downrank' it. Features designed to help people find content they'll like can too easily funnel them into a rabbit hole of disinformation. That's why platforms should examine the parts of their infrastructure that are acting as a megaphone for dangerous content and address the root cause of the problem rather than censoring users.
>
> "Transparency about how a platform's algorithms work, and tools to allow users to . . . create their own feeds, are critical . . . [Facebook's] [r]ecent transparency improvements in this area are encouraging, but don't go far enough . . . .
>
> "Users shouldn't be held hostage to a platform's proprietary algorithm. Instead of . . . giving users just a few opportunities to tweak it, platforms should open up their APIs **** to allow users to create their own filtering rules for their own algorithms. News outlets, educational institutions, community groups, and individuals should all be able to create their own feeds, allowing users to choose who they trust to curate their information and share their preferences with their communities."

Additional non-censorial approaches

In addition to the foregoing essential user empowerment strategies, other non-censorial approaches can also curb the potential adverse impact of terrorist content and misinformation more effectively than restricting such expression. These include:

- enforcing the many existing laws against actual terrorist and fraudulent conduct; and
- increasing media literacy, so consumers of on-line expression learn how to avoid terrorist and fraudulent communications, and how to find and generate effective "counterspeech," refuting and responding to such problematic communications, dissuading other people from accepting their messages, and perhaps even dissuading those who issued the communications (as has happened in significant instances).

PEN America, which advocates for writers and free speech, has issued two recent reports about fraudulent news and disinformation (in March 2019 and October 2017),[27] which strongly endorse media literacy skills as the ultimate antidote to the potential serious adverse impact of such expression. As its 2019 report concluded: "[T]he most effective proactive tactic against fraudulent news is a citizenry that is well-equipped to detect, and reject, fraudulent claims."[28] Correspondingly, that report concluded that "the spread of fraudulent news must not become a mandate for Government or corporate censorship."[29] Non-censorial steps that social media companies should take, according to PEN America, include "empower[ing] consumers with easy-to-use tools . . . to gauge the credibility of information disseminated through the platform."[30]

THIRD.—While social media companies have the legal right to engage in content moderation—including efforts to restrict terrorist content and misinformation—they

[25] Vera Eidelman, email message to Nadine Strossen, May 28, 2019.

[26] York, Greene, and Gebhart, "Censorship Can't be the Only Answer."

**** "API" is an abbreviation for "application program interface," which is a set of routines, protocols, and tools for building software applications.

[27] "Truth on the Ballot: Fraudulent News, the Midterm Elections, and Prospects for 2020," *PEN America* (Mar. 13, 2019), *https://pen.org/wp-content/uploads/2019/03/Truth-on-the-Ballot-report.pdf;* "Faking News: Fraudulent News and the Fight for Truth," *PEN America* (Oct. 12, 2017), *https://pen.org/wp-content/uploads/2017/11/2017-Faking-News–11.2.pdf.*

[28] PEN America, "Truth on the Ballot," 7.

[29] PEN America, "Truth on the Ballot," 48.

[30] PEN America, "Faking News," 27.

should do so in ways that are consistent with universal human rights norms, including those governing freedom of expression. At a minimum, they should follow procedural standards that promote accountability, fundamental fairness/due process, and transparency.

The Guiding Principles on Business and Human Rights, adopted by the United Nations Human Rights Council in 2011, urge companies to adhere to international human rights standards throughout their operations and wherever they operate.[31] Although these Principles are non-binding, the "overwhelming role" that the giant social media companies play "in public life globally argues strongly for their . . . implementation" of these Principles, according to the U.N. Special Rapporteur's Report.[32]

In terms of free speech norms, the U.N. Special Rapporteur's Report maintained that these companies should permit "users to develop opinions, express themselves freely and access information of all kinds in a manner consistent with human rights law."[33] The applicable human rights law substantially overlaps with core U.S. free speech principles; it requires that any speech restriction should be clearly and narrowly defined, and demonstrated to be both necessary and proportionate to avert specific, serious harm that the speech would directly cause. For speech that is feared to have a more indirect, speculative harmful potential, we should respond with non-censorial measures, as outlined above.

Concerning minimal procedural standards, a starting point is the "Santa Clara Principles On Transparency and Accountability in Content Moderation," which were adopted in 2018 by a group of civil liberties organizations and individual experts.[34] These minimum procedural principles have also been endorsed by the U.N. Special Rapporteur's Report, and at least their general "spirit" has been endorsed by many major social media companies, including all three companies represented at this hearing.[35]

The Santa Clara Principles spell out detailed steps that social media companies should take to pursue the following broader initiatives:

(1) Publishing the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines;

(2) Providing notice to each user whose content is taken down or whose account is suspended about the reason for such action; and

(3) Providing a meaningful opportunity for timely appeal of any content removal or account suspension.

MORE SPECIFIC POINTS THAT REINFORCE THESE MAJOR CONCLUSIONS

First.—Throughout history, we have seen a constant impulse to disproportionately blame expression for societal problems, which is understandable but misguided.

Correspondingly, it seems to be intuitively appealing to seek to suppress expression of ideas that one disfavors or fears to be potentially dangerous. As former Supreme Court Justice Oliver Wendell Holmes memorably put it:

> "Persecution for the expression of opinions seems to me perfectly logical. If you have no doubt of your premises or your power, and want a certain result with all your heart, you naturally express your wishes in law, and sweep away all opposition."[36]

Nonetheless, Holmes even more memorably explained why that tempting speech-blaming/speech-suppressive instinct is inconsistent with individual liberty and democracy, setting out the "emergency principle" that all modern Justices have embraced:

> "[W]e should be eternally vigilant against attempts to check the expression of opinions that we loathe and believe to be fraught with death, unless they so imminently threaten immediate interference with the lawful and pressing purposes of the law that an immediate check is required to save the country."[37]

[31] Kaye, *Report of the Special Rapporteur,* ¶6.

[32] Kaye, *Report of the Special Rapporteur,* ¶5.

[33] Kaye, *Report of the Special Rapporteur,* ¶15.

[34] "Santa Clara Principles on Transparency and Accountability in Content Moderation," *The Santa Clara Principles* (May 2018), *https://santaclaraprinciples.org.* The authors of the Principles were the ACLU of Northern California, The Center for Democracy & Technology, Electronic Frontier Foundation, New America's Open Technology Institute, Irina Raicu, Nicolas Suzor, Sarah T. Roberts, and Sarah Myers West.

[35] Gennie Gebhart, "Who Has Your Back? Censorship Edition 2019," *Electronic Frontier Foundation* (June 12, 2019), *https://www.eff.org/wp/who-has-your-back-2019.*

[36] *Abrams* v. *United States,* 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

[37] Ibid.

The general pattern of scapegoating expression for allegedly fostering societal problems is especially pronounced concerning expression that is conveyed by any new media. Each new, powerful communications medium raises concerns about its ability to transmit controversial expression to vulnerable individuals and groups, provoking fear—even panic—about potential ensuing harm. Accordingly, throughout history, censorial efforts have greeted each new communications medium: From the printing press to the internet.

Let me list some examples of media-blaming for a range of serious social problems, just within my adult lifetime:

- Sexual expression in all media, including on-line, has been blamed for undermining everything from women's equality and safety to the "traditional American family."
- So-called "hate radio" has been blamed for fomenting domestic extremism and terrorism.
- Violent videos have been blamed for instigating school shootings.
- Rap and rock lyrics have been blamed for instigating sexual assaults against women and also shootings of police officers.

Second.—With 20/20 hindsight, we have consistently come to realize that this scapegoating of expression as a purported major cause of social ills—and the associated calls for censorship as a purported solution—have multiple interrelated flaws.

- This approach wrongly regards individuals as passive automatons who lack autonomy to make our own choices about what expression to view and what not to, and to avoid being passively "brainwashed" by what we do view.
  - To be sure, as discussed above, social media companies and others should take affirmative steps to maximize individual freedom of choice in this sphere. We should ensure that all media consumers have the educational and technological resources to make truly independent, informed, voluntary decisions about our communications. In the social media context, this means not directing users to increasingly extreme content, unbeknownst to them. It does mean developing and deploying technology that will empower each user to make maximally individualized choices about what content to view and to communicate, and what to avoid or block.
- Scapegoating expression also diverts attention and resources from the real problems: Underlying attitudes and actual conduct. Suppressing expression is a superficial, cheap "quick fix" for complex, deep-rooted problems, which actually fixes nothing. To the contrary, pushing the feared ideas underground may well make it harder to counter those ideas, and also harder to prevent those who hold them from engaging in harmful conduct.
- This censorial approach may not only make it harder to recruit advocates of the feared ideas/actions away from their ideologies, but it may well also increase attention, sympathy, and support for such ideologies among members of the broader public. This pattern is so common that several terms have been coined to describe it, including "the forbidden fruits effect," "the boomerang effect," and "the Streisand effect" (the latter term was coined when Barbra Streisand sought to block on-line photographs of her Malibu home, thus increasing exponentially the viewing of such photographs). We should focus instead on persuading people to reject dangerous ideas, and preventing people from engaging in harmful conduct.

Third.—Social media companies are not constrained by the First Amendment's Free Speech Clause, which limits only Government actors, not private-sector actors. To the contrary, social media companies have their own First Amendment rights, including the right to decide which speakers and expression to permit—or not to permit—on their platforms. However, these platforms should provide the same free speech opportunities that Government is required to provide, consistent with both compelling policy concerns and global human rights norms applicable to business.

As I noted at the outset of this testimony, social media companies have their own First Amendment rights to adopt and enforce content moderation policies they choose, and any Government regulation of such policies—whether prescribing or proscribing—any such policies—would raise serious concerns about abridging the companies' freedom of speech. However, it is eminently appropriate for Congress and other Government actors, as well as civil society groups and users, to encourage these companies to implement content moderation policies, and to take other actions, that promote their users' free speech, as well as promoting other essential concerns, including National security and democracy.

As a practical matter, social media platforms now constitute the most important forums for exchanging information and ideas, including between "We the People" (to quote the Constitution's opening words) and the political candidates and officials who are accountable to us. In a 2017 Supreme Court decision that unanimously struck down a State law that restricted access to social media by convicted sex offenders who had served their prison terms, Justice Anthony Kennedy's majority opinion stressed the social media's stature as the preeminent platform for expression. If we do not have equal, open access to these forums, to convey and receive communications, then for all practical purposes, our freedom of speech—and, accordingly, our equal stature as sovereign citizens—is curtailed. As Justice Kennedy declared:

> "A fundamental principle of the First Amendment is that all persons have access to places where they can speak and listen, and then, after reflection, speak and listen once more. The Court has sought to protect the right to speak in this spatial context. A basic rule, for example, is that a street or a park is a quintessential forum for the exercise of First Amendment rights . . . While in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyber space . . . and social media in particular."[38]

Moreover, as discussed above, social media companies should adhere to the U.N. Human Rights Council's Guiding Principles on Business and Human Rights, which include respect for free speech. As the U.N. Special Rapporteur urged, social media companies should engage in content moderation that permits "users to express themselves freely and access information of all kinds in a manner consistent with human rights law."[39]

Fourth.—A core free speech principle that social media companies should honor, consistent with both U.S. and human rights law, is "content neutrality" or "viewpoint neutrality": That speech should not be restricted solely due its disfavored content—i.e., its viewpoint, message, or ideas.

No matter how loathed or feared such content may be, by no matter how many of us, we must respond to it with non-censorial counter measures, including education and persuasion. Measures that discriminate against speech based solely on its disfavored content or viewpoint are almost automatically un-Constitutional. The Supreme Court has hailed content neutrality as "the bedrock principle" undergirding Constitutional freedom of speech.[40]

This fundamental free speech principle is reflected in international human rights norms. Accordingly, the U.N. Special Rapporteur's Report expressly urges social media companies to enforce their content moderation policies consistent with a "non-discrimination" standard, rather than through "heavy-handed viewpoint-based regulation."[41]

Fifth.—Social media companies should additionally honor the complementary "emergency" principle, which is also integral to both U.S. and human rights law.

When we move beyond the content of speech and consider its context, speech may be restricted if it satisfies the emergency test: When, under all the facts and circumstances, the speech directly causes certain specific, imminent, serious harm, which cannot effectively be countered through non-censorial measures.

This key principle is also reflected in the global human rights requirements of "necessity" and "proportionality." As the U.N. Special Rapporteur's Report explained, proponents of any speech restriction "must demonstrate that the restriction imposes the least burden on the exercise of" free speech, "and actually protects, or is likely to protect, the legitimate . . . interest at issue." Proponents of the restriction "may not merely assert necessity but must demonstrate it, in the restriction of specific expression." Moreover, social media "[c]ompanies should . . . demonstrate the necessity and proportionality of any content actions (such as removals or account suspensions)."[42]

Applying these standards to terror content and misinformation, social media companies should not restrict such expression unless they could demonstrate that the restriction was "necessary" and "proportional" for averting the potential harms of such expression. This showing would be hard to make concerning either terror content or misinformation, in light of the evidence discussed above, which demonstrated the inherent overbreadth of such speech restrictions, and called into question whether they are even effective in averting the potential harms, let alone necessary.

Sixth.—The Supreme Court has designated several narrowly-defined categories of speech that may be restricted consistent with the content neutrality and emergency principles, including two that are pertinent to the two types of speech at issue in

[38] Packingham, 137 S. Ct. at 1735.

[39] Kaye, *Report of the Special Rapporteur,* ¶39.

[40] *Texas* v. *Johnson,* 491 U.S. 397, 414 (1988).

[41] Kaye, *Report of the Special Rapporteur,* ¶¶48, 66.

[42] Kaye, *Report of the Special Rapporteur,* ¶¶7, 28, 66

these hearings: Speech with terrorist content and misinformation. It would be appropriate for social media companies to restrict these narrowly-defined subcategories of speech with terrorist content and misinformation: Speech that satisfies the standards for punishable incitement, fraud, or defamation.

- The Court has barred Government from restricting speech that contains terrorist content or speech that is feared to potentially contribute to terrorism unless, in context, it satisfies the following, appropriately strict, standards: It intentionally incites imminent violent or criminal conduct, and it is actually likely to do so imminently. Accordingly, the Court has struck down even restrictions on explicit advocacy of violent or criminal conduct, including terrorism, when it falls short of the foregoing strict intentional incitement standard.[43]
- The Court has barred Government from punishing many kinds of "misinformation" and even outright lies, except in certain situations when intentional falsehoods directly cause certain specific imminent serious harms, including by defrauding an individual who has reasonably relied on the falsehood in a way that causes demonstrable tangible injury; or by defaming an individual about a matter of private concern in a way that injures her reputation and causes demonstrable tangible injury. When the defamatory falsehood pertains to a public official or public figure, it may not be punished unless the complainant can establish, by "clear and convincing evidence," that the speaker knowingly or recklessly lied.[44]

Seventh.—Speech that does not satisfy the emergency test may still cause serious harm; that is true for speech with terrorist content and misinformation. However, the modern Court has consistently enforced the content neutrality and emergency principles because it is even more harmful to grant enforcing authorities latitude to punish speech that does not satisfy the emergency test. This is true regardless of who the enforcing authorities are, including social media companies.

As Supreme Court Justice Oliver Wendell Holmes famously recognized, "Every idea is an incitement."[45] He did not mean that Government may therefore suppress every idea, but rather the opposite: If every idea that could potentially incite harmful conduct or consequences could be suppressed, all ideas could be suppressed. Accordingly, to shield freedom to express ideas—potentially inciting and potentially dangerous as they are—we should confine censorial power only to ideas and expression that satisfy the emergency test: Directly causing specific, imminent, serious harm.

If censorial power could be exercised under a looser, broader standard, the resulting discretion would inevitably lead to suppressing valuable speech, and would disproportionately target speech by relatively disempowered, marginalized individuals and groups, including those who challenge the status quo.

- For example, before the Supreme Court adopted the strict "intentional incitement" test, Government regularly enforced legal restrictions on "terrorist" and other feared expression against politically unpopular, relatively powerless speakers, including abolitionists, socialists, women's suffragists, pacifists, anti-war and anti-draft demonstrators, and civil rights advocates.
- Likewise, before the Supreme Court adopted strict standards limiting punishable defamation, National media outlets and civil rights leaders and organizations were regularly targeted with defamation lawsuits that (absent the Court's invalidation) would have led to speech-suppressive damage awards, preventing information and advocacy about the civil rights movement from reaching the critically important Nation-wide audience.
- When expression may be restricted short of the emergency standard, the restrictions are often counterproductive: Suppressing expression that would actually promote the countervailing goals at issue.
- As detailed above, these general, inevitable problems have—predictably—specifically afflicted social media companies' enforcement of their standards that restrict terrorist content and misinformation.

CONCLUSION

In closing, I would like to invoke an apt observation by the famed twentieth-century journalist H.L. Mencken: "For every complex problem, there is a solution that is clear, simple—and wrong."

How to effectively counter the serious potential adverse impact of terror content and misinformation is certainly a complex problem. While restricting such expres-

[43] *Brandenburg* v. *Ohio,* 395 U.S. 444 (1969).

[44] *New York Times Co.* v. *Sullivan,* 376 U.S. 254 (1964).

[45] *Gitlow* v. *New York,* 268 U.S. 652, 673 (1925) (Holmes, J., dissenting).

sion might appear to be a clear, simple solution, it is in fact neither—and, moreover, it is wrong. We must focus on the non-censorial strategies I have discussed, including user empowering education and technology. Although these approaches are also not simple, they are far more promising than censorship.

Chairman THOMPSON. I thank all the witnesses for their testimony.

I remind each Member that he or she will have 5 minutes to question the panel.

I will now recognize myself for questions.

Misinformation is some of this committee's challenges as it relates to this hearing, as well as the terrorist content. Let's take, for instance, the recent doctored video of Speaker Nancy Pelosi that made her appear to be drunk or slurring her words. Facebook and Twitter left up the video, but YouTube took it down. Everybody agreed that something was wrong with it. Facebook, again, took a different approach.

So I want Ms. Bickert and Mr. Pickles to explain how you decided the process for leaving this video up on Facebook and Twitter.

Then, Mr. Slater, I want you to explain to me why YouTube decided to take it down.

Ms. Bickert.

Ms. BICKERT. Thank you, Mr. Chairman.

Let me first say, misinformation is a top concern for us, especially as we are getting ready for the 2020 elections. We know this is something that we have to get right. We are especially focused on what we should be doing with increasingly sophisticated manipulated media.

So let me first speak to our general approach with misinformation, which is: We remove content when it violates our community standards. Beyond that, if we see somebody that is sharing misinformation, we want to make sure that we are reducing the distribution and also providing accurate information from independent fact-checking organizations so that people can put in context what they see.

To do that, we work with 45 independent fact-checking organizations from around the world, each of which is certified by Poynter as being independent and meeting certain principles. As soon as we find something that those fact-checking organizations rate "false" on our platform, we dramatically reduce the distribution, and we put next to it related articles so that anybody who shares that gets a warning that this has been rated "false." Anybody who did share it before we got the fact-checkers's rating gets a notification that the content has now been rated "false" by a fact-checker, and we are putting next to it those related articles from the fact-checking organizations.

Chairman THOMPSON. I understand. How long did it take you to do that for the Pelosi video?

Ms. BICKERT. The Pelosi video was uploaded to Facebook on Wednesday, May 22, around late morning. On Thursday around 6:30 p.m., a fact-checking organization rated it as "false," and we immediately downranked it and put information next to it.

That is something where we think we need to get faster. We need to make sure that we are getting this information to people as soon as we can. It is also a reason that at 6:30 p.m.——

Chairman THOMPSON. So it took you about a day-and-a-half.

Ms. BICKERT. Yes, it did, Mr. Chairman.

Chairman THOMPSON. Thank you.

Mr. Pickles.

Mr. PICKLES. So, as Monika said, the process for us is we review this against our rules; any content that breaks our rules we will remove. We are also very aware that people use manipulated tactics to spread this content—fake accounts, automation. So we will take action on the distribution as well as the content.

This is a policy area we are looking at right now not just in the case of where videos might be manipulated but also where the videos are fabricated, where the whole process of creating media may be artificial.

We think that the best way to approach this is with a policy and a product approach that covers in some cases removing——

Chairman THOMPSON. I understand, but just get to why you left it up.

Mr. PICKLES. So, at present, the video doesn't break our rules, and then the account posting it doesn't break our rules. But it is absolutely a policy area we are looking at right now, about whether our rules and our products are the correct framework for dealing with this challenge, which——

Chairman THOMPSON. So if it is false or misinformation, that doesn't break your rules.

Mr. PICKLES. Not at present, no.

Chairman THOMPSON. Thank you.

Mr. Slater.

Mr. SLATER. So, on YouTube, we have tough community guidelines that lay out the rules of the road, what is inbounds to be up on the platform and what is out. Violative content, when it is identified to us via machines or users, we will review and remove.

In this case, the video in question violated our policies around deceptive practices, and we removed it.

Chairman THOMPSON. So, again, our committee is tasked with looking at misinformation and some other things. We are not trying to regulate companies, but terrorist content can also be a manipulated document.

So, Ms. Strossen, talk to us about your position with that.

Ms. STROSSEN. The difficulty in—the inherent subjectivity of these concepts, Chairman Thompson, is illustrated by the fact that we have three companies that have subscribed to essentially the same general commitments and yet are interpreting the details very differently with respect to specific content. We see that over and over again.

Ultimately, the only protection that we are going to have in this society against disinformation is training and education starting at the earliest levels of a child's education in media literacy.

Because Congress could never protect against misinformation in traditional media—right?—unless it meets the very strict standards of defamation that is punishable and fraud that is punishable, content, including the Pelosi video, is completely Constitutionally protected in other media.

Chairman THOMPSON. Thank you.

I yield to the Ranking Member for his questions.

Mr. ROGERS. Thank you, Mr. Chairman.

Mr. Slater, the video I referenced in my comments with Ms. Gennai and your employee, would you like to take this opportunity—have you seen it?

Mr. SLATER. Congressman, I have not seen the full video, but I am broadly aware of what you are talking about, yes.

Mr. ROGERS. OK. Would you like to take an opportunity to respond to the comments that I offered about what was said?

Mr. SLATER. Could you be specific, Congressman? What would you like me to respond to?

Mr. ROGERS. When she basically, for example, said that we can't let Google be broken up because these smaller companies won't have the same resources we have to stop Trump from getting reelected.

Mr. SLATER. Thank you for the clarification.

So let me be clear, this employee was recorded without her consent. I believe these statements were taken out of context.

But stepping back to our policies, how we address the issue you are talking about, no employee, whether in the lower ranks, up to senior executives, has the ability to manipulate our search results or our products or our services based on their political ideology.

We design and develop our products for everyone. We mean everyone. We do that to provide relevant results, authoritative results. We are in the trust business. We have a long-term incentive to get that right.

We do that in a transparent fashion. You can read more on our How Search Works site. We have search rater guidelines that are public on the web that describe how we look at rating. We have robust systems and checks and balances in place to make sure those are rigorously adhered to as we set up our systems.

Mr. ROGERS. OK. I recognize that she was being videotaped without her knowledge, but the statements that I quoted from were full, complete statements that were not edited.

So it is concerning when you see somebody who is an executive at Google—and there were more than one in that video, by the way—making statements that indicate that it is management's policy within Google to try to manipulate information to cause one or another candidate for President of the United States—or, for that matter, any other office—to be successful or not be successful.

So that is what gave rise to my concern. Do we have reason to be concerned that Google has a pervasive nature in the company to try to push one political party over another in the way it conducts its business?

Mr. SLATER. Congressman, I appreciate the concern, but let me be clear again: In terms of what our policy is, from the highest levels on down, and what our practices and structures and checks and balances are about, we do not allow anyone—lower level, higher level—to manipulate our products in that way.

Mr. ROGERS. OK. I hope it is not the culture at any of your platforms, because you are very powerful in our country.

Ms. Strossen, you raised concerns in your testimony that, while social media companies legally can decide what content to allow on their platforms, such censorship stifles free speech and results in biased coverage.

What are your recommendations to these companies regarding content moderation without censorship?

Ms. STROSSEN. Thank you so much, Ranking Member Rogers.

I would, first of all, endorse at least the transparency that both you and Chairman Thompson stressed in your opening remarks and, in addition, other process-related guarantees, such as due process, the right to appeal, and a clear statement of standards.

I would also recommend standards that respect the free-speech guarantees not only in the United States Constitution but of international human rights that the United Nations Human Rights Council has recommended in a nonbinding way that powerful companies adopt. That would mean that content could not be suppressed unless it posed an emergency, that it directly caused certain specific, serious, imminent harm that can't be prevented other than through suppression.

Short of that, as you indicated, for example, Ranking Member Rogers, politically controversial, even repugnant, speech should be protected. We may very much disagree with the message, but the most effective as well as principled way to oppose it is through more speech.

I would certainly recommend, as I did in my written testimony, that these companies adopt user-empowering technology that would allow us users to make truly informed, voluntary decisions about what we see and what we don't see, and not manipulate us, as has been reported many times, into increasing rabbit holes and echo chambers, but give us the opportunity to make our own choices and to choose our own communities.

Mr. ROGERS. Thank you.

I yield back.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentlelady from Texas, Ms. Jackson Lee, for 5 minutes.

Ms. JACKSON LEE. I thank the Chair, and I thank the Ranking Member, the committee Members for this hearing.

Let me indicate that there is known to the public the fourth estate, and I might say that we have a fifth estate, which is all of you and others that represent the social media empire.

I believe it is important that we work together to find the right pathway for how America will be a leader in how we balance the responsibilities and rights of such a giant entity and the rights and privileges of the American people and the sanctity and security of the American people.

Social media statistics from 2019 show that there are 3.2 billion social media users world-wide, and this number is only growing. That equates to about 42 percent of the current world population. That is enormous. Certainly, I know the numbers are just as daunting in the United States.

So let me ask a few questions, and I would appreciate brevity because of the necessity to try to get as much in as possible.

On March 15, 2019, worshipers were slaughtered in the midst of their prayers in Christchurch, New Zealand. The gunman live-streamed the first attack on Facebook Live.

So my question to you, Ms. Bickert, is, can you today assure the committee that there will never be another attack of this nature that will be streamed as it is happening over Facebook Live?

You mentioned 30,000 and 300, and so I hope they may contribute to your answer. But I yield to you for your answer.

Ms. BICKERT. Congresswoman, thank you.

The video was appalling. The attack, of course, is an unspeakable tragedy. We want to make sure we are doing everything to make sure it doesn't happen again and it is not live-streamed again.

One of the things we have done is we have changed access to Facebook Live so that people who have a serious content policy violation are restricted from using it. So the person who live-streamed the New Zealand attack will not——

Ms. JACKSON LEE. What is the likelihood of you being able to commit that that would not happen again, in terms of the new structures that you have put in place?

Ms. BICKERT. Well, the technology we are working to develop—the technology is not perfect. So artificial intelligence is a key component of us recognizing videos before they are reported to us. This video was not—about fewer than 200 people saw it while it was live on Facebook. Nobody reported it.

Ms. JACKSON LEE. So can you give me—my time is short. Do you have a percentage? Fifty percent? Sixty percent?

Ms. BICKERT. With the technology, I can't give a percentage. I can say that we are working with governments and others to try to improve that technology so that we will be able to better recognize——

Ms. JACKSON LEE. Mr. Pickles and Mr. Slater, if you would, Ms. Bickert did raise the question about artificial intelligence. So, if you would respond as to the utilization of AI and individuals, as briefly as possible, please.

Mr. PICKLES. So one of the challenges Twitter has is that there is not a lot of content—280 characters, a maximum of 2-minutes-20 video. So one of the challenges in Christchurch was, we didn't see the same video uploaded. We saw different snippets that took different lengths.

So we are investing in technology to make sure that people can't re-upload content once it has been removed previously. We are also making changes to make sure that, for example, where people manipulate media we can move quicker. But this is an——

Ms. JACKSON LEE. So you are using human subjects and AI?

Mr. PICKLES. It is machine learning and humans, yes.

Ms. JACKSON LEE. All right.

Mr. Slater.

Mr. SLATER. Thank you, Congresswoman.

We use, similarly, a combination of machine learning and people to review. Speaking overall, in the first quarter of 2019, 75 percent of the 8 million videos we removed, they were first flagged by a machine, and the majority were removed before a single view.

When it comes to violent extremism, it is even stronger. So over 90 percent of the violent extremist videos that were uploaded and removed in the past 6 months were removed before a single human flag, and 88 percent with less than 10 views. That is——

Ms. JACKSON LEE. Thank you.

Let me ask a question about deepfakes—because my time is going—for each of you, in the 2020 election, what you will do to recognize the fact that deepfakes can be a distortion of an election that is really the premise of our democracy. Can you quickly answer that question?

At the same time, I just want to make mention of the fact that free speech does not allow incitement, fighting words, threats, and otherwise.

Could you just answer that, please——

Ms. BICKERT. Yes, Congresswoman.

Ms. JACKSON LEE [continuing]. The deepfakes? As briefly as you can.

Ms. BICKERT. Absolutely.

We are working with experts outside the company and others to make sure that we understand how deepfakes can be used and come up with a comprehensive policy to address them.

In the mean time, we are focused on removing fake accounts, which are disproportionately responsible for this sort of content, and also making sure that we are improving the speed at which we counter misinformation with actual factual articles and reduce the distribution.

Ms. JACKSON LEE. Mr. Pickles.

Mr. PICKLES. So, similarly, we are working on a product and policy solution. But one of the things that we already have in place is, if anyone presents any misinformation about how to vote that lends to voter suppression, we will remove that now. That policy has been in place for some time.

Ms. JACKSON LEE. Mr. Slater.

Mr. SLATER. Similarly, we are investing significantly in working with researchers and others to build capacities in this space. We have an intel desk that scans the horizon for new threats and constantly is looking at this sort of issue.

Ms. JACKSON LEE. Thank you for your courtesy. I yield back, Mr. Chairman.

Chairman THOMPSON. The Chair recognizes the gentleman from North Carolina, Mr. Walker, for 5 minutes.

Mr. WALKER. Thank you, Mr. Chairman.

While we were sitting here today, I just looked up on the internet, put in “Facebook apologizes,” “Google apologizes,” “Twitter apologizes,” and there were more pages than I could count, going through those apologies there.

I listened closely to the words or how you framed it, both Mr. Pickles and Mr. Slater, when you talked about—one of you used “hateful content”—Mr. Pickles. Mr. Slater, you used the expression “hate speech.” You listed several different people that were protected. What I did not hear you say in that group of people that you listed were those that were wanting to express their faith.

In April—one of the larger apologies I think that you guys have made—in April, Kelsey Harkness brought us to the attention of Abby Johnson’s life story in a movie called “Unplanned.” That movie has gone on to make $20 million at the box office. But Google listed that as propaganda.

My question for you today: Was that a machine that listed that, or was that an individual?

Mr. SLATER. Congressman, I am not familiar with the specific video in question. I would have to—I would be happy to go back——

Mr. WALKER. This isn't a video. It is a movie. It was one of the larger stories in April this year, a major motion picture. You are not familiar with that? It didn't come across your radar?

Mr. SLATER. No, sir, I am not familiar with that specific video.

Mr. WALKER. OK. All right.

When we talk about the difference between hateful content and hate speech, I know, Mr. Pickles, in June, just earlier this year, Marco Rubio brought the attention that Twitter was banning any kind of language that was maybe offensive to China. You later came back and apologized.

The question for you is: How does Twitter use their discretion to block information without discriminating against different individuals or groups?

Mr. PICKLES. Well, first, as you say, our rules identify hateful conduct. So we focus on behavior first. So how do two accounts interact? We look at that before we look at the speech that they are sharing.

So there are offensive views on Twitter, and there are views that people will disagree with strongly on Twitter. The difference between that and targeting somebody else is a difference between content and conduct. So our rules don't have ideology in them. They are enforced without ideology and impartially. Where we do make mistakes, I think it is important for us to recognize.

I know one of the challenges we have is that, where we remove someone from Twitter and they come back for a different purpose, our technology will recognize that person trying to come back on Twitter, and we don't want people to come back to the platform that we have removed. Sometimes that does catch people who are having a different purpose.

So there is both a value to technology, but we should recognize where we have made a mistake.

Mr. WALKER. Mr. Slater, how does Google audit their content moderation policies to ensure that they are being followed and that they are not being driven by bias?

Mr. SLATER. Thank you, Congressman, for that question.

Speaking broadly, we have a robust system of both the development and the enforcement of our policies. We are constantly reviewing and analyzing the policies themselves to understand whether they are fit for purpose, whether they are drawing the right lines.

Our reviewers go through extensive training to make sure we have a consistent approach. We draw those reviewers from around the country, around the world, and, again, train them very deeply and are constantly reviewing——

Mr. WALKER. Yes, and I appreciate it. I need to keep moving.

What type of training, if any, do you provide for your human content moderators regarding subjectivity and avoiding bias, Mr. Slater?

Mr. SLATER. Again, we provide robust training to make sure that we are applying a consistent rule.

Mr. WALKER. "Robust training," what does that mean? What is robust training?

Mr. SLATER. So, when reviewers are brought on board, before they are allowed to review, we provide them with a set of educational materials and detailed steps. In addition, they are reviewed by managers and others to make sure that they can correct mistakes, then learn from those mistakes, and so on.

Mr. WALKER. All right.

Ms. Bickert, do you think that AI will ever get to the point where you can rely solely on it to moderate content, or do you think human moderation will always play a role?

Ms. BICKERT. Thank you for the question, Congressman.

At least for the near future, human moderation is very important to this. Technology is good at some things. It is good at, for instance, matching known images of terror propaganda or child sexual abuse. It is not as good at making the contextual calls around something like hate speech or bullying.

Mr. WALKER. Uh-huh.

Final couple questions as I wind down my time.

Mr. Pickles, do you have any idea how many times Twitter apologizes per month for missing it on content?

Mr. PICKLES. Well, I know that we take action on appeals regularly. Every decision we have made——

Mr. WALKER. Do you have a number on that?

Mr. PICKLES. I don't have a number off-hand, but I can happily follow up.

Mr. WALKER. Mr. Slater, do you have any idea how many times Google apologizes for mismanaging the content per month?

Mr. SLATER. Congressman, similarly, we have an appeals process, so there are times where we don't get it right——

Mr. WALKER. Do you have a number?

Mr. SLATER. I do not today, but I would be happy to come back to you.

Mr. WALKER. Yes, I think you guys have apologized more than Kanye West has to Taylor Swift at some point.

With that, I yield back.

Chairman THOMPSON. The Chair recognizes the gentlelady from Illinois, Ms. Underwood, for 5 minutes.

Ms. UNDERWOOD. Thank you, Mr. Chairman.

In March, 2 weeks after the Christchurch terror attack, Facebook announced it would start directing users searching for white supremacist terms to Life After Hate, an organization that works to rehabilitate extremists.

Life After Hate is based in Chicago, so I met with them last month when I was at home in Illinois. They told me since Facebook's announcement they have seen, "a large bump in activity that hasn't slowed down."

Facebook and Instagram have 3 billion users combined. Life After Hate is a tiny organization whose Federal funding was pulled by this administration. They do great work and simply don't have the resources to handle every single neo-Nazi on the internet on their own.

Ms. Bickert, has Facebook considered providing continuous funding to Life After Hate for the duration of this partnership?

Ms. BICKERT. Congresswoman, thank you for that question.

Life After Hate is doing great work with us. For those who don't know, basically, we are redirecting people who are searching for these terms to this content. We do this in some other areas as well, like, for instance, with self-harm support groups.

We do see that sometimes they are under-resourced. So this is something that we can come back to you on, but we are definitely committed to making sure this works.

Ms. UNDERWOOD. OK. So right now there is no long-term funding commitment, but you will consider it.

Ms. BICKERT. I am not sure what the details are, but I will follow up with you on them.

Ms. UNDERWOOD. OK. So Facebook has made Life After Hate a significant component of its strategy against on-line extremism, and so we really would appreciate that follow-up with exact information.

Ms. BICKERT. I would be happy to.

Ms. UNDERWOOD. Mr. Slater, over the years, YouTube has put forward various policy changes in an attempt to limit how easily dangerous conspiracy-theory videos spread. For example, YouTube announced over a year ago that it would display, "information cues" in the form of links to Wikipedia next to the conspiracy videos.

Mr. Slater, in the 15 months since this policy was announced, what percentage of users who view videos with information cues actually click on the link for more information?

Mr. SLATER. Thank you for the question, and I think this is a very important issue. We do both display these sort of contextual cues to Wikipedia and Encyclopedia Britannica as well as take a number of other steps.

Ms. UNDERWOOD. Right.

Mr. SLATER. I don't have a specific percentage on how many have clicked through but would be happy to come back to you.

Ms. UNDERWOOD. OK. If you can follow up in writing, that would be appreciated.

Most Wikipedia articles can be edited by anyone on the internet. We have all seen some with questionable content. Does YouTube vet the Wikipedia articles that it links to on information cues to ensure their accuracy? Or do you all work with Wikipedia to ensure that the articles are locked against malicious edits?

Mr. SLATER. We work to raise up authoritative information and ensure that what we are displaying is trustworthy and correct any mistakes that we may make.

Ms. UNDERWOOD. So you all have corrected the YouTube—I'm sorry, the Wikipedia pages if it is incorrect?

Mr. SLATER. No. I am sorry. Before we display such things, we look to ensure that we have a robust process to make sure that we are displaying accurate information.

Ms. UNDERWOOD. The question is about what you are linking to.

Mr. SLATER. Yes.

Ms. UNDERWOOD. OK. So can you just follow up with us in writing on that one?

Mr. SLATER. Yes.

Ms. UNDERWOOD. Great.

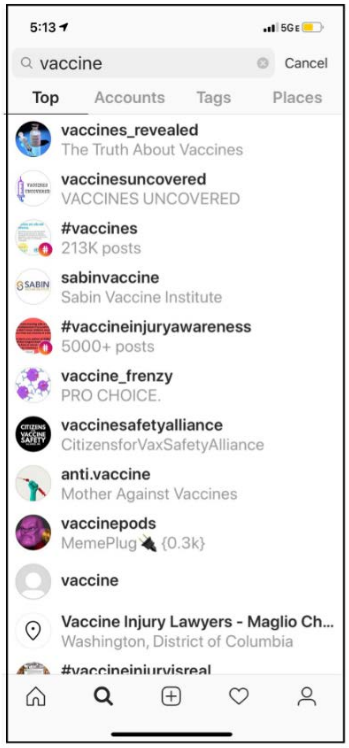
Ms. Bickert, Facebook has displayed links to additional reporting next to content that contains disinformation. What percentage of users click through to read that additional reporting?

Ms. BICKERT. I don't have that percentage for you, Congresswoman. I am sorry about that. But I will follow up in writing quickly.

Ms. UNDERWOOD. Thank you.

Mr. Chairman, at this point, I would like to ask the clerk to display the two screenshots my staff provided earlier on the TV screens.

[The information follows:]



vaccines_revealed
The Truth About Vaccines

vaccinesuncovered
VACCINES UNCOVERED

#vaccines
213K posts

sabinvaccine
Sabin Vaccine Institute

#vaccineinjuryawareness
5000+ posts

vaccine_frenzy
PRO CHOICE.

vaccinesafetyalliance
CitizensforVaxSafetyAlliance

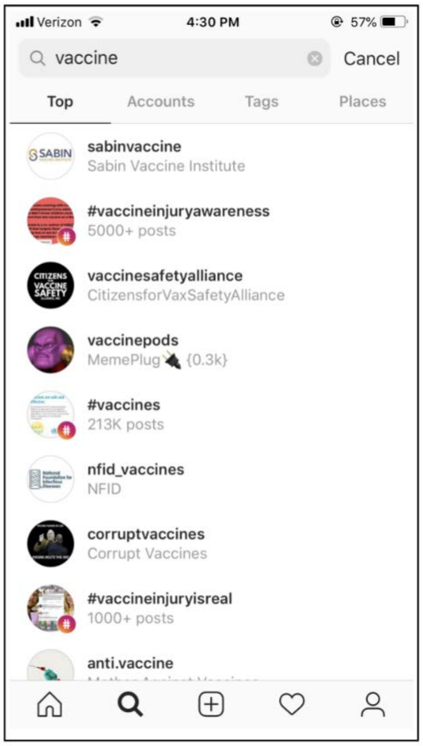
anti.vaccine
Mother Against Vaccines

vaccinepods
MemePlug 🔌 {0.3k}

vaccine

Vaccine Injury Lawyers - Maglio Ch...
Washington, District of Columbia

#vaccineinjuryisreal



sabinvaccine
Sabin Vaccine Institute

#vaccineinjuryawareness
5000+ posts

vaccinesafetyalliance
CitizensforVaxSafetyAlliance

vaccinepods
MemePlug 🔌 {0.3k}

#vaccines
213K posts

nfid_vaccines
NFID

corruptvaccines
Corrupt Vaccines

#vaccineinjuryisreal
1000+ posts

anti.vaccine

Ms. UNDERWOOD. Last month, Instagram announced that it would hide search results for hashtags that displayed vaccine disinformation. So, yesterday, I did a simple search for "vaccine" on Instagram from two different accounts. These are the top results. As you can see, the majority of these responses display anti-vax hashtags and popular accounts with titles like "CorruptVaccines," "VaccinesUncovered," and "Vaccine Injury Awareness."

These are not niche terms. This content is not hard to find, and vaccine disinformation is not a new issue.

Ms. Bickert, clearly, Instagram's efforts here have some gaps. Anti-vax content is a deadly threat to public health. What additional steps can Instagram commit to taking to ensure that this content is not promoted?

Ms. BICKERT. Congresswoman, thank you for that question. Vaccine hoaxes and misinformation are really top of mind for us. We have launched some recent measures, but I want to tell you how we are working to get better on those.

One thing we are doing is, when accounts are sharing misinformation, we are trying to downrank them and downrank them in the search results as well. That is something that is on-going. It requires some manual review for us to make sure that we are doing that right. But we are getting better at that.

Another thing is actually surfacing educational content, and we are working with major health organizations to provide that. So, when people go searching for this, at the top of the search results they will see that informational content. We are working with those health organizations right now, and we should have that content up and running soon. I can follow up with you with the details on that.

Ms. UNDERWOOD. Please. While this is a new initiative for your organization, it is critically important that that information is shared with users at the time that they search for it, which we know is on-going.

Look, everyone in this room appreciates that on-line extremism and disinformation are extremely difficult problems that require broad, coordinated solutions. But these aren't new challenges, and failing to respond seriously to them is dangerous. The research is clear: Social media helps extremists find each other, helps make their opinions more extreme, and helps them hurt our communities.

My constituents and I want strong policies from your companies that keep us safe. While I truly believe that your current policies are well-intentioned, there is a lot more that needs to be done. Frankly, some of it should have been done already. I am looking forward to working with your companies and my colleagues in Congress on broad, real solutions.

Thank you, and I yield back.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentleman from New York, Mr. Katko, for 5 minutes.

Mr. KATKO. Thank you, Mr. Chairman.

Thank you all for being here today.

It is obvious from this conversation that this is a very difficult area to maneuver in.

Ms. Strossen, I understand your concerns about First Amendment infringement, but I also understand and I applaud the companies' desire to try and find that delicate balance. Quite frankly, since you are not a Government entity, you have more flexibility in how you do that, and it is kind of up to you, as the stewards of that flexibility, to do the best job you possibly can. So I am going to get back to you in a minute with a couple of questions.

But I just want to follow up with Mr. Slater to make sure I am perfectly clear with what you are saying here. I am well aware from your testimony previously of what the policies and practices are at Google. But that video that Mr. Rogers did reference did show that people were—looked like they were talking about a very serious political bias and their intent to implement that bias in their job. Whether or not that happened, I don't know.

I am not asking about the policies and practices. I am asking you if you personally have ever been made aware of anyone that has done that, used political bias at Google to alter content, or whether they—first of all, have you ever heard that within Google? I know what your policy and practices are, so I don't want a long answer. I just want to know if you have heard that.

Mr. SLATER. Congressman, I am not aware of any situation. Our robust checks and balances and processes would prevent that.

Mr. KATKO. OK. So you personally have not ever heard of that, ever, since your time at Google.

Mr. SLATER. Correct.

Mr. KATKO. OK. The allegation that Congressman Walker referenced about the abortion movie, you have heard nothing about people limiting contact with respect to that, as well—context—excuse me—content?

Mr. SLATER. Congressman, I am not familiar with that video, no.

Mr. KATKO. OK. All right. You have never heard anybody limiting content in that regard for any sort of issue-oriented things?

Mr. SLATER. Again, we would remove content where it violates our policies, but not—but our policies with regard to ranking——

Mr. KATKO. I am aware of your policy and practices. I'm just saying, have you ever heard that yourself? There is a difference. You understand the difference? It is not what your policy and practices are; it is what you are personally aware of.

Mr. SLATER. Yes, Congressman, I believe I understand. I am not aware of any situation like that.

Mr. KATKO. OK. Thank you.

Now, I want to talk to Mr. Slater—all of you here today. This internet forum, G-I-F-C-T, GIFCT—which is the lamest acronym ever, by the way—Global Internet Forum to Counter Terrorism.

Can someone just—Mr. Pickles, perhaps, could you just give me a little detail of what exactly the goal is of this forum?

Mr. PICKLES. Sure. Equally, as Facebook and Google have both chaired the organization, happy for them to add.

I think the critical thing is, GIFCT is about bringing together 4 companies who have expertise and investment on countering terrorism but recognizing that the challenge is far bigger.

So the 3 strands. Support small companies. As we remove content, it goes across the internet, and we need to help those small companies. Fund research, so we can better understand, so we have a research network. Then, finally, sharing technical tools. So you have heard people reference these digital fingerprints to make sure that, whether it is a fingerprint or—in Twitter's case, we share the URL. So, if we take down an account for spreading a terrorist manual and we see it is linked to a company, we will tell the other company, "Hey, a terrorist account is linked to something on your service. You should check it out."

Mr. KATKO. It is similar to what you do in the malware arena, correct?

Mr. PICKLES. Yes. So industry collaboration is really at the heart of it.

Mr. KATKO. OK. Now, what companies are members of this? Is there a whole bunch, or is there just a limited number?

Mr. PICKLES. So, when we founded it, it was Google, Twitter—sorry—YouTube, Twitter, Microsoft, and Facebook. Dropbox has now joined.

One of the things we have is a partnership with Tech Against Terrorism, which allows small companies to go through a training process, so they learn things like how to write their terms of service, how to enforce their terms of service. By mentoring them, that is where—we are hopeful that we will have more companies joining and growing this, but the hash-sharing consortium has many members, 15 members. We share URLs with 13 companies.

So it is broad, but we want it to have a high standard. We want membership to be the companies who are doing the best, and that is why we want to keep a high bar and bring people in.

Mr. KATKO. I understand.

Now, as far as the encrypted messaging platforms, I take it they are not all members of this, they are not all participants on this, are they?

Mr. PICKLES. I am probably not the best person to answer that question.

Mr. KATKO. Would you know, Ms. Bickert?

Ms. BICKERT. Sure. Thank you for the question, Congressman. So the main members are, as Mr. Pickles mentioned, those 5 companies. Now, in terms——

Mr. KATKO. I understand.

Ms. BICKERT [continuing]. Of the smaller companies who have been trained, that does include some of the encrypted messaging services. Because some of this is about just understanding what are the right lines to draw, how to work with law enforcement authorities, which encrypted communication services can definitely do.

Mr. KATKO. Some of the—my biggest concern is that, while the big players in this field, all of you at the table, seem to be endeavoring to try and do the right thing, especially with respect to counterterrorism, that the encrypted messaging platforms, by and large, have a much broader field to play in, and there doesn't seem to be much we can do to stop their content from spreading their filth and their violence.

So I would love to hear any suggestions—I know my time is up, so perhaps in writing—as to how we could try and entice some of them to be part of this effort. The encryption is obviously a breeding ground for white supremacists, violence of all sorts. Trying to get the companies to be more responsible and just not worried about their bottom-line profit-making would be—would be great to hear from you guys. So thank you.

I yield back.

Chairman THOMPSON. The Chair recognizes the gentlelady from Michigan, Ms. Slotkin, for 5 minutes.

Ms. SLOTKIN. Good morning. Thanks for being here.

I wanted to switch gears for just a second and talk about the influence and the spread of foreign-based information, foreign-based political ads in particular, in our political process.

Many of us read the Mueller report page by page, and I was interested, Ms. Bickert, that the Facebook general counsel stated for the record that, for the low, low price of $100,000, the Russian-associated Internet Research Agency got to 126 million American eyeballs.

I am interested in this because the political ads that they put forward were specifically targeted to swing States, and Michigan is one of those States, so we saw an overabundance of these ads. They were specifically paid for by foreign entities, and they were advocating for or against a candidate in our political process. I have a serious problem with that.

So, separate from the issues of speech and what an American does or does not have the right to say, can you speak specifically to Facebook's reaction to the fact that they spread foreign, purchased information—and it doesn't matter to me that it was Russian; it could be Chinese or Iranian—and what steps you have taken since 2016 to prevent the spread of foreign information?

Ms. BICKERT. Absolutely, Congresswoman. Thank you for the question. Where we were in 2016—I mean, we are in a much, much better place. So let me share with you some of the steps we have taken.

First of all, all of those ads came from fake accounts. We have a policy against fake accounts, but we have gotten much better—and we had it then, but we have gotten much better at enforcing it. Now we are actually stopping more than a million accounts, fake accounts, per day at the time of upload. We publish stats on how many fake accounts we are removing every quarter, and you can see how much better we have gotten in the past 2 years.

Another thing that we are doing with political ads specifically is we are requiring unprecedented levels of transparency. Now, if you want to run a political or political issue ad in the United States, you have to first verify your identity. You have to show you are an American, which means we actually send you something—because we have seen fake IDs uploaded from advertisers—we send you something through the mail, and you actually then get a code, and you upload for us the government ID. So we verify that you are a real American.

Then we also put a "paid for" disclaimer on the political ad, and we put it in an ads library we have created that is visible to everybody. So, even if you don't have a Facebook account, you can go and see this ads library. You can search what type of political ads are appearing, who is paying for them, and other information about how they are being targeted and so forth.

Ms. SLOTKIN. That is good to hear. I am glad to hear it. I would love to see—if there are reports, I would love to just be directed to them so I can see them.

For the others at the table, can you talk about your specific—and brief, please—your specific policy on the spread of foreign political ads for or against a candidate running for office in the United States?

Mr. PICKLES. So the first thing we did was to ban Russia Today and all of its associated entities from using any of our advertising products going forward.

We took all of the revenue from Russia Today and their associated entities and are funding research and partnerships with organizations like the Atlantic Council, like the DisinfoLab in Brussels, to research better how we can prevent against this.

We then took the unprecedented step of publishing every tweet, not just the paid-for ones, every tweet that was produced by a foreign influence operation in a public archive. So you can now access more than 30 million tweets that runs to more than a terabyte of videos and photographs in a public archive. Those include operations from Russia, Iran, Venezuela, and other countries.

Ms. SLOTKIN. Mr. Slater.

Mr. SLATER. Thank you for the question.

Looking backward at 2016, we found very limited improper activity on our platforms. That is a product of our threat analysis group and our other tools to root out that sort of behavior.

Looking forward, we continue to invest in that, as well as our election transparency efforts, requiring verification of advertisers for Federal candidates, disclosure in the ads, and then a transparency report.

Ms. SLOTKIN. Great.

What about the spread of information through bots? What kind of disclosure requirement so that when someone is receiving or viewing something they have some way of knowing who produced it, who is spreading it, whether it is a human being, a machine?

Why don't we start with Facebook.

Ms. BICKERT. Thank you, Congresswoman.

One of our policies is that you have to have your real name and be using an account authentically. So, when we are removing bot accounts, we are removing them for being fake accounts. Those are all numbers that we publish.

Mr. PICKLES. Every week, we challenge between 8 million and 10 million accounts for breaking our rules on suspicious activity, including malicious automation. So we are removing those accounts. About 75 percent of those 8 million to 10 million challenge, fail those challenges, and they are removed every week.

Mr. SLATER. Congresswoman, for our part, we have strict policies about misrepresentation in ads, impersonation. We are looking out, again, through our threat analysis group, for coordinated, inauthentic behavior and will take action where appropriate.

Ms. SLOTKIN. Thank you.

I know my time has expired. Thank you.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentleman from Louisiana for 5 minutes, Mr. Higgins.

Mr. HIGGINS. Thank you, Mr. Chairman.

Mr. Slater, are you ready? Get your scripted answers ready, sir.

Google and YouTube are developing quite a poor reputation in our Nation. A clear history of repetitively silencing and banning voices. Conservatives or liberal, doesn't concern me right now. We are talking about freedom of speech and access to open communications.

We are here today to discuss extremist content, violent threats, terrorist recruiting tactics, and instigation of violence. To get the same justification your platform uses to quell true extremism is often used to silence and restrict the voices that you disagree with, and we don't like it.

For example, Prager University, a series of 5-minute videos which discuss political issues, religion, economic topics from a conservative perspective, has had over 50 of their videos restricted.

Some of their restricted videos include "Why America Must Lead." Perhaps that is a question that should be directed to the mirror. America leads because of our stance for freedom, for all voices to be heard. "The Ten Commandments/Do Not Murder" video—pulled by your people. What is wrong with the 10 commandments, might I ask? "Why Did America Fight the Korean War?", a legitimate reflection on a significant part of the history of our Nation—pulled.

Additionally, YouTube removed a video from Project Veritas which appears to show a senior Google executive acknowledging politically-motivated search manipulation with an intent to influence election outcomes. None of us here want that, on either side of this aisle. I don't know a man or woman present that is not a true patriot and loves their country. We have varying ideological perspectives, yes, but we love our country, and we will stand for freedom, including against Google.

A frequent reason provided by YouTube is that the content in question harmed the broader community. What could be more harmful to the broader community than the restriction of our free speech and open communications, regardless of our ideological stance?

Please define for America, what do you mean by "harmed the broader community" as it is used to justify restricting the content on Google or YouTube? And point out, is harm limited to physical threats and the incitement of violence, as it should be, or is it a convenient justification to restrict the content that you deem needs to be restricted?

Please explain to America how you determine what is "harmed the broader community," what does that mean. Let's have your scripted answer.

Mr. SLATER. Congressman, thank you for the question. I appreciate the concern and the desire to foster robust debate. We want YouTube to be a place where everyone can share their voice and get a view of the world.

Mr. HIGGINS. But you don't allow everyone to share their voice. I have given examples in my brief time—and thank you, Mr. Chairman, for recognizing my time.

The First Amendment protects Americans' right to express their viewpoints on-line. Is something that offends an individual or something an individual agrees with, does that meet your company's definition of extreme?

Mr. SLATER. We have community guidelines that lay out the rules of the road about what is not permitted on the platform, including incitement to violence, hate speech, harassment, and so on. If you can clarify what you are asking about specifically, I would be happy to try and answer.

Mr. HIGGINS. Mr. Slater, God bless you, sir. Google is in a bind. Today, America is watching. Today, America is taking a step back. We are looking at the services, we are looking at the platforms that we use, and we are finding, to our horror, that they can't be trusted.

Today, America is looking carefully at Google, and a word reverberates through the minds of America: Freedom. Shall it be protected, shall it be preserved, or shall it be persecuted and subject to the will and whim of massive tech companies?

Mr. Chairman, thank you for recognizing my time, and I yield the balance. Thank you for holding this hearing today.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentlelady from New York for 5 minutes, Ms. Clarke.

Ms. CLARKE. Thank you very much, Mr. Chairman.

I thank our panelists for appearing before us today.

I want to go into the issue of deepfakes, because I have recently introduced legislation, the first ever in a House bill, to regulate the technology. If my bill passes, what it would do is it would make sure that deepfake videos include a prominent, unambiguous disclosure as well as a digital watermark that can't be removed.

The question I have is, when it comes to your attention that a video has been substantially altered or entirely fabricated, how your companies decide whether to do nothing, label it, or remove it? That is for the panel.

Ms. BICKERT. Thank you for the question, Congresswoman.

Ms. CLARKE. Sure.

Ms. BICKERT. So, when it comes to deepfakes, this is a real top priority, especially because of the coming elections.

Right now, our approach is, we try to use our third-party fact-checking organizations. There are 45 of them world-wide. If they rate something as being false—they can also tell us that something has been manipulated. At that point, we will put the information from the fact-checking organization next to it. So, much like the label approach, this is a way of actually letting people understand that this is something that is, in fact, false. We also reduce the distribution of it.

We are also looking to see if there is something we should do specifically in the area of deepfakes. We don't want to do something in a one-off way; we want to have a comprehensive solution. Part of that means we have to get a comprehensive definition of what it means to actually have a deepfake. Those are conversations that we look forward to having with you.

Ms. CLARKE. Yes. My bill would require that there is a digital watermark and that it shows how—similar to how your companies do sort of a hash of terrorist content. If there was a central database of deceptive deepfake hashes, would you agree to utilize that?

Mr. Pickles.

Mr. PICKLES. I am happy to pick up on that and the previous question.

I was at a conference in London a few weeks ago hosted by the BBC and an NGO called Digital Witness, and they actually work on issues around verifying media from war zones of war crimes. So I think, actually, as Monika says, this policy goes from a whole spectrum of content, from synthetic to edited to manipulated. So I think, certainly, from our point of view, every partnership is one we want to explore to make sure that we have all the information.

I think your framing of how in some circumstances there may be situations to remove content, in other circumstances it is about providing context to the user and giving them more information, that is the best balance, I think, of making sure that we have all the tools available to us. That is the approach that we are developing now.

Ms. CLARKE. Yes. Time is not your friend here. What we are trying to find is something universal that creates transparency, respects the First Amendment, but also makes sure that, you know, it is something that, you know, as Americans whose eyes are constantly on video, something you can identify right away. If you have to go through all of these sources to determine—and each platform has a different way of indicating—it almost nullifies that.

So I wanted to put that on your radar, because I think that there needs to be some sort of a universal way in which Americans can detect immediately that what they are seeing is altered in some form or fashion. That is what my bill seeks to do.

Imagine if Russia, just days before the 2020 election, released a fake video of a Presidential candidate accepting a bribe or committing a crime.

If your companies learn of a deepfake video being promoted by a foreign government to influence our election, will you commit to removing it? How would you handle such a scenario?

Have you thought about it? Give us your thoughts. I don't have a whole lot of time.

Ms. BICKERT. Congresswoman, we do have a real name requirement on Facebook, and we also have various transparency requirements that we enforce. So if it is shared by somebody not in a real name or otherwise violating our transparency requirements, we would simply remove it.

Mr. PICKLES. We have a clear policy on affiliated behavior, so activity affiliated with an entity we have already removed. As I said, we have removed millions of tweets connected with the Internet Research Agency. We would remove any activity affiliated with that organization.

Mr. SLATER. Thank you for the question. It is a critical issue. I think we would evaluate such a video under our policies, including our deceptive practices policies, and look as we would at any sort of foreign interference.

Ms. CLARKE. Thank you very much.

Mr. Chairman, I yield back.

I look forward to talking to you further about this, because we have to get to that sweet spot, and we are not there, it is very clear from your testimony.

Chairman THOMPSON. Thank you.

The Chair recognizes——

Ms. CLARKE. Thank you, Mr. Chairman.

Chairman THOMPSON [continuing]. The gentlelady from Arizona, Mrs. Lesko, for 5 minutes.

Mrs. LESKO. Thank you, Mr. Chairman.

Years ago, required reading I had was the book "1984." This committee hearing is scaring the heck out of me, I have to tell you. It really is. Because here we are talking about, you know, if somebody googles "vaccines," the answer was, "Oh, we are going to put above what the person is actually looking for what we think is best." Who are the people judging what is best, what is accurate?

This is really scary stuff and really goes to the heart of our First Amendment rights. So I don't always agree with the ACLU—and you are the past president of ACLU, Ms. Strossen, but I agree with you wholly on this.

We have to be very careful, my colleagues, on this. Because what you deem as inaccurate I do not deem as inaccurate or other people may not deem. In a previous briefing on this issue, one of the Members said, "Well, I think President Trump's tweets incite terrorism." Well, are we now going to ban what President Trump says because somebody thinks that it incites terrorism?

This is some really scary stuff, and I am very concerned. I am glad I am part of this, because, boy, we need more of us standing up for our rights, whether it is what you believe or what I believe.

I have a specific question, and this is to Mr. Slater.

In this Project Veritas video, which I did watch last night, they allege that there are internal Google documents, which they put on the video, and this is what it said:

"For example, imagine that a Google image query for 'CEOs' shows predominantly men. Even if it were a factually accurate representation of the world, it would be algorithmic unfairness. In some cases, it may be appropriate to take no action if the system accurately reflects current reality, while in other cases it may be desirable to consider how we might help society reach a more fair and equitable state via product intervention."

What does that mean, Mr. Slater?

Mr. SLATER. Thank you, Congresswoman, for the question.

I am not familiar with the specific slide, but I think what we are getting at there is, when we are designing our products, again, we are designing for everyone. We have a robust set of guidelines to ensure we are providing relevant, trustworthy information. We work with a set of raters around the world, around the country, to make sure that those search rater guidelines are followed. Those are transparent and available for you to read on the web.

Mrs. LESKO. All right. Well, I personally don't think that answered the question at all, but let me go to the next one.

You asked, Mr. Clay Higgins, a specific example. So, Mr. Slater, he was talking about Prager University. I just googled—and I used Google—on "Prager University," and it came up. On the Prager University website, it says, "Conservative ideas are under attack. YouTube does not want young people to hear conservative ideas. Over 10 percent of our entire library is under 'restricted mode.' "

Why are you putting Prager University videos about liberty and those type of things on restricted mode?

Mr. SLATER. Thank you, Congresswoman. I appreciate the question.

To my knowledge, Prager University is a huge success story on YouTube, with millions of views, millions of subscribers, and so on. Remains so to this day.

There is a mode that users can choose to use called "restricted mode," where they might restrict the sorts of videos that they see. That is something that is applied to many different types of videos from across the board, consistent not with respect to political viewpoints but applied to, for instance, "The Daily Show," other sorts of channels as well.

To my knowledge, it has been applied to a very small percentage of those videos on Prager University. Again, that channel has been a huge success story, I think, with a huge audience on YouTube.

Mrs. LESKO. Mr. Pickles, regarding Twitter, President Trump has said, I think on multiple occasions, that—he has accused Twitter of, you know, people having a hard time—being deleted from followers. This actually happened to my husband. He followed Donald Trump, and then, all of a sudden, it was gone.

So can you explain that? What is happening there? Why does that happen? Because I tell you, a lot of conservatives really think there is some conspiracy going on here.

Mr. PICKLES. Well, I can certainly look into the case of your husband to make sure there wasn't an issue there.

What I can say is that President Trump is the most followed head of state anywhere in the world. He is the most talked-about politician anywhere in the world on Twitter. Although he did lose some followers when we recently undertook an exercise to clean up compromised accounts, President Obama lost far more followers in the same exercise.

So I think people can look at the way that people are seeing President Trump's tweets widely and be reassured that the issues that you are outlining there are not representative in Twitter's approach.

Mrs. LESKO. Mr. Chairman, I ran out of time, but if we have another round, I really want to hear from Ms. Strossen. I want to hear her views, because she hasn't had a lot of time to speak, so I hope some of my fellow colleagues ask her.

Thank you.

Ms. STROSSEN. Thank you.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentlelady from California, Ms. Barragán, for 5 minutes.

Ms. BARRAGÁN. Thank you very much, Mr. Chairman.

This is for Ms. Bickert, Mr. Pickles, and Mr. Slater. I want to talk a little bit about your relationship with civil society groups that represent communities targeted by terrorist content, including white supremacist content. I am specifically referring to content that targets religious minorities, ethnic minorities, immigrants, LGBTQ, and others.

Can you help by describing your engagement with civil society groups in the United States to understand the issues of such content and develop standards for combating this content?

Ms. BICKERT. Thank you for the question, Congresswoman.

Yes, any time that we are evolving our policies, which we are doing constantly, we are reaching out to civil society groups, not just in the United States but around the world. I have a specific team under me, actually, called Stakeholder Engagement. That is what they do.

When they are doing this, one of their jobs is to make sure—let's say we are looking at our hate speech policies. One of their jobs is to make sure that we are talking to people across the spectrum, so different groups that might be affected by the change, people who will have different opinions. All of those people are brought into the conversation.

Mr. PICKLES. Well, similarly, we have teams around the world who are speaking to civil society groups every day. Something we are also doing is training them, and I think it is really important. Because Twitter is a unique public platform and a public conversation, when people actually challenge hatred and offer a counternarrative, offer a positive narrative, their views can be seen all over the world.

So, you know, "Je Suis Charlie" was seen all over the world after an attack in Paris. Similarly, after Christchurch, "Hello Brother," or even "Hello Salam," which was a gentleman in Kenya who challenged a terrorist who was trying to separate Christians and Muslims.

So we talk to civil society groups both about our policies but also how they can use our platform more to reach more people with their messages.

Ms. BARRAGÁN. OK.

Then, Mr. Slater, before you start, because I want to make sure you incorporate this, one of my concerns is the onus to report the hateful content is placed on the very communities that are targeted by the hateful content. That can make social media platforms hostile places for people in targeted communities. So can you also tell us what your companies are doing to alleviate this burden?

So Mr. Slater, and then I would like to hear from the two of you on that.

Mr. SLATER. Sure.

Speaking of how we enforce our community guidelines, including against hate speech, including, again, as we said, we have updated our hate speech policies to deal with people expressing superiority to justify discrimination and so on. We use a combination of machines and people—machines to scan across for broad patterns and so on, compared to previous violative content.

So we do take our responsibility here very seriously, our ability to detect that first, review it before it has been flagged. You know, we are making great strides in that.

We also do rely on flags from users, as well as flags from trusted flaggers—that is, civil society groups, other experts, who we work with very closely both in the development of our policies and then again in flagging those sorts of videos.

Ms. BARRAGÁN. Yes.

So, just to the two of you, about the burden?

Mr. PICKLES. This is something that we have said previously; there was too much burden on victims. A year ago, 20 percent of the abuse we removed was surfaced proactively. That is now 40 percent. So, in a year, we have been able to double the amount of content that we find proactively without waiting for a victim to review it. We are continuing to invest to raise that number further.

Ms. BARRAGÁN. Can the three of you provide an example where you had community engagement and, because of that feedback, there was a policy change that you made?

Mr. PICKLES. Let me share a slightly different example, which is how we write a better policy to prevent that.

So, when we were crafting a policy on nonconsensual intimate imagery, that covers not just media shared by an ex-partner, but it might share creep shots, which I think have been—so various countries start asking, do you have a policy on creep shots? Because we had spoken to those very civil society groups, our policy from the beginning was reaching broadly enough to capture not just the original problem but all those different issues.

Ms. BARRAGÁN. Ms. Bickert.

Ms. BICKERT. Yes. Let me address the second question that you asked about, putting the burden on the victims.

We have invested a lot in artificial intelligence. So there are certain times when artificial intelligence has really helped us and other areas where it is very much in its infancy. With hate speech, over the past few years, we have gone from zero proactive detection to now, in the first quarter of this year, the majority of content that we are removing for violating our hate speech policies we are finding using artificial intelligence and other technology.

So huge gains there. There is still a long way to go because all of those posts, after they are flagged by technology, have to be reviewed by real people who can understand the context.

Ms. BARRAGÁN. Right.

Ms. BICKERT. In terms of where our engagement has led to concrete changes, one thing I would point to is the use of hate speech in imagery. The way that we originally had our policies on hate speech, it was really focused on what people were saying in text. It was only through working with civil society partners that we were able to see how we needed to refine those policies to cover images too.

Another thing I would point to is, a lot of groups told us it was hard to know exactly how we defined hate speech and where we drew the line. That was a contributing factor, among many others, in why a couple years ago we published a very detailed version of our community standards, where now people can see exactly how we define hate speech and how we implement it.

Ms. BARRAGÁN. Great. Thank you.

I yield back.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentleman from Texas for 5 minutes, Mr. Crenshaw.

Mr. CRENSHAW. Thank you, Mr. Chairman.

Thank you for some of the thoughtful discussion on how you combat terrorism on-line. I think there are worthy debates to be had there. There are good questions on whether, you know, some of this content provides education so that we know of the bad things out there or whether it is radicalizing people. Those are hard discussions to have, and I don't know that we are going to solve them today.

But the problem is that the testimony doesn't stop there; the policies at your social media companies do not stop there. It doesn't stop with the clear-cut lines of terrorism and terrorist videos and terrorist propaganda. Unfortunately, that is not exactly what we are talking about. It goes much further than that. It goes down the slippery slope of what speech is appropriate for your platform and the vague standards that you employ in order to decide what is appropriate.

This is especially concerning given the recent news and the recent leaked emails from Google. They show that labeling mainstream conservative media as "Nazis" is a premise upon which you operate. It is not even a question, according to those emails. The emails say, given that Ben Shapiro, Jordan Peterson, and Dennis Prager are Nazis, given that that is a premise, what do we do about it?

Two of three of these people are Jewish, very religious Jews, and yet you think they are Nazis. It begs the question, what kind of education do people at Google have so they think that religious Jews are Nazis?

Three of these people had family members killed in the Holocaust. Ben Shapiro is the No. 1 target of the alt-right, and yet you people operate off the premise that he is a Nazi. It is pretty disturbing.

It gets to the question, do you believe in hate speech—how do you define that? Can you give me a quick definition right now? Is it written down somewhere at Google? Can you give me a definition of hate speech?

Mr. SLATER. Congressman, yes. So hate speech, again, as updated in our guidelines, now extends to superiority over protected groups that justify discrimination, violence, and so on based on a number of defining characteristics, whether that is race, sexual orientation, veteran status——

Mr. CRENSHAW. Do you have an example of Ben Shapiro or Jordan Peterson or Dennis Prager engaging in hate speech? Do you have one example off the top of your head?

Mr. SLATER. So, Congressman, we evaluate individual pieces of content based on that content rather than based on the speaker.

Mr. CRENSHAW. OK. Let's get to the next question. Do you believe speech can be violence? All right, now, not can you incite violence; that is very clearly not protected. But can speech just be violence? Do you believe that speech that isn't specifically calling for violence can be labeled violence and, therefore, harmful to people? Is that possible?

Mr. SLATER. Congressman, I am not sure I fully understand the distinction you are drawing. Certainly, again, incitement to violence or things that are——

Mr. CRENSHAW. Right.

Mr. SLATER [continuing]. Encouraging dangerous behavior, those are things that would be against our policies.

Mr. CRENSHAW. Here is the thing. When you call somebody a Nazi, you can make the argument that you are inciting violence, and here is how. As a country, we all agree that Nazis are bad. We actually invaded an entire continent to defeat the Nazis. It is normal to say hashtag-punch-a-Nazi, because there is this common thread in this country that they are bad and that they are evil and that they should be destroyed.

So, when you are operating off of that premise—and, frankly, it is a good premise to operate on—well, what you are implying, then, is that it is OK to use violence against them. When you label them, one of the most powerful social media companies in the world, labels people as Nazis, you can make the argument that is inciting violence. What you are doing is wholly irresponsible.

It doesn't stop there. A year ago, it was also made clear that your fact-check system is blatantly targeting conservative newspapers. Do you have any comments on that? Are you aware of the story I am talking about?

Mr. SLATER. I am not familiar with necessarily the specific story, Congressman. I am aware that, from all political viewpoints, we sometimes get questions of this sort. I can say that our fact-check labels generally are done algorithmically based on a markup and follow our policies——

Mr. CRENSHAW. For the record, they specifically target conservative news media. Oftentimes they don't even—they have a fact-check on there that doesn't even reference the actual article, but Google makes sure that it is right next to it so as to make people understand that that one is questionable, even though, when you actually read through it, it has nothing to do with it.

You know, a few days ago—and this goes to you, Ms. Bickert—one of my constituents posted photos on Facebook of Republican women daring to say that there are women for Trump. Facebook took down that post right away, with no explanation. Is there any explanation for that?

Ms. BICKERT. Without seeing it, it is hard for me to opine. That doesn't violate our policies. But I am happy to follow up on the specific example with you.

Mr. CRENSHAW. Thank you.

Listen, here is what it comes down to. If we don't share the values of free speech, I am not sure where we go from here. You know, this practice of silencing millions and millions of people, it will create wounds and divisions in this country that we cannot heal from.

This is extremely worrisome. You have created amazing platforms; we can do amazing things with what these companies have created. But if we continue down this path, it will tear us apart.

You do not have a Constitutional obligation to enforce the First Amendment, but I would say that you absolutely have an obligation to enforce American values. The First Amendment is an underpinning of American values that we should be protecting until the day we die.

Thank you.

Thank you for indulging me, Mr. Chairman.

Chairman THOMPSON. Thank you.

Ms. Strossen, the Chair is going to take prerogative and allow you to make a comment if you would like.

Ms. STROSSEN. Oh, thank you so much for protecting my free-speech rights, Mr. Chairman.

The main point that I wanted to make is that, even if we have content moderation that is enforced with the noblest principles and people are striving to be fair and impartial, it is impossible. These so-called standards are irreducibly subjective. What one person's hate speech is—and an example was given by Congressman Higgins—is somebody else's cherished, loving speech.

For example, in European countries, Canada, Australia, New Zealand, which generally share our values, people who are preaching religious texts that they deeply believe in and are preaching out of motivations of love are prosecuted and convicted for engaging in hate speech against LGBTQ people. Now, I obviously happen to disagree with those viewpoints, but I absolutely defend their freedom to express those viewpoints.

At best, these so-called standards—and I did read every single word of Facebook's standards. The more you read them, the more complicated it is. No two Facebook enforcers agree with each other, and none of us would either.

So that means that we are entrusting to some other authority the power to make decisions that should reside in each of us as individuals, as to what we choose to see and what we choose not to see and what we choose to use our own free-speech rights to respond to.

On that, I think these platforms have—I cannot agree more about the positive potential, but we have to maximize that positive potential through user empowerment tools, through radically increased transparency.

One of the problems of this——

Chairman THOMPSON. I am not going to limit your speech; I am going to limit your time.

Ms. STROSSEN. Thank you.

Chairman THOMPSON. Congressman Correa for 5 minutes.

Mr. CORREA. Thank you, Chairman Thompson and the Ranking Member, for holding this very critical hearing on very interesting, very important issues.

I want to turn a little bit to the Russian interference in 2016. The Mueller report outlines indictment of 13 Russians, 3 companies for conspiring to subvert our election system. In 2018, we saw indications that, again, Russians were at it again. In 2020, former Secretary of Homeland Security Nielsen, before she was resigned—she resigned—brought up the fact that the Russians were at it for 2020 again. There are other countries also trying to affect our election system.

So I am hearing your testimony, and my question, of course, Ms. Strossen, addressing the issue of First Amendment: Does the First Amendment cover fake videos on-line?

We talked a little bit about the Pelosi fake video, and maybe you say "yes." I probably say "probably not." I will tell you why. Because that is a damaging video with false content. Although you may be private companies, when I hear my children tell me, I saw it on this platform, the assumption is that it is factual.

Ms. Bickert, it took you 24 hours to take that video down. The others didn't take it down.

You are essentially a messenger, and when your information shows up on-line, this population believes that you are credible and that the information on there is probably credible too. That is what is damaging to our country, to our democracy.

Moving forward, we have another election happening now, and if this information continues to be promulgated through your social media, through your companies, we have a First Amendment issue, but we have an issue, also, of democracy and keeping it whole.

Any thoughts?

Ms. Bickert.

Ms. BICKERT. Thank you for the question, Congressman.

We share the focus on making sure that we are ready——

Mr. CORREA. Twenty-four hours is not fast enough. So are we playing here defense or offense? Are we reacting? Are you being proactive so the next Nancy Pelosi video is something you can take down essentially faster than 24 hours?

Ms. BICKERT. Congressman, we are being proactive. I do agree that there is a lot that we can do to get faster.

Our approach when there is misinformation is making sure that people have the context to understand it. We don't want people seeing it in the abstract. We want to make sure we are informing people, and we have to do so quickly. So that is something that we are focused on getting better at.

Mr. CORREA. So let me ask you something. On the Pelosi video, who put it up?

Ms. BICKERT. It was uploaded by a regular person with a regular account.

Mr. CORREA. So somebody at home with some very smart software and a good platform was able to put together a fake video and put it up.

Ms. BICKERT. Congressman, the technique that was used was to slow down the audio, which is the same thing we see a lot of comedy shows, frankly, do——

Mr. CORREA. OK.

Ms. BICKERT [continuing]. With a lot of politicians——

Mr. CORREA. So what were the consequences to this individual for putting up essentially a video of somebody, defaming, you know, hurting her reputation?

Ms. BICKERT. Congressman, that video—our approach to misinformation is we reduce the distribution, and then we put content from fact-checkers next to it so that people can understand that the content is false or has been manipulated.

Mr. CORREA. Mr. Pickles.

Mr. PICKLES. Well, one of the things we talked about earlier was how to provide context to users. So our focus now is developing——

Mr. CORREA. Well, are your policies changing so that you will be able to take it down next time, or are you just going to let it ride?

Mr. PICKLES. Well, we are looking at all of our policies in this area.

Mr. CORREA. Are you going to look at taking it down, or are you going to let it ride? A "yes" or a "no."

Mr. PICKLES. Well, I think we are looking at both how do you give more——

Mr. CORREA. Mr. Slater, what are you going to do?

I didn't get an answer.

Mr. Slater, what are you going to do next time you see a video like this?

Mr. SLATER. With respect to that video, to be clear, we took it down, under our deceptive practices policy.

Mr. CORREA. Ms. Strossen, not to, you know, violate your freedom of speech here, do you think these false videos on-line are Constitutionally protected?

Ms. STROSSEN. There is a very strict definition of false speech that is Constitutionally unprotected. The Supreme Court has repeatedly said that blatant, outright lies are Constitutionally protected unless——

Mr. CORREA. So let me switch in my 7 seconds I have left. Will you write policy so outright lies do not have the devastating effect on our voters that they had in the 2016 election?

Mr. PICKLES. As I said, we are looking at the whole issue.

Mr. CORREA. Thank you.

Ms. Bickert, Mr. Slater, any thoughts?

Ms. BICKERT. We, too, Congressman, are making sure that we have the right approach for the election.

Mr. CORREA. Thank you.

Mr. Slater.

Mr. SLATER. Absolutely. We want to raise up authoritative content, reward it, and then demote borderline content, harmful misinformation, and remove violative content.

Ms. STROSSEN. If I may say, this is exactly the reason why President Trump wants to change the libel laws, because it is now legal to lie about politicians and Government officials.

Mr. CORREA. Maybe there is an area we will work together on some issues, huh?

Mr. Chairman, I yield.

Chairman THOMPSON. Thank you.

The Chair now recognizes the gentlelady from New Jersey, Mrs. Watson Coleman, for 5 minutes.

Mrs. WATSON COLEMAN. Thank you very much, Mr. Chairman. Thank you for being here. This has been very informative.

Let me ask you a really quick question, "yes" or "no." The GIFCT—is that right?—GIFCT, your collaboration, does keeping your trade secrets secret interfere with your sharing standards and, you know, working together to——

Mr. PICKLES. I don't——

Mrs. WATSON COLEMAN. "Yes" or "no"?

Mr. PICKLES. I don't think it has to, no.

Mrs. WATSON COLEMAN. OK.

I know you use this platform for terrorism. Do you use that platform at all for, sort-of, hate groups?

Mr. PICKLES. Not at present, but, certainly, after New Zealand, that highlighted that we do need to broaden our approach to different issues.

Mrs. WATSON COLEMAN. Uh-huh.

So, in my briefing, dog whistling has been mentioned as a certain kind of political messaging strategy that employs coded language to send a message to certain groups that flies under the radar. It is used by white supremacist groups often. It is rapidly evolving on social media platforms and has its—a space and targeting of racism and other sort-of -isms that we find abhorrent in this country.

How do you solve the challenge of moderating dog-whistle content on your platform, especially when it is being used to encourage these -isms that we abhor so much?

Mr. PICKLES. I am happy to start and then let others finish.

Mrs. WATSON COLEMAN. I would—yes. I will take 1, 2, 3.

Mr. PICKLES. Well, first, we enforce our rules, and one of the things that our rules are is about behavior. So, if you are targeting somebody because of their membership of a protected characteristic, that is the important factor. The words come secondary.

GIFCT has an entire stream of research, and one of the reasons for having that research stream is so that we can investigate what are the latest trends, what are the things we need to be learning about those kind of terms.

Then, finally, when we see, whether it is different kinds of extremist groups, speaking for Twitter, we have banned more than 180 groups from our platform for violent extremism across the spectrum, both in the United States and globally.

So we have a policy framework and also the industry sharing.

Mrs. WATSON COLEMAN. Thank you.

Ms. BICKERT. Thank you, Congresswoman.

I would echo that a lot of this is about getting to the groups. We do have a hate speech policy, but, beyond that, we know that sometimes there are groups that are just engaging in bad behavior. So we ban not only violent groups but also white supremacist groups and other hate groups. We have removed more than 200 of them from our platform to date.

Mr. SLATER. Thank you for the question.

We do, as I said, remove hate speech on our platform. The sort of concerns you are talking about is what motivated the more recent changes.

We also recognize that things may brush up against those policies, be borderline, but not quite cross them. For those, we do work to reduce, demote them in the frequency and recommendations and so on.

Ms. STROSSEN. Congresswoman, if I could have just 10 seconds——

Mrs. WATSON COLEMAN. I am going to ask you a question, so you can have a little bit more than that.

Ms. STROSSEN. Thank you.

Mrs. WATSON COLEMAN. This is a very quick question. Ms. Bickert, did you bring any staff here with you today, any employees from your——

Ms. BICKERT. Yes, we did.

Mrs. WATSON COLEMAN. Could you please have them stand up? For those that have accompanied Ms. Bickert, could you please stand up?

Two.

Thank you very much.

Mr. Pickles, you?

Mr. PICKLES. Yes.

Mrs. WATSON COLEMAN. Thank you.

Mr. Slater.

Thank you very much.

A couple of things that you mentioned. You talked about making sure that the people are real and that they are American when they are going to do advertising. You said we are going to send information to you, you have to send it back, and it just simply proves that you are maybe pretending to be an American living—and really living here or having a domicile here, an address here, still doesn't necessarily guarantee that they are legitimate. So that is a challenge, I think, that we might have.

Is that understandable, Mr. Slater, or am I confusing you?

Mr. SLATER. If you could clarify the question, I would appreciate it.

Mrs. WATSON COLEMAN. It is not a question; it is a statement. We were talking earlier about making sure that people who are doing political advertising, et cetera, are not foreign nationals, that they are Americans. Did we not have this discussion about this advertising? It was stated by somebody there—thank you.

Ms. BICKERT. That's right.

Mrs. WATSON COLEMAN [continuing]. That you do verification to make sure that the person is an American, does live in America, and isn't this false whatever coming from another nation. I said, that really doesn't necessarily prove that, as far as I am concerned.

Ms. BICKERT. Congresswoman, just to clarify, that is Facebook's approach. We do verify——

Mrs. WATSON COLEMAN. Right. I have to give her the 10 percent. I have to give her the——

Ms. BICKERT. Oh, sorry. We also—we look at a Government ID.

Mrs. WATSON COLEMAN. Because my question to you is, are there trigger words that come out of some of this speech that you think should be protected that needs to be taken down because it incites?

Ms. STROSSEN. All of them, it is a problem.

I wanted to give an example from a story in *Bloomberg News* today that talked about YouTube's recent new policy of broadening the definition of unprotected hate speech. On the very first day that it went into effect, one of the people that was suppressed was an on-line activist in the United Kingdom against anti-Semitism. But, in condemning anti-Semitism, he was, of course, referring to Nazi expression and Nazi insignia, and, hence, he was kicked off.

Mrs. WATSON COLEMAN. So there are no trigger words. It seems to me that—I think it was Mr. Pickles. Did you do the definition of hate speech for us earlier?

Mr. PICKLES. That was the hateful conduct under Twitter's——

Mrs. WATSON COLEMAN. Yes. I think that that probably covers the President of the United States of America, unfortunately.

Thank you, Mr. Chairman. I yield back.

Chairman THOMPSON. The Chair recognizes the gentleman from New York, Mr. Rose, for 5 minutes.

Mr. ROSE. Mr. Chairman, thank you.

Thank you all for being here.

Two months ago, in the immediate aftermath of the Christchurch incident, we sent out a letter to you all, asking, how much money are you spending on counterterrorist screening, and how many people do you have allocated to it? We have had interesting conversations over those ensuing months.

The 3 basic problems that you have brought to me are that, No. 1, that oversimplifies it because there is also an AI component to this. Well, yesterday, we did a hearing that showed AI alone cannot solve this, impossible, and not into the future. You all agree with that.

The second thing, though, that you have all said to me is that this is a collective action problem, we are all in this together, and we have the GIFCT. So I have some very basic questions about the GIFCT. I would appreciate it if you could just immediately answer "yes" or "no," and then we can get into the details.

First question: Does the GIFCT have any full-time employees?

Ms. Bickert. Does the GIFCT have a full-time employee dedicated to it to run it?

Ms. BICKERT. No. We have people at Facebook full-time dedicated to GIFCT.

Mr. ROSE. OK.

Mr. PICKLES. The same. We have people at Twitter working with GIFCT, but we don't have—GIFCT doesn't have staff.

Mr. ROSE. OK.

Mr. SLATER. Yes, our answer is the same.

Mr. ROSE. Does the GIFCT have a brick-and-mortar structure? If I want to go visit the GIFCT, could I do so?

Ms. Bickert.

Ms. BICKERT. No, Congressman.

Mr. ROSE. OK.

Ms. BICKERT. We do host the database physically at Facebook.

Mr. ROSE. OK.

Mr. Pickles.

Mr. PICKLES. No. Our collaboration is 4 companies working together. We meet in person; we have virtual meetings. It is about collaboration, not about a physical building.

Mr. ROSE. OK.

Mr. Slater.

Mr. SLATER. That is right. Nothing further to add.

Mr. ROSE. So no brick-and-mortar structure, but I presume you have a Google Hangout or maybe a Facebook hangout. I don't know how you would decide that.

But Adhesive and Sealant Council, an association located in Bethesda, Maryland, at the Adhesive and Sealant Council, it has 5 full-time staff, it has a brick-and-mortar structure. You all cannot get your act together enough to dedicate enough resources to put full-time staff under a building dealing with this problem.

I think it speaks to the ways in which we are addressing this with this technocratic, libertarian elitism. All the while, people are being killed. All the while, there are things happening that are highly preventable.

AI. Are there any AI systems that any of you all have that are not available to the GIFCT?

Ms. BICKERT. Congressman, yes, depending on how our products work. They all work differently, so artificial intelligence works differently.

What we have—and we actually worked for some time on doing this. We had to come up with one common technical solution that everybody could use. We now have that for videos, and we do give it for free to smaller companies. But that is but one technique we have.

Mr. ROSE. OK.

Please, just keep it—I just want to know if you have any AI not—that the GIFCT doesn't have, though.

Mr. PICKLES. Well, I would also say that this isn't just AI. That is why we share URLs—very low-tech, lo-fi. But if you are a small company and someone gives you a URL to content, you don't need AI to look at that. So I think that is why it is a combination solution.

Mr. ROSE. Uh-huh.

Mr. SLATER. Nothing further to add to those comments.

Mr. ROSE. OK.

My understanding is that there were no officially declared POCs for the GIFCT that were made public from each company until after the Christchurch shooting. I know that they were there, but they were not declared established POCs at each of your companies until after the Christchurch shooting 2 months ago. Is this the case?

Ms. BICKERT. Congressman, we have a channel that people can use that gets routed to whoever is on-call from our team.

Mr. ROSE. But is that the case, that there were no established POCs—and this is the information you all have given me already; I am just asking you to put it on the record—no established POCs at the GIFCT until after the Christchurch shooting? Is that correct?

Ms. BICKERT. Perhaps not publicly listed, but certainly people know who to call——

Mr. ROSE. No established public POCs until after the Christchurch shooting.

Mr. PICKLES. Well, I would draw a distinction between the POCs and the companies. We work together every week, every day. I think the point you are getting at is crisis response is——

Mr. ROSE. I am getting to the fact that you are not taking it seriously, because there is no public building, there is no full-time staff, there were no public POCs until after the Christchurch shooting.

Mr. PICKLES. Well, I think——

Mr. ROSE. That is what I am speaking to. How is anyone supposed to think that you all take this collective action problem seriously if you have no one working on it full-time?

This is not something that technology alone can solve. This is a problem that we are blaming the entire industry for, rightfully so. There are the smallest of associations in this town and throughout the country that do so much more than you do.

It is insulting—it is insulting that you would not at least apologize for saying that there were no established POCs prior to the Christchurch shooting. It was a joke of an association, it remains a joke of an association, and we have got to see this thing dramatically improved.

Last, if there were terrorist content shown to be on your platforms by a public entity, would you take it down?

So, Ms. Bickert, why when the whistleblower association reveals that you Facebook is establishing through its AI platform al-Qaeda community groups, such as this one, a local business, al-Qaeda in the Arabian Peninsula, with 217 followers—I have it right here on my phone—by the whistleblower association. It is considered the most active of al-Qaeda's branches, or franchises, that emerged due to weakening central leadership. It is a militant Islamist organization primarily active in Yemen and Saudi Arabia. Why is this still up?

We have every right right now to feel as if you are not taking this seriously. By "we," I do not mean Congress; I mean the American people.

Thank you.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentlelady from Florida, Mrs. Demings, for 5 minutes.

Mrs. DEMINGS. Thank you so much, Mr. Chairman.

We have already talked about the massacre at Christchurch, and we also know that it was law enforcement who notified Facebook about what was going on.

Ms. Bickert, I would like to know if you could talk a little bit about your working relationship with law enforcement and share some of the specific things that you are doing to further enhance your ability to work with law enforcement to continue to work to prevent incidents like this from happening again.

Ms. BICKERT. Thank you, Congresswoman.

We have a special point of contact from our law enforcement engagement team, so people from within our company, usually former law enforcement, who are assigned to each company. Those relationships are well-functioning and are the reason that New Zealand law enforcement were able to reach out to us. Once they did, within minutes——

Mrs. DEMINGS. You surely believe that they would have been able to reach out to you if you didn't have a law enforcement team, right? Wouldn't that have been part of their responsibility, any law enforcement agency that saw what was happening live on your platform, to notify you?

Ms. BICKERT. Congresswoman, we want to make it very easy, if they see something, that they know exactly where to go.

It is also a reason—so, here, with New Zealand, when they reached out to us, we responded within minutes. We also have an on-line portal through which they can reach us, and that is manned 24 hours a day. So if there is any kind of an emergency, we are on it.

Finally, if we see that there is an imminent risk of harm, we proactively reach out to them.

I will also tell you, any time that there is a terror attack or some sort of mass violence in the world, we proactively reach out to law enforcement to make sure that if there are accounts we should know about or names of victims, any sort of action that we should be taking, that we are on it immediately.

Mrs. DEMINGS. OK.

Moving right along, Mr. Pickles, you said that we will not solve the problems by moving content alone. Is that correct, what you said?

Mr. PICKLES. Yes.

Mrs. DEMINGS. OK. I know that most companies do a pretty good job in terms of combating or fighting child exploitation or pornography. I would just like to hear you talk a little bit about your efforts to combat terrorism and share some of the similarities. Because we can't solve the problems by just taking down the content alone.

So if you could just show some of the similarities in terms of your efforts of combating terrorism along with your efforts to combat child pornography. I know you put a lot of resources in combating child pornography, rightfully so. But could you talk about the similarities in the two goals?

Mr. PICKLES. Absolutely. There are similarities and differences. In the similarities space, we are able to use similar technology to look for an image we have seen before. If that appears again, we can proactively detect that image and stop it being distributed and then, critically, work with law enforcement to bring that person to—so we work with the National Center for Missing and Exploited Children, who work with law enforcement around the world.

So that process of discovering content, working with law enforcement is seamless. Because I think, particularly for child sexual exploitation but also for violent threats, we——

Mrs. DEMINGS. So what about for——

Mr. PICKLES [continuing]. We need people——

Mrs. DEMINGS [continuing]. Combating terrorism?

Mr. PICKLES. So I think, in either case, if someone is posting that content, removing the content is our response, but there is a law enforcement response there, as well, which holds people to account, potentially prosecutes them for criminal offenses. And that working in tandem between the two is very important.

We have a similar industry body that shares information. We also work with governments to share threat intelligence and analysis of trends so that we can make sure we are staying ahead of bad actors.

But the biggest area of similarity is, the bad actors never stay the same. They are constantly evolving. So we have to constantly be looking for the next opportunity to improve——

Mrs. DEMINGS. OK. All right. Thank you.

At the beginning of this conversation, the Chairman asked a question about—or referenced the video of the Speaker and why some of you removed it and some did not.

Mr. Slater, I was so pleased to hear your answer, which was—you look for deceptive practices? If it was deceptive, you removed it, correct?

Could you just talk a little bit more about—it seemed like such a—and, Ms. Strossen, I believe you said that the social media platforms' free-speech right is their ability to decide what is posted and what is not posted.

Ms. STROSSEN. Exactly.

Mrs. DEMINGS. It is just that simple, right? They can decide what is posted and what is not posted.

So, Mr. Slater, if you could just talk a little bit about your process, and it was deceptive, you took it down.

Mr. SLATER. I would be happy to, Congresswoman. It is an important question.

We have community guidelines. One of those guidelines is about deceptive practices. We review each bit of content thoroughly to make sure whether it is violative or whether it may fit into an exception—education, documentary, and so on and so forth—and do that on an individualized basis to see if the context has been met.

We present those guidelines publicly on our website for anyone to read.

Mrs. DEMINGS. Thank you very much.

Mr. Chair, I yield back.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentleman from Texas, Mr. Taylor, for 5 minutes.

Mr. TAYLOR. Thank you, Mr. Chairman.

Just a quick question. So is Google an American company?

Mr. SLATER. Congressman, we are headquartered in California, yes.

Mr. TAYLOR. Are you loyal to the American Republic? I mean, is that something you think about? Or do you think of yourselves as an international company?

Mr. SLATER. We build products for everyone. We have offices all across this country, have invested heavily in this country, and are proud to be founded and headquartered in this country.

Mr. TAYLOR. So, if you found out that a terrorist organization was using Google products, would you stop that? Would you end that?

Mr. SLATER. We have a policy, Congressman, of addressing content from designated terrorist organizations, to prohibit it, make sure it is taken down.

Mr. TAYLOR. I am not asking about content. I am saying, if you found that al-Nusrah terrorist organization was using Gmail to communicate inside that terrorist organization, would you stop that? Do you have a policy on that?

If you don't have a policy, that is fine. I am just trying to—where are you on this?

Mr. SLATER. Certainly. Where appropriate, we will work with law enforcement to provide information about relevant threats, illegal behavior, and so on. Similarly, we will respond to valid requests for information from law enforcement.

Mr. TAYLOR. I am not asking if you respond to subpoenas. I appreciate that. It is good to hear that you deign to be legal.

What I am asking is, if a terrorist organization uses a Google product and you know about that, do you allow that to continue? Or do you have a policy?

Mr. SLATER. Under the appropriate circumstances and where we have knowledge, we would terminate a user and provide information to law enforcement.

Mr. TAYLOR. OK. So you will forgive me for not—your answer is a little opaque. I am still trying to figure this out.

So, if a terrorist organization is using a Google product, do you have a policy about what to do about that?

Mr. SLATER. Thank you, Congressman. I am attempting to articulate that policy. I would be happy to come back to you with further information if it is unclear.

Mr. TAYLOR. OK. Do——

Mrs. DEMINGS. Would the gentleman yield?

Mr. TAYLOR. Sure.

Mrs. DEMINGS. Listen to the answer about referring it to law enforcement. I think that is an appropriate response, because if there is a suspicion that criminal activity is afoot, you would want to refer it to law enforcement and law enforcement make the call on that.

Mr. TAYLOR. Sure.

Mrs. DEMINGS. So just to kind-of——

Mr. TAYLOR. OK.

Mrs. DEMINGS [continuing]. Maybe help you a little bit with that particular portion of it. But——

Mr. TAYLOR. Thanks, Chief.

Mrs. DEMINGS [continuing]. Back to the policy. Thank you.

Mr. TAYLOR. Appreciate it.

Just to kind-of follow up with that, so the Islamic Republic of Iran is the largest state sponsor of terrorism in the world, right? They are a terrorist—and, you know, pieces of the Islamic Republic are terrorist organizations. Do you have a specific ban on that terrorist organization and their ability to use your Google products?

Mr. SLATER. Congressman, we have prohibitions on designated terrorist organizations using products, uploading content, and so on.

Mr. TAYLOR. OK. So you seek to ban terrorist organizations from using Google products?

I am not trying to put words in your mouth. I am just trying to understand your position on this.

Mr. SLATER. Designated terrorist organizations, we have prohibitions on that sort of organization, correct?

Mr. TAYLOR. I am not just asking about content. I am asking about the services you provide. Right? You provide Gmail, you provide iCalendar, you provide a whole host of different services that people can use. I am trying to ask about the services, not the content. I realize that the focus of this hearing is about content, which is why you are here, but I am asking about the actual services.

Mr. SLATER. To the best of my knowledge, if we were to have knowledge—and, again, as my colleagues have said, these bad actors are constantly changing their approaches, trying to game the system, and so on. But we do everything we can to prohibit that sort of illegal behavior from those sorts of organizations.

Mr. TAYLOR. Do you have screens set up to try to figure out who the users are, to try to, you know, pierce the veil, so to speak, into an anonymous account, figure out where that is or who that might be, where it is sourcing from? Are you looking at that? Is that something, a part of how you operate as an organization, that Google does?

Mr. SLATER. Absolutely, Congressman. We use a combination of automated systems, threat analysis to try and ferret out behaviors that may be indicative in that way.

Mr. TAYLOR. All right. Thank you. I appreciate your answers.

With that, Mr. Chairman—and I appreciate the panel for being here. This is an important, important topic, and thank you.

Thank you, Mr. Chairman.

Chairman THOMPSON. Thank you very much.

The Chair now recognizes the gentlelady from Nevada, Ms. Titus, for 5 minutes.

Ms. TITUS. Thank you, Mr. Chairman.

We have heard a lot about incidents, but we haven't mentioned much about one that occurred in my district of Las Vegas. This was the deadliest shooting in the United States in modern history. October 1, 2017, a gunman opened fire on a music concert, a festival. After that attack, there was a large volume of hoaxes, conspiracy theories, misinformation that popped up all across your platforms, including about the misidentity of the gunman, his religious affiliation, and some of the fake missing victims. Some individuals even called it a false flag.

In addition, when you put up a search Safety Check site on Facebook, where loved ones could check in to see who was safe and who wasn't, there were all kind of things that popped up, like links to spam websites that solicited Bitcoin donations, they pedaled false information, claiming that the shooter was associated with some anti-Trump army—just a lot of mess there, where people were trying to make contact.

I wonder if you have any specific policy or protocols or algorithms to deal with the immediate aftermath of a mass shooting like this. All three of you.

Ms. BICKERT. Thank you, Congresswoman.

Let me say that the Las Vegas attack was a horrible tragedy. We think we have improved since then, but I want to explain what our policies were even then and how we have gotten better.

So, with the Las Vegas attack, we remove any information that is praising that attack or the shooter, and we also took steps to protect the accounts of the victims. Sometimes in the aftermath of these things, we will see people try to hack into accounts or do other things like that, so we take steps to protect the victims. We also worked very closely with law enforcement.

Since then, one area where we have gotten better is crisis response in the wake of a violent tragedy. So, for instance, with Christchurch, you had these companies at the table and others communicating real-time, sharing with one another URLs, new versions of the video of the attack, and so forth to make sure—and it was literally a real-time, for the first 24 hours, operation where we were sharing. In that first 24 hours, on Facebook alone, we were able to stop 1.2 million versions of the video from hitting our site.

So we have gotten a lot better technically, but this is an area where we will continue to invest.

Mr. PICKLES. Thank you.

As you have just heard, I think one of the challenges we have in this space is different actors will change their behavior to try and get around our rules.

One of the things that we saw after Christchurch which was concerning was people uploading content to prove the event had happened. So the suggestion that because companies like us were removing content at scale, people were calling that censorship, so there were people uploading content to prove the attack had happened.

That is a challenge that we haven't had to deal with before, and it is something we are very mindful of. We need to figure out what is the best way to combat that challenge.

We have policies against the abuse and harassment of the survivors and victims and their families. So if someone is targeting someone who has been a victim or a survivor and is denying the event took place or is harassing them because of another factor, like political ideology, we would take action for the harassment in that space.

Then, finally, the question of how we work with organizations to spread the positive message going forward. So that is where, you know, if there are groups in your communities who are affected by this and working with the victims to show the kind of positivity of your community, then we would be keen to work with those organizations, wherever they are in the United States, to spread that message of positivity.

Ms. TITUS. Mr. Slater.

Mr. SLATER. Yes. Thank you, Congresswoman. This is of the utmost seriousness. It was a tragic event, I think, for our country, for society. Personally, as someone who lived in both Las Vegas and New Zealand, both of these events I hold deeply in my heart.

We take a threefold approach to the sort of misinformation and other conduct that you were talking about:

We try and, on YouTube, raise up authoritative sources of information, particularly in those breaking news events, to make sure that authoritative sources outpace those who might wish to misinform and so on.

We will strike, remove denials of well-documented violent events or people who are spreading hate speech toward the survivors of that event.

We will also seek to reduce exposure to content that is harmful misinformation, including conspiracies and the like.

Ms. TITUS. Well, these people have already been victimized in the worst sort of way. You hate to see them then become victims of something that occurs over the internet.

One thing we heard from law enforcement was that you might think about—and I think this relates to kind-of what you were saying, Mr. Slater—using your algorithms to elevate posts that come from law enforcement, so people seeking help go to those first as opposed to some of this other information just that comes in randomly.

In your work with law enforcement, have you considered that? I know you were addressing the chief's questions earlier. Ms. Bickert.

Ms. BICKERT. Thank you, Congresswoman.

That is something that we can explore with law enforcement. We certainly try to make sure that people have accurate information after attacks. Our systems didn't work the way we wanted them to after Las Vegas. We learned from that, and I think we are in a better place today.

Ms. TITUS. I would appreciate it if you would look into that. I think law enforcement would too.

Thank you, Mr. Chairman.

Chairman THOMPSON. Thank you.

The Chair recognizes the gentleman from Mississippi for 5 minutes.

Mr. GUEST. Thank you, Mr. Chairman.

First of all, to our representatives from Facebook, Google, and Twitter, I want to thank you for being here today. I want to thank you for previously appearing for a closed briefing that we had earlier this year.

So we seek to continue to examine this complex issue of balancing First Amendment rights against making sure that content that is on social media does not promote terroristic activity.

Professor Strossen, you were not here during that closed briefing, so I want to ask a couple questions to you.

During your testimony, your written testimony, you highlight the potential dangers associated with content moderation, even when done by private companies in accordance with their First Amendment rights. You make a case for social media companies to provide free-speech protections to users.

You even state in the conclusion of your written testimony—you say, "How to effectively counter the serious potential adverse impact of terror content and misinformation is certainly a complex problem. While restricting such expressions might appear to be a clear, simple solution, it is, in fact, neither, and, moreover, it is wrong."

Now, I know that was the conclusion of an 11-page report that you provided, but could you just briefly summarize that for the purpose of this hearing?

Ms. STROSSEN. Thank you so much, Congressman Guest.

Yes, the problem is the inherent subjectivity of these standards. No matter how much you articulate them—and I think it is wonderful that Facebook and the other companies have now, fairly recently, shared their standards with us—you can see that it is impossible to apply them consistently to any particular content.

Reasonable people will disagree. The concept of "hate," the concept of "terror," the concept of "misinformation" are strongly debated. One person's fake news is somebody else's cherished truth.

Now, a lot of attention has been given to the reports about discrimination against conservative viewpoints in how these policies are implemented. I want to point out that there also have been a lot of complaints from progressives and civil rights activists and social justice activists complaining that their speech is being suppressed.

What I am saying is that, no matter how good the intentions are, no matter who is enforcing it, whether it be a Government authority or whether it be a private company, there is going to be, at best, unpredictable and arbitrary enforcement and, at worst, discriminatory enforcement.

Mr. GUEST. Let me ask you, as an expert in the First Amendment, do you feel that content moderation by social media companies has gone too far?

Ms. STROSSEN. I think that, you know, first of all, they have a First Amendment right. I think that is really important to stress.

But given the enormous power of these platforms—which, as the Supreme Court said in a unanimous decision 2 years ago, that this is now the most important forum for the exchange of information and ideas, including with elected officials, those who should be accountable to we, the people.

So if we do not have free and unfettered exchange of ideas on these platforms for all practical purposes, we don't have it. That is a threat to our democratic republic as well as it is to individual liberty.

There is a lot that these platforms can do in terms of user empowerment so that we can make our own choices about what to see and what not to see and, also, information that will help us evaluate the credibility of the information that is being put out there.

Mr. GUEST. Finally, Ms. Strossen, do you have any recommendations that you feel would help balance individuals' First Amendment rights versus trying to protect social media from terrorists being able to use that as a platform that you would recommend, first, to the social media companies? Then are there any recommendations that you would have of this body, things that Congress should consider, that would help us as we navigate this very difficult situation?

Ms. STROSSEN. I think that Congress's oversight, as you are exercising very vigorously, is extremely important. I think encouraging, but not requiring, the companies to be respectful of all of the concerns—human-rights concerns of fairness and transparency and due process, as well as free speech, but also concerns about potential terrorism and dangerous speech.

I actually think that the U.S. Supreme Court and international human rights norms, which largely overlap, have gotten it right. They restrict discretion to enforce standards by insisting that, before speech can be punished or suppressed, that there has to be a specific and direct, tight causal connection between the speech in that particular context which causes an imminent danger.

We can never look at words alone in isolation, to get back to the question that I was asked by the Congresswoman, because you have to look at context. If in a particular context there is a true threat, there is intentional incitement of imminent violence, there is material support of terrorism, there is defamatory statements, there is fraudulent statements, all of that can be punished by the Government, and, therefore, those standards should be enforced by social media as well. That would give us—in my view, that is exactly the right way to strike the balance here.

Mr. GUEST. Thank you, Mr. Chairman. I yield back.

Chairman THOMPSON. Thank you very much.

The Chair recognizes the gentleman from Missouri, Reverend Cleaver, for 5 minutes.

Mr. CLEAVER. Thank you, Mr. Chairman.

I am going to have a little different approach than my colleagues.

Ms. Strossen, in 1989, I was a member of the city council in Kansas City, and the Klan had planned a big march in Swope Park. All this is still on-line; you can look at it. I fought against them. The ACLU supported their right to march and that, if I had passed an ordinance—I was also vice mayor at the time—if I passed an ordinance, they would challenge it in court.

I am not mad. I am not upset. I am a former board member of the ACLU. So I think that free speech has to be practiced even when it is abhorrent.

Now, for everybody else—and, in some ways, I kind-of feel sorry for you, not enough to let you out without, you know, beating up on you a little bit. But, you know, we—I am afraid for our country. I mean, we have entered an age where people respect an alternative truth. It is just so painful for me to watch it. I don't think I am watching it in isolation. Alternative truths, where people just will say something that is not true and continue to say it. It doesn't matter.

I saw it last night, where the President said, "Barack Obama started this border policy, and I am correcting it." What they did—and this is what I want you to consider. What one of the TV networks did is put up people making statements about what was happening. They showed Jeff Sessions, when he had first announced the separation policy and so forth.

You know, the problem is that—Churchill said that a lie can travel halfway around the world before the truth puts on its shoes. That is true. If we started a 20th-Century new bible, that should be one of the scriptures, because it is a fact. The truth cannot always be uncontaminated with sprinkles of deceit.

So you guys have a tough job. I don't want to make it seem like it is something that you can do easily.

Our system of government, I think, even beyond that, our moral connections are dependent a lot more—and I didn't realize this. I spent 3½ years in the seminary. I didn't even realize this until recently. But we depend significantly on shame. I mean, there are some things that laws can't touch, and so our society functions on shame. So, when shame is dismembered, I am not sure what else we have left.

But what I would like for you to react to and maybe even consider is, you know, instead of taking something down in some instances, why not just put up the truth next to it? I mean, the truth. I am not talking about somebody else's response. I am talking about the truth, where you, you know, like the video—I wish I could have brought it to you, where they said, here is the lie, and here is the truth.

Can anybody help me?

OK. All right.

Anybody else?

Mr. SLATER. So this is a very important issue, Congressman. And——

Mr. CLEAVER. Yes, that is why—yes.

Mr. SLATER. Absolutely. So one of the things we have been trying to do is two-fold with respect to harmful misinformation.

So one is, where there is a video that says, say, the moon landing didn't happen——

Mr. CLEAVER. My grandmother says that.

Mr. SLATER [continuing]. Or the Earth is flat, the video may be up, but you will see a box underneath it that says, here is a link to the Wikipedia page about the moon landing, or the Encyclopedia Britannica page, where you can go learn more. I think that speaks to——

Mr. CLEAVER. Yes, that is what I am talking about. Yes.

Mr. SLATER [continuing]. The sort of feature that you are talking about.

The other thing we try and do——

Mr. CLEAVER. You do that now?

Mr. SLATER. We do that today, yes, sir.

The other thing we try and do is reduce the exposure of the frequency of the recommendations to information that might be harmful misinformation, such as those sorts of conspiracies.

Mr. CLEAVER. Thank you.

Mr. PICKLES. Well, I think you rightly highlighted the interplay between what is on social media companies, the news media, what is on TV. How that cycle of information works together is a critical part of solving this.

I think the one thing that, for Twitter, because we are a public platform, very, very quickly people are able to challenge, to expose, to say, "That is not true. Here is the evidence. Here is the data."

There is something incredibly important about these conversations taking place in public. That, I think, is something, as we move into the information century, we need to bear in mind.

Mr. CLEAVER. Thank you.

Ms. BICKERT. Congressman, thank you.

Similar to what my colleague referenced, if there is something like misinformation that a third-party fact-checking organization has debunked—and we work with 45 of these organizations worldwide. They all meet objective criteria; they are all Poynter-certified. What we do is we actually take the articles from those fact-checkers and put it right next to the false content so that people have that context. If you go to share some of that content, we say, "This content has been rated false by a fact-checker," and we link them to it.

Similarly, when it comes to things like misinformation about vaccines, we are working with organizations like the CDC and the World Health Organization to get content from them that we can actually put next to vaccine-related misinformation on our site.

We do think this is a really important approach. It obviously takes a lot of resources.

Another thing we are trying to do is—I guess what I would say is empower those—and this is similar to what Mr. Pickles mentioned—empower those who have the best voices to reach the right audience on this. So we invest heavily in promoting counterspeech and truthful speech.

Mr. CLEAVER. Thank you.

Thank you, Mr. Chairman.

Chairman THOMPSON. Thank you very much.

Before we close, I would like to insert into the record a number of documents.

The first is several letters from stakeholders, addressed to Facebook as well as Twitter and YouTube, about hateful content on their platform.

The second is a joint report from the Center for European Policy Studies and the Counter Extremism Project.*

The third is a statement for the record from the Anti-Defamation League.

*The information has been retained in committee files and is available at *https://www.ceps.eu/ceps-publications/germanys-netzdg-key-test-combatting-online-hate/*.

The fourth are copies of community standards as of this day for Facebook, Twitter, and Google.[**]

Without objection, so ordered.

[The information referred to follows:]

October 30, 2017.

Mr. Mark Zuckerberg, *Chief Executive Officer,*
Ms. Sheryl Sandberg, *Chief Operating Officer,*
Facebook, Inc., 1 Hacker Way, Menlo Park, CA 94025.

DEAR MR. ZUCKERBERG AND MS. SANDBERG:

We, the undersigned civil rights, interfaith, and advocacy organizations write to express our deep concern regarding ads, pages, and hateful contenton your platform used to divide our country, and in particular, to promote anti-Muslim, anti-Black, anti-immigrant, and anti-LGBTQ animus. We thank you for recent meetings with some of our organizations representing communities that were directly affected by the material on your platform. We appreciate that senior members of your team—including you, Ms. Sandberg—have facilitated these meetings, and we hope that these conversations are the beginning of a serious and ongoing dialog. Now, it is necessary for Facebook to take critical steps to address the bigotry and discrimination generated on your platform.

As you know, we do not yet have access to all the divisive content targeting communities we represent; therefore, we are only able to cite to the few examples that were leaked to the media.

For example, Russian operatives set up misleading accounts impersonating or posing as American individuals and groups on Facebook to promote Russian propaganda during the American election season. Reports indicate that a Russian Facebook account called "Secured Borders" posed as a group of US citizens concerned about the increased number of refugees in America. This fake account not only promoted anti-immigrant messaging online, but also managed to organize an in-person anti-refugee rally in Twin Falls, Idaho in August 2016.[1]

In addition, a Facebook page entitled "United Muslims of America" was an imposter account traced back to Russia[2]—the real United Muslims of America is a California-based interfaith organization working at the local level to promote dialog and political participation.[3] The imposter account smeared political candidates and promoted political rallies aimed at Muslim audiences.[4] In another example, the Internet Research Agency in Russia promoted an anti-Muslim rally thousands of miles away in Houston, Texas where individuals protested outside of a mosque.[5] Additional reports indicate that Facebook offered its expertise to a bigoted advocacy group by creating a case study testing different video formats, and advising on how to enhance the reach of the group's anti-refugee campaign in swing States during the final weeks of the 2016 election.[6] These examples of content on Facebook were not only harmful, but also used to rile up supporters of President Trump.

Furthermore, it has been reported that Russian operatives purchased Facebook ads about Black Lives Matter—some impersonating the group and others describing it as a threat.[7] This included ads that were directly targeted to reach audiences in Ferguson, Missouri and Baltimore, Maryland. CNN reports that the Russian Internet Research Agency used these ads in an attempt to amplify political discord and

[**] The information has been retained in committee files and is available at *https://www.facebook.com/communitystandards/*, *https://help.twitter.com/en/rules-and-policies/twitter-rules*, and *https://www.youtube.com/about/policies/#community-guidelines*, respectively.

[1] Geoffrey Smith, "Russia Orchestrated Anti-Immigrant Rallies in the U.S. via Facebook Last Year," Fortune, Sept. 12, 2017, available at *http://fortune.com/2017/09/12/russia-orchestrated-anti-immigrant-rallies-in-the-u-s-via-facebook-last-year/*.

[2] Dean Obeidallah, "How Russian Hackers Used My Face to Sabotage Our Politics and Elect Trump," The Daily Beast, Sept. 27, 2017, available at *https://www.thedailybeast.com/how-russian-hackers-used-my-face-to-sabotage-our-politics-and-elect-trump*.

[3] United Muslims of America "About" page, available at *http://www.umanet.org/about-us*.

[4] Obeiallah, supra note 1.

[5] Tim Lister & Clare Sebastian,"Stoking Islamophobia and secession in Texas—from an office in Russia," CNNPolitics, Oct. 6, 2017, available at *http://www.cnn.com/2017/10/05/politics/heart-of-texas-russia-event/index.html*.

[6] Melanie Ehrenkranz, "Facebook Reportedly Used Anti-Muslim Ad as Test Case in Video Formats," Gizmodo, Oct. 18, 2017, available at *https://gizmodo.com/facebook-reportedly-used-anti-muslim-ad-as-test-case-in-1819645900*.

[7] Adam Entous, Craig Timberg, & Elizabeth Dwoskin,"Russian operatives used Facebook ads to exploit America's racial and religious divisions," The Washington Post, Sept. 25, 2017, available at *https://www.washingtonpost.com/business/technology/russian-operatives-used-facebook-ads-to-exploit-divisions-over-black-political-activism-and-muslims/2017/09/25/4a011242-a21b-11e7-ade1-76d061d56efa__story.html?tid=sm__tw&utm__term=.e49cecc1a834*.

create a general atmosphere of incivility and chaos.[8] This included a fake ad containing an image of an African-American woman dry-firing a rifle, playing on the worst stereotypes regarding African-Americans as threatening or violent.[9]

We were alarmed to see your platform being abused to promote bigotry, and especially disappointed that it has taken media exposure and Congressional oversight to give a degree of transparency into your practices. It is important to keep in mind that pervasive bigotry has long existed on your platform, and the Russian operatives simply exploited the hateful content and activity already present. We are concerned about how a platform like Facebook's could operate without appropriate safeguards that take into account how it could be manipulated to further sow divisions in our society.

As a company and social network platform whose mission is "to give people the power to build community and bring the world closer together,"[10] we hope that you understand the gravity of this hateful rhetoric and behavior. During a time when anti-Muslim, anti-Black, anti-LGBTQ, and anti-immigrant sentiment has swept the nation, it is more important than ever for companies like yours to take an unequivocal stance against bigotry.

Over the years, many of us have raised concerns about how your platform may have a negative impact on our communities, with disappointing results. For example, we have requested that you address attacks on African Americans and Muslims, organizing by hate groups, and the censorship of Black, Arab, Muslim, and other marginalized voices. As a result of the pervasive presence and organizing by hate groups on your platform—some could not exist as national level entities without it—we have repeatedly requested that you convene a gathering with civil rights organizations to discuss appropriate and strategic responses. While you were unable to sufficiently respond to the concerns raised above, Facebook participated in and organized events that stigmatized Muslims and other communities such as a recent convening called "Tech Against Terrorism."

Though in the past you have displayed a willingness to listen to our concerns, we have yet to see meaningful change. It is our hope that recent developments will mark a new chapter in Facebook's commitment to protecting the rights of all who use your platform.

As we continue this important dialog, we urge you to:

1. Fully disclose to the public all of the ads, pages, events, accounts, and posts you have traced back to Russian operatives targeting African American, LGBTQ, and Muslim communities. In particular, we believe that Facebook has a special responsibility to notify those individuals and organizations who have been impersonated or misrepresented.
2. Bring on an independent third-party team to conduct a thorough and public audit of the civil rights impact of your policies and programs, as well as how the platform has been used by hate groups, political entities, and others to stoke racial or religious resentment or violence. Other leading companies in the industry like Airbnb have made the decision to conduct such an assessment, and we hope you will follow their lead.
3. Regularly convene a new working group of a diverse group of civil rights organizations working to counter bigotry, and solicit input on policies and processes from this group. And, integrate addressing hate into Facebook's corporate structure by:
    a. Assigning a board committee with responsibility for assessing management efforts to stop hate groups, State actors, and individuals engaged in hate from using your platform and tools;
    b. Assigning a senior manager who is a member of Facebook's Executive Team with authority to oversee addressing hate company-wide and name that person publicly and employing staff with expertise in this area to vet advertisements and develop process and procedures the address this issue; and,
    c. Creating a committee of outside advisors with expertise in identifying and tracking hate who will be responsible for producing an annual report on the effectiveness of steps taken by Facebook.

[8] Dylan Byers, "Exclusive: Russian-bought Black Lives Matter ad on Facebook targeted Baltimore and Ferguson," CNN Media, Sept. 28, 2017, available at *http://money.cnn.com/2017/09/27/media/facebook-black-lives-matter-targeting/index.html.*

[9] Adam Entous, Craig Timberg, & Elizabeth Dwoskin, "Russian Facebook ads showed a black woman firing a rifle, amid efforts to stoke racial strife," The Washington Post, Oct. 2, 2017, available at *https://www.washingtonpost.com/business/technology/russian-facebook-ads-showed-a-black-woman-firing-a-rifle-amid-efforts-to-stoke-racial-strife/2017/10/02/e4e78312-a785-11e7-b3aa-c0e2e1d41e38__story.html?utm__term=.aa2267a2f46c.*

[10] Facebook "About" page, February 4, 2004, available at *https://www.facebook.com/pg/facebook/about/?ref=page__internal.*

4. Develop, with input from diverse civil rights groups and experts, and make public a clear process for how Facebook:

a. Reviews content constituting hate speech;

b. Reviews efforts to use Facebook as a platform to stoke identity-based, racial, or religious resentment or violent actions; and,

c. Responds to complaints about content that reasonably creates fear and chills speech on Facebook.

5. Make public detailed information regarding training and support for anti-immigrant, anti-Muslim, anti-black, and anti-LGBTQ organizations, including the monetary value of these services; and establish a fund to provide grants to organizations combating hatred and bigotry.

Thank you in advance for your consideration.

Please contact Naheed Qureshiat [sic] with any questions.

We look forward to your reply.

Sincerely,

Arab American Institute (AAI)
Asian Americans Advancing Justice/AAJC
Center for Media Justice
Center for New Community
Color of Change
CREDO
Human Rights Campaign (HRC)
The Leadership Conference on Civil and Human Rights
League of United Latin American Citizens (LULAC)
MoveOn.org
Muslim Advocates
NAACP
NAACP Legal Defense and Educational Fund, Inc. (LDF)
National Center for Lesbian Rights
National Hispanic Media Coalition
National LGBTQ Task Force
National Sikh Campaign
Sikh Coalition
Southern Poverty Law Center

February 22, 2018.

Ms. Monica Bickert,
Head of Product Policy and Counterterrorism, Facebook, 1 Hacker Way, Menlo Park, CA 94025.

Ms. Juniper Downs,
Director, Public Policy and Government Relations, YouTube, 901 Cherry Ave., San Bruno, CA 94066.

Mr. Carlos Monje, Jr.,
Director, Public Policy and Philanthropy, U.S. & Canada, Twitter, 1355 Market Street, San Francisco, CA 94103.

Ms. Bickert, Ms. Downs, and Mr. Monje: The undersigned civil rights and advocacy organizations write to share our concerns regarding your recent testimony at the United States Senate Committee on Commerce, Science, and Transportation hearing titled, “Terrorism and Social Media: #IsBigTechDoingEnough?” Many of the undersigned organizations have had on-going conversations with your companies regarding the spread of hateful and dangerous content online, and in light of this, we watched your testimony regarding extremist content online with great interest. We were alarmed by the continuous conflation of Muslims and violent extremism at the hearing and the extent to which testimony focused on conduct by Muslims, with comparatively almost no mention about violent actions by white supremacists who target members of the African American, LGBTQ, immigrant, Latino, Asian, Jewish, Sikh and Muslim communities. These omissions are particularly striking in light of the recent tragic attacks in New Mexico, Portland, and Charlottesville.

To no avail, several of the signatories below reached out to you prior to the hearing to request that your companies avoid stigmatizing and singling out the Muslim community by failing to address other forms of extremism in your testimony. All three of your statements for the record failed to do so; they referenced only violent extremism committed by those claiming to be acting in the name of Islam and highlighted efforts at countering extremism that focus on Muslim communities. Facebook’s written testimony, for example, did not mention white supremacist violence, but repeatedly cited ISIS and al-Qaeda.[1] And, in response to questioning from Senator John Thune (R–SD), the Facebook representative volunteered Boko Haram—another group claiming to act in the name of Islam—as an example of a group whose content has been banned by their company.[2] Later, when Senator Tom Udall (D–NM) directly asked the Facebook witness what the company is doing to curtail the explosion of white supremacists online, once again, Facebook failed to mention white supremacy in the response. In fact, in response to questioning regarding domestic extremism—specifically violence by white nationalists and white supremacists—the Google witness was the only panelist to specifically mention "white supremacy," albeit briefly.[3] It is striking that such complex questions seem to consistently elicit simple, and near-uniform answers.

Furthermore, it was very unhelpful that each of your companies chose to highlight your support or participation in violent extremism initiatives designed to target Muslims and others as examples of the work you are doing to fight extremism. For example, Twitter's testimony stated that the company has participated in more than 100 CVE trainings over the last few years including summits at the White House. We are concerned that most of these events were focused primarily on activities by Muslims. In addition, all three companies continue to emphasize their sponsorship of Tech Against Terrorism events, one of which, in the San Francisco Bay Area, focused exclusively on extremism by Muslims. Other Tech Against Terrorism events have given some attention to white supremacy, but not nearly enough and not on a par with the attention given to Muslims and extremism. In one example, the Southern Poverty Law Center (SPLC), one of our Nation's leading experts on hate groups and white supremacy, was invited to a Tech Against Terrorism conference in Brussels and given less than a week's notice of the event. When SPLC requested to participate via video conference due to the short notice, they received no response. If there is a true commitment by the companies to address white supremacy and other forms of violent extremism unrelated to Islam through this initiative, more lead time is necessary to appropriately engage relevant experts and stakeholders. Additionally, as recently as last week, presentations by Facebook, Google, and Twitter at an event organized by the Department of Homeland Security focused heavily on activities designed to address extremism by those claiming to act in the name of Islam.

At a time when anti-Muslim, anti-Black, anti-LGBTQ, anti-immigrant and anti-Jewish sentiment have fueled a marked increase in violent attacks on individuals in each of these communities, a responsible discussion regarding violent extremism must include a focus on white supremacists and other non-Muslim violent extremists. On far too many occasions, discussions about terror do not acknowledge that no ideology owns violent extremism. The failure to recognize white supremacy and other forms of violent extremism unrelated to Islam in discussions regarding extremism is irresponsible and reckless, and your failure to adequately address this publicly during the Senate hearing stigmatizes Muslims and other affected communities when the facts on this issue are clear. In their 2017 annual report on extremism in the United States, the Anti-Defamation League (ADL) concluded that the number of murders committed by white supremacists in the United States doubled from the previous year, nothing 71 percent of extremist-related murders in the past decade have been carried out by right-wing extremists, a majority of whom were born in the United States.[4] And in 2017, 53 percent of extremist-related murders in the United States were perpetrated by white supremacists.[5]

We have raised at least some of our concerns either with your parent companies, or with your companies directly. One recent example, is the letter sent on October 31st, 2017, by 19 civil rights groups to Facebook citing the company's inadequate response to hate speech and bigotry directed toward members of the African-American, LGBTQ, and Muslim community on its platform, as well as problematic CVE

[1] Bickert, M. (2018, January 17). Hearing before the United States Senate Committee on Commerce, Science, and Transportation. Retrieved January 22, 2018, from *https://www.commerce.senate.gov/public/__cache/files/a9daccb8-5f07-42a6-b4c3-20ad0b9ba26d/FC0A5B87F787273A7FA793B458C03E41.bickert-testimony-final.pdf*.

[2] Terrorism and Social Media: #IsBigTechDoingEnough. (2018, January 17). Retrieved January 22, 2018, from *https://www.c-span.org/video/?c4709695%2Fbickert-response*.

[3] Terrorism and Social Media: #IsBigTechDoingEnough. (2018, January 17). Retrieved January 22, 2018, from *https://www.c-span.org/video/?c4709693%2Fms-bickert-response*.

[4] ADL Report: White Supremacist Murders More Than Doubled in 2017. (2018, January 17). Retrieved January 22, 2018, from *https://www.adl.org/news/press-releases/adl-report-white-supremacist-murders-more-than-doubled-in-2017*.

[5] Williams, J. (2017, October 02). White American men are a bigger domestic terrorist threat than Muslim foreigners. Retrieved January 22, 2018, from *https://www.vox.com/world/2017/10/2/16396612/las-vegas-mass-shooting-terrorism-islam*.

activities.[6] Given your companies' size, influence, and role in all discussions of hateful and violent content on-line, we again call on you to join us in a comprehensive and inclusive dialog on extremism and extremist violence.

As we continue this important dialog, we urge each of your companies to:

- Submit amended testimony for the hearing regarding the dangers posed by white supremacist groups and the measures your organization will be taking as a result;
- Bring on an independent third-party team to conduct a thorough and public audit of the civil rights impact of your policies and programs, including an assessment of processes related to addressing extremism by white supremacists and other hate-based content on your platforms that encourages harassment and violence towards many communities;
- Assign and publicly name a senior member of the executive team with authority to oversee addressing hate on the platform company-wide;
- Hire or contract with a diverse team of experts on white supremacist groups to develop methods for detecting and responding to such groups, and to address hateful conduct and content by these groups;
- Create a committee of outside advisors with expertise in identifying and tracking hate who will be responsible for producing an annual and publicly available report on the effectiveness of the steps taken by the company; and,
- Disclose publicly any new plans that have been developed to address extremism, including whether those plans will target Muslims or seriously address white supremacists.

Thank you for your consideration of our views. We look forward to hearing from you.

Sincerely,

Arab American Institute
Bend the Arc Jewish Action
Center for Media Justice
Color of Change
CREDO
Emgage
Media Matters for America
MoveOn.Org
Muslim Advocates
NAACP
NAACP Legal Defense and Educational Fund, Inc. (LDF)
National LGBTQ Task Force
National Sikh Campaign
Sikh Coalition
Southern Poverty Law Center

December 18, 2018.

Mark Zuckerberg,
Chairman and Chief Executive Officer, Facebook, 1 Hacker Way, Menlo Park, CA 94025.

Dear Mr. Zuckerberg: We write to express our profound disappointment regarding Facebook's role in generating bigotry and hatred toward vulnerable communities and civil rights organizations. For years, many of us have engaged directly with your company in good faith, seeking change from within the company that we hoped would address a range of civil rights, privacy, and safety problems resulting from abuse and mismanagement of the platform, including engaging in an on-going audit of the civil rights impact of your policies and programs, as well as how the platform has been used by hate groups, political entities, and others to stoke racial or religious resentment or violence. In particular, we asked you to take immediate action to stop abuse of the platform. Recent news demonstrates, however, that Facebook was not only looking the other way in response to our concerns, but also has been actively working to undermine efforts by those who seek to hold the company responsible for abuses on the platform.

As you know, a recent investigation by the *New York Times*[1] details information about Facebook's responses to a series of crises—including crises around how the

[6] Simpson, S. (2017, October 31). Civil Rights Groups Urge Facebook to Address Longstanding Issues with Hate Speech and Bigotry. Retrieved January 22, 2018, from *https://www.muslimadvocates.org/19civilrightsgroupslettertofacebook/*.

[1] "Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis," *The New York Times,* November 14, 2018.

company manages and responds to hateful content. In the face of clear evidence that Facebook was being used to broadcast viral propaganda and inspire deadly bigoted campaigns, the company's leadership consistently either looked the other way, or actively worked to lobby against meaningful regulation, shifted public opinion against its allies, and personally attacked its critics.

Though Facebook has had significant time, opportunity and the benefit of input from experts and advocacy groups to address the problems on the platform, your company chose to target civil rights groups and our allies instead of changing the way you do business. Compounding this mistake, you retained the services of Definers Public Affairs to investigate, undermine, and attack our allies, mimicking the tactics of the worst, disreputable political operatives and hate groups. Out of your need to treat those leveling legitimate critiques against Facebook as your enemies, you jeopardized the safety and security of people who have dedicated their lives to the common good. This decision crossed all lines of common decency.

Furthermore, it's an absolute disgrace that Facebook sought to deflect criticism and discredit advocates by exploiting anti-Semitic campaigns against philanthropist George Soros. A research document circulated by Definers wrongfully identified Mr. Soros as the force behind a broad anti-Facebook movement. According to the *Times,* Definers urged reporters to explore the financial connections between Mr. Soros's family or philanthropy and progressive groups hoping to somehow use this information to undercut advocates pursuing accountability for bigotry on the platform. Unbelievably, Facebook sought to have their cake and eat it too; while you weaponized anti-Semitism directed at Mr. Soros, you attacked legitimate criticism of the company as anti-Semitic.

Equally troubling are your claims over the years that problems with the platform or the company's approach have been inadvertent, and that, per a statement quoted in the article, "our entire management team has been focused on tackling the issues we face." What is now clear, however, is direct evidence of malicious and calculated campaigns to undermine Facebook's critics.

Your response as the company's chairman and CEO was also disconcerting. You plead ignorance, that you had no idea that this was happening. But the public has given your company the benefit of the doubt for far too long and ignorance is no longer an excuse. It's become abundantly clear that, as currently constituted, your leadership team is unable to adequately address the valid concerns of the civil rights community. It is now time for significant changes in, not only your policies, but also your leadership structure. At this time, we demand that Facebook immediately:

1. Reorganize Facebook's board in order to enable greater accountability of the leadership team and to allow more diverse voices at the decision-making table. Specifically:
   a. You, Mr. Zuckerberg, should step down as chairman of the board as long as you serve as the chief executive officer to allow the board to provide independent oversight and guidance for the management team.
   b. Sheryl Sandberg should step down from the board of directors as long as she serves as chief operating officer in order to allow the board to provide independent oversight and guidance for the management team.
   c. Facebook should expand its board of directors by at least three members to diversify the board; these new members should reflect the diversity of your global community of users.
   d. The board should appoint an independent and permanent civil rights ombudsman to conduct consistent and on-going reviews of the civil rights implications of Facebook's policies and practices; this ombudsman shall also serve as a member of the board of directors.
2. Publicly identify and apologize to all organizations targeted by Definers Public Affairs. In the spirit of transparency, release all internal documents pertaining to opposition research generated by Definers, including all research on civil rights and advocacy organizations.
3. Remove Facebook's Vice President of Global Public Policy, Joel Kaplan, from his position.
4. Make public all findings and recommendations of the civil rights audit without revisions or redactions by January 31, 2019.

Thank you in advance for your consideration. We would like to meet with you to discuss our concerns and recommendations. Please contact Naheed Qureshi of Muslim Advocates at [sic] with any questions and to coordinate a meeting.

Sincerely,

Muslim Advocates
Arab American Institute
Asian Americans Advancing Justice—Atlanta
Bend the Arc Jewish Action
Center for Human Technology
Center for Media Justice
Community Responders Network
CreaTV San Jose
CREDO
Emgage
Equality Labs
Freedom From Facebook
HOPE not hate
Interfaith Center on Corporate Responsibility
MCN—Muslim Community Network
Media Matters for America
Million Hoodies Movement for Justice
MomsRising
MoveOn
MPower Change
Muslim Youth Collective
NAACP
National LGBTQ Task Force
National Network for Arab American Communities/The Campaign to TAKE ON HATE
South Asian Americans Leading Together (SAALT)
Southern Poverty Law Center
The Sikh Coalition
UltraViolet
United We Dream
Urbana-Champaign Independent Media Center
Voting Rights Forward
Women's Alliance for Theology, Ethics, and Ritual (WATER)

cc: Sheryl Sandberg, Marc Andreessen, Erskine B. Bowles, Kenneth I. Chenault, Susan Desmond-Hellmann, Reed Hastings, Peter A. Thiel, Jeffrey Zients.

Statement of the ADL (Anti-Defamation League)

June 26, 2019

ABOUT ADL

Since 1913, the mission of the Anti-Defamation League (ADL) has been to "stop the defamation of the Jewish people and to secure justice and fair treatment for all." For decades, ADL has fought against bigotry and anti-Semitism by exposing extremist groups and individuals who spread hate and incite violence.

Today, ADL is the foremost non-governmental authority on domestic terrorism, extremism, hate groups, and hate crimes. Through our Center on Extremism (COE), whose experts monitor a variety of extremist and terrorist movements, ADL plays a leading role in exposing extremist movements and activities, while helping communities and Government agencies alike in combating them. ADL's team of experts—analysts, investigators, researchers, and linguists—use cutting-edge technologies and age-old investigative techniques to track and disrupt extremists and extremist movements world-wide. ADL provides law enforcement officials and the public with extensive resources, such as analytic reports on extremist trends and Hate Symbols[1] and Terror Symbols databases. Through our Center for Technology and Society (CTS), ADL serves as a resource to tech platforms, civil society organizations and government, and develops proactive solutions to the problems of cyber hate, on-line harassment, and misuses of technology. Launched in 2017 and

[1] ADL, Hate on Display™ Hate Symbols Database, available at *https://www.adl.org/hatesymbolsdatabase.*

headquartered in Silicon Valley, CTS aims for global impacts and applications in an increasingly borderless space. It is a force for innovation, producing cutting-edge research to enable on-line civility, protect vulnerable populations, support digital citizenship and engage youth. CTS builds on ADL's century of experience building a world without hate and supplies the tools to make that a possibility both on-line and off-line.

On October 27, 2018, Robert Bowers perpetrated the deadliest attack against Jews in American history when he stormed a Pittsburgh synagogue armed with an assault rifle and three handguns.[2] Shouting "All Jews must die," Bowers killed 11 people in their place of worship. Less than 5 months later, Brenton Tarrant perpetrated the deadliest attack against Muslims in New Zealand's history, slaughtering 50 people who had gathered for prayer at two mosques.[3] A little over a month later, John Earnest attacked the Jewish community at a synagogue in Poway, California killing 1 congregant and injuring 3 others.[4] In the wake of these horrific crimes, Jewish and Muslim communities world-wide and concerned citizens across the globe began searching for clues about attacks that seemed to come out of nowhere.

In hindsight, however, these killings are wholly unsurprising, given that both attackers were enmeshed in on-line communities that exposed them to content designed to teach and amplify hate and make them potentially violent. Bowers was an engaged and active member of a fringe on-line community called Gab, which, like similar on-line forums, is a bastion of hatred and bigotry.[5] Gab has seen a surge in racist and anti-Semitic postings since the 2016 Presidential election. Tarrant and Earnest, too, were part of a fringe on-line community called 8chan, one of the most notoriously hateful on-line communities on the internet.[6]

ADL has been researching white supremacy and other forms of hate on-line. We have been working to identify how to more effectively address this growing threat.

ADL AND ON-LINE HATE

ADL has been working to combat on-line hate since 1985, with its "Computerized Networks of Hate" report which explored how dial-up computer bulletin boards served as a communications tool for white supremacists who have a modem and a home computer.

Since then, ADL has worked with the technology industry at each turn of its rapid expansion to help counter hate and extremism on-line. In the 1990's, ADL published reports on the state of on-line hate such as "The Web of Hate: Extremists Exploit the Internet"[7] and "Poisoning the Web: Hatred on-line." In 2012, ADL convened the "Anti-Cyberhate Working Group" which consisted of leading stakeholders from both technology companies in Silicon Valley as well as civil society discuss the burgeoning problem of hate on social media, and explored ways to mitigate this threat through policy. In 2014, inspired by this group, ADL released "Best Practices for Responding to Cyberhate," which was endorsed by leading tech companies and became a guidepost for the industry.[8]

ADL has continued to consult with technologists and policy makers on issues of on-line hate in the years following, and, in 2017 ADL launched the Center for Technology and Society (CTS).[9] CTS is the leading advocacy and research center headquartered in Silicon Valley focused on fighting hate and harassment on-line. Since its launch, CTS has contributed considerably to the advocacy on cyber hate spearheaded by ADL, convening the Cyberhate Problem Solving lab with Facebook,

[2] Jay Croft and Saeed Ahmed, "The Pittsburgh synagogue shooting is believed to be the deadliest attack on Jews in American history, the ADL says," CNN, October 28, 2018, available at *https://www-m.cnn.com/2018/10/27/us/jewish-hate-crimes-fbi/index.html.*

[3] ADL, "White Supremacist Terrorist Attack at Mosques in New Zealand," March 15, 2019, available at *https://www.adl.org/blog/white-supremacist-terrorist-attack-at-mosques-in-new-zealand.*

[4] *https://www.adl.org/blog/poway-attack-illustrates-danger-right-wing-extremists-pose-to-jews-muslims*.

[5] ADL, "Gab Was Down For a Week, Forcing Extremists to Consider Their Alternatives," November 5, 2018, available at *https://www.adl.org/blog/gab-was-down-for-a-week-forcing-extremists-to-consider-their-alternatives.*

[6] *https://www.adl.org/news/article/how-facebook-and-twitter-help-amplify-fringe-websites*.

[7] ADL, "The Web of Hate: Extremists Exploit the Internet," available at *https://www.adl.org/sites/default/files/documents/assets/pdf/combating-hate/ADL-Report-1996-Web-of-Hate-Extremists-exploit-the-Internet.pdf.*

[8] ADL, "Best Practices for Responding to Cyberhate," available at *https://www.adl.org/best-practices-for-responding-to-cyberhate*.

[9] *https://www.adl.org/news/press-releases/adl-to-build-silicon-valley-center-to-monitor-fight-cyberhate-omidyar-network-2*.

Microsoft, Twitter, and Google.[10] The lab includes managers at these companies from both the policy teams as well as the engineering teams that put policies into practice. CTS has significantly expanded on ADL's research work on on-line hate. This includes projects like the on-line Hate Index with UC Berkeley's D-Lab—a cutting-edge project that combines social science and machine learning techniques to develop a new way for AI to understand language and context in order to help identify and measure on-line hate speech.[11] CTS also worked with our Belfer Fellow, Samuel Woolley, to produce original research on how disinformation tactics were used to spread anti-Semitism on-line in advance of the 2018 midterm elections.[12] Moreover, in an effort to keep up with new forms of interactive digital technology, CTS collaborated with Implosion Labs to understand the potential for hate in the emerging ecosystem of social virtual reality.[13] CTS has also expanded its focus to fight hate, bias, and harassment in video games. CTS has worked with our Belfer Fellow Dr. Karen Schrier on ways in which games can foster empathy and reduce bias, developed a guide for game developers to explore issues of identity through game design,[14] and recently released a white paper exploring research at the intersection of identity, bias, empathy, and game design.[15]

In February 2019, CTS released a survey report that focused on the American experience of on-line hate and harassment.[16] The report considered how people were harassed on-line, which individuals and groups were targeted, the serious and lasting impact and effect on targets' lives, and how respondents want Government and the tech industry to address this pervasive and important issue.

INTERNAL PROCESSES: MANAGING HATE ON A SOCIAL MEDIA PLATFORM

A significant number of Americans use mainstream social media platforms as a part of their day-to-day life. As we have outlined, the terrorist attack in Christchurch highlighted the role that mainstream social media companies play in amplifying the spread of violent, extreme, and hateful content on-line. It is clear that the public, Government, and civil society lack important knowledge about social media platforms' ability and responsibility to detect and decrease violent, extremist, and hateful content.

In the wake of this horrific tragedy and others that have involved social media, we know more attention needs to be paid to the following issues: The process by which mainstream social media platforms manage violent and hateful content; the limitations hindering our ability to understand and evaluate platforms' efforts to decrease the prevalence of hate on-line; the weaknesses on each platform that allow for hateful, extreme, and violent content to reach users despite efforts by the platforms; and the need for more information and transparency regarding how effective tech companies' current practices are in countering hate, violence, and extremism on their platforms.

When we refer to "mainstream social media companies," we are primarily referring to Facebook (which owns Instagram), Google (which owns YouTube), Twitter, and Snapchat. These American companies own and operate platforms that have the largest number of monthly active users.[17]

Mainstream social media platforms are not bound by the laws of a particular country. Instead, when moderating hate, violence and extremism, each of the mainstream social media platforms is governed by two sets of rules. First is the individual platform's forward-facing rules, often called "community guidelines." Second is an internal, more expansive and granular set of unnamed rules, used to review and analyze content through a process called "content moderation."[18] These content moderation guidelines are typically developed, maintained, reviewed, and revised by the Policy, Security or Trust and Safety teams at a tech company.

[10] *https://www.adl.org/news/press-releases/facebook-google-microsoft-twitter-and-adl-announce-lab-to-engineer-new.*

[11] *https://www.adl.org/resources/reports/the-on-line-hate-index.*

[12] *https://www.adl.org/resources/reports/computational-propaganda-jewish-americans-and-the-2018-midterms-the-amplification.*

[13] *https://www.adl.org/resources/reports/hate-in-social-virtual-reality.*

[14] *https://www.adl.org/media/12529/download.*

[15] *https://www.adl.org/designing-ourselves.*

[16] *https://www.adl.org/on-lineharassment.*

[17] Aaron Smith and Monica Anderson, "Social Media Use in 2018," Pew Research Center, March 1, 2018, available at *https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.*

[18] Kate Klonick, "The New Governors: The People, Rules, And Processes Governing on-line Speech," available at *https://docs.house.gov/meetings/IF/IF00/20180905/108642/HHRG-115-IF00-20180905-SD011.pdf.*

The internal, more expansive rules governing content moderation are enforced continuously by staff or contractors on an operations team. In April 2018, Facebook became the first tech company to state that they were publicly releasing their internal community guidelines; however, this claim is unverifiable.[19]

Historically, the majority of content reviewed through the content moderation process is reported to the platform by its users. If the content moderation process is predominantly reactive—meaning problematic activity is only flagged once it is reported to the platform by users—the burden is placed entirely on users and the platform is merely providing customer service in addressing—and selectively at that—user reports of hateful content. (User flagging of problematic content has also been employed to address other types of problematic content, such as copyright violations.) In the mean time, as a result of their business models and algorithms many of the larger platforms continued to monetize and promote harmful content in search of increasing user engagement. Ultimately, this model allowed platforms to de-prioritize addressing the existence of hateful content on their platforms.

Notably, when mandated by law or when trying to avoid Government regulation, tech companies have shown the ability to coordinate and take proactive measures to moderate certain kinds of objectionable content. For example, in the areas of child pornography and international terrorism, the mainstream tech platforms have worked together—using technology that allows them to tag and catalog certain images and videos and coordinate across platforms—to proactively remove problematic content.[20] A recent working paper questioned the efficacy of such efforts following the events of Christchurch.[21] Nevertheless, tech companies have shown meaningful success in terms of mitigating ISIS-related terrorism content.[22]

It is worth noting that the proliferation of harmful content and the ineffectiveness to date of the tech companies' responses have led to calls to weaken or eliminate Section 230 of the Communications Decency Act of 1996. That is the law that protects tech platforms from being liable for content posted by users—so called user-generated content. This law is the fundamental bedrock for much of what has been transformative for good in the development of an open internet promoting free speech, community, access to knowledge, education, and creativity. For example, Section 230 enabled platforms like Wikipedia, the on-line encyclopedia, to be created and to thrive. Without the protections of Section 230 many innovations and smaller companies, including not-for-profit sites like Wikipedia, likely could not exist. So ADL is not calling for the elimination or "swiss-cheesing" of Section 230 protections.

At the same time, immunity from liability for user-generated content—along with a dominant business model that monetizes engagement (and often harmful but impactful content)—as well as the lack of other regulations or meaningful self-governance helps foster a purely reactive culture among large social media platforms. That places the onus on users to bring problematic content to the attention of the companies. And as we now know, that model failed egregiously to find and mitigate harmful content and did not adequately protect our democracy from manipulation.

However, one-size-fits-all-regulation concerning content moderation will have unintended consequences—including removing extremist and unlawful content to places where it cannot easily be found and preempted or prosecuted by law enforcement. It will have serious potential unintended consequences. In addition, it will almost certainly make it very expensive to comply with internet regulations and thus lead to the ironic effect of consolidating monopoly market positions of the very tech giants whose behavior has rightly concerned Congress, since few companies would be able to afford to comply with regulations or defend against countless lawsuits based on user content.

Turning back to content moderation as it works when the onus is on users, once a user (or a group like ADL) has reported a piece of content, the process by which a company decides whether that piece of content violates the platform's guidelines,

[19] Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Facebook Newsroom, April 24, 2018, available at *https://newsroom.fb.com/news/2018/04/comprehensivecommunity-standards/*.

[20] 9 [sic] Kaveh Waddell, "A Tool to Delete Beheading Videos Before They Even Appear online," The Atlantic, June 22, 2016, available at *https://www.theatlantic.com/technology/archive/2016/06/a-tool-to-delete-beheading-videosbefore-they-even-appear-on-line/488105/*.

[21] *https://www.ivir.nl/publicaties/download/Hash__sharing__Heller__April__2019.pdf*.

[22] Nitasha Tiku, "Tech Platforms Treat White Nationalism Different From Islamic Terrorism," WIRED, March 20, 2019, available at *https://www.wired.com/story/why-tech-platforms-dont-treat-all-terrorism-same/?utm__source=twitter&utm__medium=social&utm__campaign=wired-&utm__brand=wired&utm__socialtype=owned;* J.M. Berger, "Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS On-line Social Media Networks," GW Program on Extremism, September 2016, available at *https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/Nazis%20v.%20ISIS.pdf*.

and what actions to take as a result, is unclear. It is completely a black box, and is not made transparent in the companies' transparency reports or in any other way. What is clear is that the final decision regarding what constitutes a violation of platform rules, including determinations regarding classifying specific content as anti-Muslim, anti-Semitic, or racist, is made solely and independently by the platform.

Some platforms have also provided the ability for users to appeal decisions made by the platform,[23] which will result in a second review of the content, and second determination by the platform—or if Facebook's new initiative gets off the ground,[24] by a sort of independent Supreme Court—as to whether that piece of content violates the platform rules. In the case of the New Zealand massacre, Facebook stated that 200 people watched the live-streamed video of the attack, and no one reported it to the platform, until it was brought to the company's attention by the authorities in New Zealand.[25] Indeed, the entire 17 minutes of the video remained up as it was being live-streamed. While Facebook has recently made some changes to its live-streaming function as a result of this,[26] the efficacy of those efforts is not clear.

Some of this process of content moderation can be automated through technology, including machine learning. This is especially true when there is little ambiguity regarding the nature of the content, as in the case of spam or child pornography. Tech companies also regularly include "violent content" as one category of content that can be easily caught through automated technological methods. The attack in New Zealand and social media's role in the spread of the live-streamed video of the event calls this claim into question. CTS recently wrote a piece discussing efforts companies can undertake to improve their efforts around live-streaming.[27]

Because hate speech or coded extremist activity require more context and nuance in review, this content is typically reviewed by a human moderator. Oftentimes, the automated tool and the human moderators work in tandem to detect and review complicated and nuanced hateful content. There, the tool will identify content that requires human judgment and will then route the content to the human reviewer. Once routed, the human reviewer will analyze the instance and either make a decision according to the platform rules or escalate for further review.

Once a piece of content is determined to have violated the rules of a particular platform, the platform then decides the appropriate consequences for violating the rules. Consequences range from the individual piece of a content being removed, to the user being suspended for a certain period of time, to the user (or community) being banned from the platform altogether. Each platform has its own unique approach to analyzing violations of its platform rules and implementing consequences. Certain platforms have also experimented with alternate types of consequences. For example, Reddit has explored placing offensive or objectionable communities in quarantine, making it harder for users not explicitly seeking out that content to find it, such as in the case of 9/11 Truthers and Holocaust Deniers.[28] Most recently, YouTube has taken a number of types of content—including white supremacist and Holocaust denial content—which it previously handled by placing in "limited state" and instead decided to remove these categories of content from the YouTube platform.[29] This may speak to the lack of efficacy of this particular alternative method of "limited state," if reducing the ability of this content to be found was not enough to reduce the harm caused by it. Barring more information as to the nature of policy change, we can only surmise as to the effectiveness of these alternative approaches.

Each company has specific and unique methods when handling hateful content on its platforms—from the definition of what is hateful and other rules, to content moderation practices, to actions and consequences. And each company shares its prac-

[23] Ian Wren, "Facebook Updates Community Standards, Expands Appeals Process," NPR, April 24, 2018, available at *https://www.npr.org/2018/04/24/605107093/facebook-updates-community-standards-expands-appeals-process;* Jacob Kastrenakes, "Twitter promises to respond more quickly to people who report abuse," The Verge, October 17, 2017, available at *https://www.theverge.com/2017/10/17/16492100/twitter-updated-abuse-hate-harassment-ban-rules;* Claudine Beaumont, "YouTube users can appeal against 'violations,'" The Telegraph, April 30, 2019, available at *https://www.telegraph.co.uk/technology/google/7876926/YouTube-users-can-appeal-against-violations.html.*

[24] Chris Sonderby, "Update on New Zealand," Facebook Newsroom, March 18, 2019, available at *https://newsroom.fb.com/news/2019/03/update-on-new-zealand/.*

[25] Ibid.

[26] *https://newsroom.fb.com/news/2019/05/protecting-live-from-abuse/.*

[27] *https://www.adl.org/news/article/livestreaming-hate-problem-solving-through-better-design.*

[28] Xeni Jardin, "Reddit 'quarantines' white supremacist, incel, holocaust denier, and other gross subreddits," Boing, September 28, 2018, available at *https://boingboing.net/2018/09/28/reddit-quarantines-major-w.html.*

[29] *https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html.*

tices with varying degrees of openness or transparency. That said, this overview should provide a surface-level understanding of how mainstream social media platforms function in terms of managing hate, violence, and extremism on-line, when there is no legal (or compelling business) mandate to do so.

EVALUATING EFFORTS BY COMPANIES

Evaluating the effectiveness of mainstream social media platforms' content moderation processes, however, is hard to gauge, especially in light of the scale at which these platforms operate. Platform content moderation is taking place on an on-going basis and will only become more important and more difficult as these platforms continue to grow. How well content moderation can scale is an open question, as platforms grow, as billions of new users come onto the internet, as what might be otherwise praiseworthy privacy innovations have the unintended consequence of making content moderation harder, and as disruptive technologies come on-line—such as virtually undetectable "deep fakes" that generate hate and violence while defeating detection. Already, in January 2019, it was reported that Facebook had 2.27 billion monthly active users globally, while YouTube had 1.9 billion.[30] As of December 2018, Facebook reported that of the 30,000 employees or contractors working on safety and security at Facebook, half of those are focused on reviewing content.[31] In late 2017, YouTube stated that it was increasing the number of individuals reviewing content to 10,000 people.[32]

At-scale information on the effectiveness of these efforts is currently only available via self-reported statistics from the companies, each with varying degrees of opacity and no form of on-going, independent, external auditing. More research on the nature of the problem is available from outside academics and civil society; however, this research also has no agreed-upon definitions or set of metrics to evaluate hateful and extreme content. Further, these groups have limited access to platform information or data, including on the prevalence of hateful content on a given platform. Some of the researchers are bound by non-disclosure agreements.

In spite of these limitations, there are two limited methods for understanding mainstream social media companies' efforts to address hateful, violent, and extreme content: Reports released by the tech companies, and external studies from academics and civil society.

REPORTING BY COMPANIES

One method of company reporting on hate is the transparency reports that tech companies release on a regular basis, without being legally required to do so. Transparency reports contain a set of metrics, set by each tech company, regarding moderation practices across self-selected content areas on their platforms. For example, Facebook's first transparency report in 2013 reported solely on the number of times governments asked Facebook for information on users, and the number of times Facebook responded.[33] Google's first transparency report from 2010 provided similar statistics, focused on Government requests to Google's suite of products, which included but did not disaggregate YouTube.[34] In 2018, both Facebook and Google/YouTube provided their first public statistics regarding their content moderation practices related to the enforcement of their community guidelines.[35]

They have since provided several updates on the enforcement of their community guidelines, which contain most of the same shortcomings CTS articulated earlier

[30] Statista, "Most famous social network sites 2019, by active users," January 2019, available at *https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.*

[31] Ellen Silver, "Hard Questions: Who Reviews Objectionable Content on Facebook—And Is the Company Doing Enough to Support Them?" Facebook Newsroom, July 26, 2018, available at *https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/.*

[32] Sam Levin, "Google to hire thousands of moderators after outcry over YouTube abuse videos," The Guardian, December 5, 2017, available at *https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos.*

[33] Lorenzo Franceschi-Bicchierai, "Facebook Releases First Transparency Report," Mashable, August 27, 2013, available at *https://mashable.com/2013/08/27/facebook-transparency-report/#j4SvneYzzPqx.*

[34] Google, Transparency Report, available at *https://web.archive.org/web/20100924044616/http://www.google.com:80/transparencyreport/.*

[35] 2 [sic] Guy Rosen, "Facebook Publishes Enforcement Numbers for the First Time," Facebook newsroom, May 15, 2018, available at *https://newsroom.fb.com/news/2018/05/enforcement-numbers/;* Catherine Shu, "YouTube releases its first report about how it handles flagged videos and policy violations," TechCrunch, available at *https://techcrunch.com/2018/04/23/youtube-releases-its-first-report-about-how-it-handles-flagged-videos-andpolicy-violations/.*

this year:[36] The limited, vague, and sometimes nonexistent, metrics in these transparency reports do not provide material information either to users, looking to frequent a particular platform, or to external researchers in academia or civil society looking to understand and combat the phenomena of hate on-line. For example, none of the figures provided by the companies can answer basic questions such as: "How much hate is there on platform X? Are there indications that the approaches to mitigating this problem by the company are working?" or "Is this platform a safe space for people who identify as X?" More concerning is the fact that these metrics are self-chosen and self-reported, so that there is no independent verification that they are either meaningful or accurate.

Additional reporting related to hate on-line has been conducted by the companies in relation to the "Code of Conduct on Countering Illegal Hate Speech On-line" which was signed by Facebook, Twitter, Google/YouTube and Microsoft and the European Union in 2016.[37] In this agreement tech platforms agreed to work to together with the European Union to address terrorism and hate speech on-line. Most notably, the code of conduct stipulates a 24-hour turnaround on reports of "illegal hate speech." In February 2019, the tech companies reported that 89 percent of flagged content is assessed within 24 hours and 72 percent of the content deemed to be illegal hate speech is removed. This is compared to 40 percent and 28 percent respectively when the Code was launched in 2016.[38] Once again, there is no information available about what communities are being impacted and how these figures relate to the prevalence of hate across an entire platform, let alone across platforms. Additionally, once again, these figures are self-reported by the companies, and are not verified by any independent third party. Nor are there agreed-upon and externally audited metrics about resultant (or corollary) reductions in the impact, in addition to incidence of hateful content.

EXTERNAL STUDY

The other limited pathway available to help understand the phenomena of hate on mainstream social media platforms is through external studies conducted by academic researchers and/or civil society. The advantage to this kind of study is that it exists outside of the corporate structure of tech companies, and thus can engage more freely in research and public communication regarding findings. However, because the phenomena of hateful content is so context dependent, there are currently no common frameworks, metrics or definitions to apply to these studies, thus making it hard to compare results. For example, in 2018, reports were released by ADL, the Community Security Trust (CST) in the United Kingdom and the World Jewish Congress on the nature of anti-Semitism on-line. Each report had its own methodology in terms of defining anti-Semitism and each were looking at different and incomplete parts of various on-line platforms.

At present, most studies of these kinds are based on data available from Twitter and Reddit. Both Twitter and Reddit provide the public with access to a portion of their data, while still respecting user privacy. This allows researchers to independently examine and measure the nature of hate on these platforms. However, the scale of the platforms is so vast and the resources of these external groups and individuals so limited that conducting any kind of analysis that is generalizable to any platform as a whole is extremely difficult and time-consuming.

ON-LINE HATE AND HARASSMENT: THE AMERICAN EXPERIENCE

Since its launch, CTS has taken an extensive look at the phenomena of hate on-line. Through various independent studies, CTS has worked to increase the public's understanding of how hate manifests on-line and has provided new ways to think about potential solutions to this monumental and multi-faceted problem.

In February 2019, CTS released the results of its survey on on-line hate and harassment in the United States.[39] The survey found that 53 percent of Americans experienced some form of on-line harassment, whereas 37 percent of Americans reported experiencing severe harassment, which includes physical threats, sexual harassment, stalking and sustained harassment. Of people who were targeted by harassment on-line based on their identity, the most targeted communities were the

[36] *https://www.adl.org/blog/we-need-real-transparency-about-hate-on-social-media.*

[37] European Commission, Code Of Conduct On Countering Illegal Hate Speech on-line *http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf;* Commission of the European Communities Press.

[38] Commission of the European Communities Press Release, IP/19/805, February 4, 2019, available at *http://europa.eu/rapid/press-release_IP-19-805_en.htm.*

[39] ADL, "On-line Hate and Harassment: The American Experience," available at *https://www.adl.org/on-lineharassment.*

LGBTQ community (63 percent), Muslims (35 percent), Latinx (30 percent), African-Americans (27 percent) and women (24 percent)

Notably, an overwhelming majority of respondents from across the political spectrum supported strengthening laws against perpetrators of on-line hate and harassment, strengthening laws applying to platforms, and providing more training to law enforcement on how to handle on-line hate and harassment.

The survey model of understanding the problem of hate on-line avoids the limitations of the data provided by the platforms and the definitions agreed to by external researchers by allowing respondents to self-report and define their own experience of on-line hate. For example, the platform where respondents said they most often experienced hate and harassment was Facebook, followed by Twitter and YouTube. Getting this kind of cross-platform comparative results on the experience of users regarding hate on-line through the data publicly available from these companies and platforms is currently impossible. A survey-based approach, however, is a very broad measure, and cannot get at the absolute level of prevalence of hate and harassment on any particular on-line platform at any one time.

COMPUTATIONAL PROPAGANDA, JEWISH-AMERICANS AND THE 2018 MIDTERMS: THE AMPLIFICATION OF ANTI-SEMITIC HARASSMENT ON-LINE

In November 2018, ADL released the report "Computational Propaganda, Jewish-Americans and the 2018 Midterms: The Amplification of Anti-Semitic Harassment on-line." The report focused on how tactics of disinformation, such as the use of automated accounts or bots to spread content, are being utilized to spread anti-Semitism on-line. The study consisted of both qualitative interviews with leaders and public figures in the Jewish community and a quantitative analysis of 7.5 million Twitter messages from August to September 2018.

The report found that nearly 30 percent of accounts engaging in anti-Semitic behavior were in fact bots, and that those bots made up over 40 percent of the anti-Semitic content in that time period. The qualitative results found that for the Jewish public figures who participated in the study, experiencing threats of violence and deluges of anti-Semitism had become part of their internal calculus for engaging in public life. For some, it drove them to speak out more loudly and vigorously; others, often citing concern over the harassment of family members, friends and romantic partners, sought to make adjustments to their on-line activity.

This type of study shows both the strengths and limits of studying on-line hate with data currently available from the companies. Given the data available from Twitter, the author, Sam Woolley, was able to look deeply at a particular moment in time on a platform, leading up to a particular event and to perform analysis of certain activities on the platform related to hate within that time frame. The limitation of this study is that we cannot generalize to the whole of one platform, such as Twitter, even within the narrow subject matter of how disinformation tactics spread anti-Semitism. To do so would require significantly more effort and resources. Without getting a great deal closer to understanding the prevalence and impact of particular hateful content, among other data points, it is difficult to devise the best mitigation tactics or to measure their effectiveness.

THE ON-LINE HATE INDEX

In an effort to provide a set of metrics or a common language as to the nature of hate on-line, ADL has been working in partnership with UC Berkeley's D-Lab on a project called the on-line Hate Index. The on-line Hate Index combines social science practices with machine learning techniques to create an academically rigorous way to understand the phenomena of hate on-line.

For a machine learning algorithm to actually learn, it requires large amounts of data that is labeled as to what it is or what it is not. For a machine learning algorithm to learn to detect hate speech on-line, it would need a large data set with some comments labeled as hateful and some labeled as not. At present, there are not many datasets that exist like this, and the ones that do exist are not very expansive. The on-line Hate Index project is working to provide a tool which will be available to the public that allows on-line comments to be labeled systematically and rigorously from a social science perspective, incorporating a myriad of community perspectives, so that there is a clear set of metrics and understandings as to the nature of hate on-line not only from the perspective of the speaker, but from the targets.

This approach is novel in the sense that it is not directly engaging in research on the problem, but rather creating a methodology whereby future research of hate on-line can be conducted in a more systematic and uniform way. The issue here is, again, the tool will only be as good as the data to which it has access. The limited data currently provided by tech platforms limits the ability of innovative researchers such as the team at UC Berkeley's D-Lab from creating a shared understanding of the problem in the research community.

POLICY RECOMMENDATIONS

The challenges discussed above are complex and wide-ranging and require a whole-of-society response. There is no magic bullet. Indeed, there is not even a collective set of magic bullets. A constantly iterative interdisciplinary approach will be needed, drawing upon education in K–12 and in universities, engagement of various professions and industries (including, perhaps venture capital firms), a change in the divisive and polarizing rhetoric that has become mainstream at the highest levels of government, the training of engineers, including those developing games, the creation of tools and better coordination between humans and those tools, the inclusion of targeted communities and the groups that represent them, innovative marketing campaigns, litigation, and legislation, and reform in self-governance, to name a handful. How to balance content moderation and privacy and free expression concerns will remain challenging, to say the least.

Nonetheless, we must start somewhere, and quickly. Below are some initial recommendations for government and tech platforms to address hate on-line and the threat that it poses to Americans and people around the world.

ON-LINE HATE: POLICY RECOMMENDATIONS FOR GOVERNMENT

Strengthen Laws Against Perpetrators of On-line Hate

- Hate and harassment have moved from on the ground to on-line, but our laws have not kept up. Many forms of severe on-line misconduct are not consistently covered by current cyber crime, harassment, stalking, and hate crime laws. While many of these issues can and should be enacted and enforced at the State level, Congress has an opportunity to lead the fight against cyber hate by increasing protections for targets as well as penalties for perpetrators of on-line misconduct.
- Some actions Congress can take include:
  - Revising Federal law to allow for penalty enhancements based on cyber-related conduct
  - Updating Federal stalking and harassment statutes' intent requirement to account for on-line behavior where intent or targeting is not present in the traditional sense but the harm to the individual is just as devastating;
  - Legislating specifically on cyber crimes such as doxing, swatting, and non-consensual pornography. ADL endorsed the on-line Safety Modernization Act, which was introduced in the last Congress to fill these gaps.

Urge Social Media Platforms to Institute Robust Governance

- Government officials have an important role to play in encouraging social media platforms to institute robust and verifiable industry-wide self-governance. This could take many forms, including Congressional oversight or passage of laws that require certain levels of transparency and auditing. As noted, one-size-fits-all laws specifying particular types of content moderation are unlikely to be effective. The internet plays a vital role in allowing for innovation and democratizing trends, and that should be preserved. At the same time the ability to use it for hateful and severely harmful conduct needs to be effectively addressed. An escalating series of regulations, depending upon a platform's successful self-regulation, may be an option. There are other areas of law to which we can look to find systems that allow individual companies to meet required thresholds in the ways best suited for the manner in which they operate.

Improve Training of Law Enforcement

- Law enforcement is a key responder to on-line hate, especially in cases when users feels they are in imminent danger. Increasing resources and training for these departments is critical to ensure they can effectively investigate and prosecute cyber cases and that targets know they will be supported if they contact law enforcement.

ON-LINE HATE RECOMMENDATIONS FOR INDUSTRY

Enhance Transparency

- Platforms must report meaningful statistics to the public about the prevalence of hate on their platforms. The metrics of these reports should be determined in consultation with trusted third parties so that they will be of value to the communities most impacted by hate on-line.

Improve Accountability

- Any public reporting done by tech companies regarding hate on-line, whether through transparency reports or reporting through other initiatives, should be reviewed and verified by trusted third parties. Additionally, platforms should submit to an external audit of hate on their platforms, to allow for a fully independent analysis of the effectiveness of a company's policies and practices in terms of mitigating hate on-line.

Provide Data

- Platforms should, while respecting the privacy of their users, provide meaningful data to external researchers to advance understanding of the problem of hate on-line and to promote innovation in solutions to mitigate the problem.

Ensure Strong Policies Against Hate

- Privacy-by-design has become a best practice over the past years. At the risk of being a bit facile, so must "anti-hate-by-design." Every social media platform must have clear and transparent terms of service that address hateful content and harassing behavior, and clearly define consequences for violations. These policies should include, but should not be limited to:
  - Making clear that the platform will not tolerate hateful content or behavior on the basis of protected characteristics.
  - Prohibiting abusive tactics such as harassment, doxing, and swatting.
  - Establishing an appeal process for users who feel their content was flagged as hateful or abusive in error.

Strengthen Enforcement of Policies

- Social media platforms should assume greater responsibility to enforce their policies and to do so accurately at scale. This means:
  - Improving the complaint process so that it provides a more consistent and speedy resolution for targets. We know from research that content moderators regularly make mistakes when it comes to adjudicating hateful content.
  - Relying less on complaints from individual users, and instead proactively, swiftly, and continuously addressing hateful content using a mix of artificial intelligence and humans who are fluent in the relevant language and knowledgeable in the social and cultural context of the relevant community.

Design to Reduce Influence and Impact of Hateful Content

- Social media companies should design their platforms and algorithms in a way that reduces the influence of hateful content and harassing behavior. Steps should include:
  - Making hateful content more difficult to find in search and algorithmic recommendations. This means, for example, never recommending hatemongers' tweets, suggesting them as friends, or auto-playing their videos.
  - Removing advertisements from hateful content.
  - Not allowing hateful content to be monetized for profit.
  - Labeling content suspected to be from automated "bot" accounts, given the use of bots for spreading hate.

Expand Tools and Services for Targets

- Given the prevalence of on-line hate and harassment, platforms should offer far more user-friendly services, tools, and opportunities for individuals facing or fearing on-line attack. This includes:
- Greater filtering options that allow individuals to decide for themselves how much they want to see likely hateful comments. What goes into default settings should also be considered.
- Protections for individuals who are being harassed in a coordinated way.
- User-friendly tools to help targets preserve evidence and report problems to law enforcement and companies.
- Enhanced industry support for counter-speech initiatives, including fostering, aggregating and promoting positive messages responding to offensive content.

CONCLUSION

We implore you and all public leaders to consistently call out bigotry and extremism at every opportunity. We all have a responsibility to make clear that America is no place for hate.

We at ADL look forward to working with Members of the committee and the tech industry to understand and combat hate on-line, and to ensure justice and fair treatment to all in digital spaces.

Chairman THOMPSON. I thank the witnesses for their valuable testimony and Members for their questions.

The Members of the committee may have additional questions for the witnesses, and we ask that you respond expeditiously in writing to those questions.

The other point I would like to make, for Facebook, you were 30 hours late with your testimony. Staff took note of it. For a company your size, that was just not acceptable for the committee. So I want the record to reflect that.

Without objection, the committee record shall be kept open for 10 days.

Hearing no further business, the committee stands adjourned.

[Whereupon, at 12:30 p.m., the committee was adjourned.]

APPENDIX

QUESTIONS FROM CHAIRMAN BENNIE G. THOMPSON FOR MONIKA BICKERT

Question 1. Does your company currently make data on on-line terror content and misinformation—including the amount and types of content you remove—available for academics and other stakeholders to research? In what ways or what partnerships? Will you commit to making such data available to researchers?

Answer. To track our progress and demonstrate our continued commitment to make Facebook safe and inclusive, we regularly release our Community Standards Enforcement Report. This report shares metrics on how Facebook is performing in preventing and removing content that goes against our Community Standards. We also release a "prevalence" metric that estimates how much violating content has been posted on the platform. The report is focused on the following categories:

- Terrorist propaganda (ISIS, al-Qaeda and affiliates)
- Adult nudity and sexual activity
- Bullying and harassment
- Child nudity and sexual exploitation of children
- Fake accounts
- Hate speech
- Regulated goods (drugs and firearms)
- Spam
- Violence and graphic content.

For the first time in our May 2019 report, we also began sharing data on our process for appealing and restoring content to correct mistakes in our enforcement decisions. That report can be viewed at *https://transparency.facebook.com/community-standards-enforcement.* We continue to look for ways to expand and enhance the report moving forward.

We have also launched the Content Policy Research Initiative (CPRI), which invites experts and researchers to help inform development of our content policies and assess possible product solutions to countering hateful and harmful content. At present, CPRI is focused on:

- Hate speech and harmful speech
- Preventing off-line harm
- Bullying and harassment
- Fairness in global enforcement.

CPRI is comprised of both funding opportunities to support external research on content policy issues, as well as workshops where we openly share internal research methodology, discuss how we measure violations on the platform, explain policy making and processes, and work to identify areas that are ripe for collaboration with the research community. For more information about CPRI, see *https://research.fb.com/programs/content-policy-research/#About.*

Question 2. Private posts containing objectionable content pose a unique challenge for moderation. How does your company reconcile data privacy with the challenges of moderating content that violates your terms of service, including terrorist content and misinformation?

Answer. Although the visibility of material varies for the general public based on the setting in which it is posted, our systems can proactively detect and remove violating content, including terrorist content, to help improve on-line safety. We do this by analyzing specific examples of bad content that have been reported and removed to identify patterns of behaviors. Those patterns can be used to teach our software to proactively find other, similar problems.

Question 3. Does your company currently share AI training data related to counterterrorism with other social media companies? If not, is there a plan to share such data in the future?

Answer. We are 1 of 4 founding members of the Global Internet Forum to Counter Terrorism (GIFCT). As part of this industry partnership, we are jointly focused on tech innovation—one of the key pillars of GIFCT's work. The partnership is crucial to combating terrorist content on on-line platforms. GIFCT is committed to working on technological solutions to help thwart terrorists' use of our services, including through a shared industry hash database, where companies can create "digital fingerprints" for terrorist content and share it with other participating companies. The database, which became operational in the spring of 2017, now includes 15 companies that contribute to it, more than 200,000 visually distinct image hashes, and more than 10,000 visually distinct video hashes. It allows the 15 member companies to use those hashes to identify and remove matching content—videos and images—that violate our respective policies or, in some cases, block terrorist content before it is even posted. Each company has different policies, practices, and definitions as they relate to extremist and terrorist content. If content is removed from a company's platform for violating that platform's individual terrorism-related content policies, the company may choose to hash the content and include it in the database. GIFCT also has created an on-line resource for smaller tech companies to seek support and feedback.

We recognize that our work is far from done, but we are confident that we are heading in the right direction. We will continue to provide updates as we forge new partnerships and develop new technology in the face of this global challenge.

Question 4. How is your company working together with other companies to share technology, information, and resources to combat misinformation? Is there an analogue to the Global Internet Forum to Counter Terrorism (GIFCT) for misinformation on your platforms?

Answer. We believe that tech companies, media companies, newsrooms, and educators all need to work together to address the societal problem of misinformation. We are engaged with partners across these industries to help create a more informed community.

In doing so, we have greatly expanded our efforts to fight false news: We are getting better at removing fake accounts and coordinated inauthentic behavior; we are using both technology and people to fight the rise in photo- and video-based misinformation; we have deployed new measures to help people spot false news and get more context about the stories they see in News Feed; and we have grown our third-party fact-checking program to include 54 certified fact-checking partners who review content in 42 languages. And we are making progress. Multiple research studies suggest that these efforts are working and that misinformation on Facebook has been reduced since the U.S. Presidential elections in 2016.

But misinformation is a complex and evolving problem, and we have much more work to do. With more than a billion things posted to Facebook each day, we need to find additional ways to expand our capacity. The work our professional fact-checking partners do is an important piece of our strategy. But there are scale challenges involved with this work. There are simply not enough professional fact-checkers world-wide, and fact-checking—especially when it involves investigation of more nuanced or complex claims—takes time. We want to be able to tackle more false news, more quickly.

As we have worked to expand our misinformation efforts over the past 2 years, we have also been doing extensive research and talking to outside experts to identify additional approaches that might bolster our defenses. One promising idea we have been exploring involves relying on groups of people who use Facebook to point to journalistic sources that can corroborate or contradict the claims made in potentially false content.

We are also consulting a wide range of academics, fact-checking experts, journalists, survey researchers, and civil society organizations to understand the benefits and risks of ideas like this. We are going to share with experts the details of the methodology we have been thinking about to help these experts get a sense of where the challenges and opportunities are, and how they will help us arrive at a new approach. We will also share updates from these conversations throughout the process and find ways to solicit broader feedback from people around the world who may not be in the core group of experts attending these roundtable events.

We all must work together to find industry solutions that strengthen the on-line news ecosystem and our own digital literacy. That is why we are collaborating with others who operate in this space. For instance, through the Facebook Journalism Project, we seek to establish stronger ties between Facebook and the news industry. The project is focused on developing news products, providing training and tools for journalists, and working with publishers and educators on how we can equip people with the knowledge they need to be informed readers in the digital age.

Taking the fight against misinformation to the next level is an important task for us. There are elections around the world month after month, only adding to the everyday importance of minimizing false news. We plan to move quickly with this work, sharing some of the data and ideas we have collected so far with the experts we consult so that we can begin testing new approaches as soon as possible.

Question 5. What is your company doing to ensure that your human content moderators are provided with all the resources they need in order to carry out their jobs, including mental health resources and adequate pay?

Answer. We are committed to providing support for our content reviewers, as we recognize that reviewing certain types of content can be hard. That is why everyone who reviews content for Facebook goes through an in-depth, multi-week training program on our Community Standards and has access to extensive support to ensure their well-being. This includes on-site support with trained practitioners, an on-call service, and health care benefits from the first day of employment.

Facebook actively requests and funds an environment that ensures this support is in place for the reviewers employed by our partners, with contractual expectations around space for resiliency and wellness, wellness support, and benefits including health care, paid time off, and bonuses.

In 2015, we introduced a new set of standards for people who do contract work in the United States and since 2016, we have also required vendors in the United States to provide comprehensive health care to all of their employees assigned to Facebook. In the years since, it has become clear that $15 per hour does not meet the cost of living in some of the places where we operate. After reviewing a number of factors including third-party guidelines, we are committing to providing compensation that reflects local costs of living. This means a raise to a minimum of $20 per hour in the San Francisco Bay Area, New York City, and Washington, DC, and $18 per hour in Seattle. We will be implementing these changes by mid-next year, and we are working to develop similar standards for other countries.

For workers in the United States that review content on Facebook, we are raising wages even more. Their work is critical to keeping our community safe, and it is often difficult. That is why we have paid content reviewers more than minimum wage standards, and why we will surpass this new living wage standard as well. We will pay at least $22 per hour to all employees working for our vendor partners based in the Bay Area, New York City, and Washington, DC; $20 per hour to those living in Seattle; and $18 per hour in all other metro areas in the United States. As with all people who do contract work, we are working to develop similar international standards. This work is on-going, and we will continue to review wages over time.

In addition to pay, we collaborate with our vendor partners to ensure they are providing a holistic approach to well-being and resiliency that puts the needs of their employees first. We have a team of clinical psychologists across three regions who are tasked with designing, delivering, and evaluating resiliency programs for everyone who works with objectionable content. This group works closely with our partners and each of their dedicated resiliency professionals to help build resiliency programming standards for their teams and share best practices. These programs are important as support and resiliency is so personal to each and every person. Everyone has their own way to build resilience and we, and our partners, work hard to ensure that resources are in place to help do that.

We are also employing technical solutions to limit exposure to graphic material as much as possible. For the first time, we have added preferences that let reviewers customize how they view certain content. For example, they can now choose to temporarily blur graphic images by default before reviewing them. We made these changes after hearing feedback that reviewers want more control over how they see content that can be challenging.

In April, we hosted all of our vendor partners at a summit to discuss on-going improvement and commitment to the work in these areas. We also formed an Advisory Working Group specific to resiliency issues. The group includes a subject-matter expert from each vendor partner to ensure that we are sharing across partners and setting standards going forward.

Content review at our size can be challenging and we know we have more work to do. This is an important issue and we are committed to getting this right and to supporting our content reviewers in a way that puts their well-being first.

Question 6. Prior to introducing new policies or products on your platforms, what processes does your company have in place to anticipate and plan for unintended consequences, harmful side effects, or exploitation by bad actors, including terrorists and those seeking to spread misinformation?

Answer. Our Community Standards are a living set of guidelines—they must keep pace with changes happening on-line and in the world. The core of our policy development process is a twice-monthly, global meeting where we debate and discuss potential changes to our Community Standards. 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 both a range of perspectives and opinions across the globe, as well as unintended consequences and efforts to thwart our policies. When our policies are written or updated, we share those updates on our Community Standards website. More information about this process is available at *https://newsroom.fb.com/news/2019/04/insidefeed-community-standards-development-process/* and *https://www.facebook.com/communitystandards/additional__information.*

Question 7. Facebook's latest Transparency Report contains some metrics for understanding Facebook's efforts to combat hate speech on its platform, but the report says that Facebook can't measure the prevalence of hate content. This is concerning because users, advocacy organizations, and Congress can only make sense of Facebook's enforcement performance if it can be compared to the prevalence of hate on the platform. What is preventing Facebook from reporting on the prevalence of hate content on your platform? Can Facebook report on U.S.-specific data? Does Facebook plan to report on this in the future?

Community Standards Enforcement Report: Hate Speech, Facebook, *https://transparency.facebook.com/community-standards-enforcement#hate-speech* (accessed July 9, 2019).

Answer. We cannot currently estimate the prevalence of hate content on Facebook. But our prevalence measure is expanding to cover more languages and regions and to account for cultural context and nuances for individual languages.

Measuring prevalence is difficult in some scenarios because it requires sampling content randomly. This prevalence methodology requires a very large number of content samples to estimate a precise measure for violations that are viewed very infrequently, such as Terrorist Propaganda. For these types of violations, we can only estimate the upper limit of violating views—meaning that we are confident that the prevalence of violating views is below that limit.

Question 8. How do you anticipate making GIFCT effective in the next 5 years? Is there a plan to increase the resources devoted to GIFCT? Is there a plan to have a permanent staff and a physical location?

Answer. The Global Internet Forum to Counter Terrorism (GIFCT) was founded to improve the ability of technology companies to identify and remove terrorist content. It is not a panacea, and has never been presented as such. Many of its programs are designed to help smaller technology companies improve their enforcement efforts.

When GIFCT was founded in 2017, we worked in 3 workstreams—employing and sharing technology, facilitating knowledge sharing, and supporting research. We have made major progress in these areas. Our shared hash-database includes hashes from more than 200,000 visually distinct pieces of content; we have engaged more than 120 technology companies in 11 workshops on 4 continents; and the research network we sponsored has published 7 papers on a range of issues, including lessons learned from regulation of the financial industry and a comparative study of how countries in the Five Eyes define and counter terrorism. Over the course of 2019, we will be holding workshops in Jordan, California, India, and the United Kingdom, along with a high-level event with the United Nations' General Assembly in September.

In the wake of the Christchurch attacks, we made the decision to add a fourth major workstream: Developing the ability to cooperate in a crisis. That commitment was drawn from elements of the Christchurch Call and led to an announcement on July 24 at the GIFCT Annual Summit about a new Content Incident Protocol to be used by the 4 founding GIFCT companies. The Protocol includes the ability to quickly spin up dedicated mechanisms within our hash-sharing database to share information relevant to a crisis scenario.

Strengthening GIFCT is critical going forward and we are working with our partner companies to consolidate the consortium and ensure it can play a stronger role in the years to come. At the same time, policy makers must understand that the vast majority of the work that Facebook does to identify and remove terrorist content is not dependent on GIFCT. Rather, it relies on internal efforts and technology. GIFCT is a critical tool to leverage those efforts across industry. But the most important enforcement work we do on Facebook is driven internally.

More information about GIFCT is available on its website: *www.gifct.org.*

Question 9. Have members of GIFCT agreed on a common standard for prohibited terrorist content? If so, what are those standards and how do you ensure they are updated?

Does GIFCT meet regularly to discuss trends in terrorist content? In addition to combatting ISIS and al-Qaeda content, does GIFCT focus on content related to other designated foreign terrorist organizations as well as right-wing extremist groups, such as white supremacist extremists? If not, is there a plan to do so?

Answer. GIFCT members must prohibit terrorism in their Terms of Service, enable user reports of terrorism, agree to collaborate on technology, and commit to transparency reports. GIFCT also supports Tech Against Terrorism, an NGO that provides coaching for companies that need to develop these elements.

The United Nations has been debating the definition of "terrorism" for decades and even U.S. agencies define terrorism differently. The hash-sharing database is structured around the United Nations' Consolidated Sanctions list, with the exception of material produced during a crisis. Hash-sharing is ultimately a referral mechanism, but each company enforces against content per its own policies.

Facebook's internal definitions of terrorism are available in our public-facing Community Standards. We remove content that supports Foreign Terrorist Organizations and define terrorism based on the behavior of groups and individuals. This means that we have long listed a wide-range of organizations—jihadis, right-wing, and left-wing—as terrorists.

Questions From Honorable Lauren Underwood for Monika Bickert

Question 1. During the hearing, you committed to providing in writing details on Facebook's partnership with Life After Hate, including financial support, and on any plans Facebook has to provide Life After Hate with continuous funding for the duration of Facebook's partnership with them. Please provide this information.

Answer. We support Life After Hate's mission and, as Monika Bickert testified, they are "doing great work with us." We provided Life After Hate with an initial grant when we set up the redirect initiative. We are currently working with the organization to help it manage the increased volume as a result of our productive partnership. We are awaiting their proposal to upscale our support and expect to have additional funding for them in the near future.

Question 2a. During the hearing, you committed to provide in writing the percentage of Facebook users who click on links to "additional reporting" that Facebook displays next to content that contains disinformation.

Please provide this information, broken down by each month since Facebook began displaying links to additional reporting next to content that contains disinformation.

Question 2b. Please provide a complete list, as well as a breakdown by percentage, of the websites that Facebook's suggested "additional reporting" links to.

Answer. We do not capture metrics that would allow us to determine what percentage of Facebook users click on links that provide additional reporting. But we recognize that this is an important issue. False news is bad for people and bad for Facebook. Therefore, to help people make informed decisions about what to read, trust, and share, we built products that give people more information directly in News Feed. We also demote false news, which is one of our best weapons because demoted articles typically lose 80 percent of their traffic.

We are continuing to evaluate these methods to ensure that they are providing a clear signal to people about the credibility of fact-checked content when users encounter such content on Facebook.

We have found this strategy to be successful. For example, we saw that when we started showing related articles to people—and in doing so, made the context from fact-checkers front and center in News Feed—people were less likely to share the false stories.

We know there is more to do, and we are prioritizing fighting misinformation. We would be happy to brief you or your staff to provide you with more information about our efforts in this area.

Question 3a. During the hearing, you stated that Facebook has "launched some recent measures" to combat vaccine hoaxes and disinformation, and that you would have additional measures in partnership with "major health organizations" in place "soon."

Please provide a detailed description of all measures that Facebook currently has in place to combat vaccine hoaxes and disinformation on Facebook, Instagram, and WhatsApp.

Question 3b. Please provide a detailed description and an exact date of implementation for those additional future measures.

Question 3c. Please provide a list of the "major health organizations" that Facebook is working with to combat vaccine disinformation and hoaxes.

Answer. We are working to tackle vaccine misinformation on Facebook by reducing its distribution and providing people with authoritative information on the topic. Our efforts include:

- Reducing the ranking of groups and pages that spread misinformation about vaccinations in News Feed and Search. These groups and pages are not included in recommendations or in predictions when you type into Search.
- When we find ads that include misinformation about vaccinations, we reject them. We also have removed related targeting options, like "vaccine controversies." For ad accounts that continue to violate our policies, we may take further action, such as disabling the ad account.
- We do not show or recommend content that contains misinformation about vaccinations on Instagram Explore or hashtag pages.
- We are exploring ways to share educational information about vaccines when people come across misinformation on this topic.
- We have also removed access to our fundraising tools for pages that spread misinformation about vaccinations on Facebook.

As part of our effort to combat vaccine misinformation, we work with global health organizations, such as the World Health Organization and the U.S. Centers for Disease Control and Prevention, which have publicly identified verifiable vaccine hoaxes. If these vaccine hoaxes appear on Facebook, we take action against them. For example, if a group or page admin posts vaccine misinformation, we exclude the entire group or page from recommendations, reduce these groups' and pages' distribution in News Feed and Search, and reject ads with this misinformation.

We also believe in providing people with additional context so they can decide whether to read, share, or engage in conversations about information they see on Facebook. We are exploring ways to give people more accurate information from expert organizations about vaccines at the top of results for related searches, on pages discussing the topic, and on invitations to join groups about the topic.

Question 4. Before Facebook announced its digital currency, Libra, did Facebook evaluate or otherwise conduct "red-teaming" to assess potential use of Libra for gang activity, terrorism, child abuse, or by other bad actors?

Answer. We made the deliberate decision to announce the plans for Libra early after an initial consultative phase with regulators, central banks, and other organizations. The time between now and launch is designed to be an open, collaborative process. We know that we cannot do this alone and that engaging with regulators, policy makers, and experts is critical to Libra's success. We will take the time to get this right. The Libra Association will set standards for its members to maintain anti-money-laundering and anti-fraud programs, and to cooperate with legitimate law enforcement investigations. The Association will also develop monitoring programs and work with vendors who have expertise in identifying illicit activity on public blockchains. That said, most of the work of preventing illicit activity will happen at the service-provider level. These service providers will include payment services and marketplaces that are already trusted today by millions of people to complete their transactions safely, and that have major investments in people and technology to fight fraud and prevent illicit activity.

The service provider for which Facebook will be responsible is Calibra, a Facebook subsidiary. Calibra will incorporate know-your-customer and anti-money-laundering methodologies used around the world, including those focused on customer identification and verification, and risk-based customer due diligence, while developing and applying technologies such as machine learning to enhance transaction monitoring and suspicious activity reporting. Calibra's efforts will be commensurate with its risk profile based on several factors, such as Calibra's product features, customer profiles, geographies, and transaction volumes.

QUESTIONS FROM RANKING MEMBER MIKE ROGERS FOR MONIKA BICKERT

Question 1. During the hearing, Professor Strossen raised a number of concerns about the dangers of moderating speech. Rather than censorship, she offered a number of suggestions to empower users. What, if anything, are your companies doing to provide more filtering tools to enhance the ability of users to control the content they see?

Answer. The goal of News Feed is to connect people with the posts they find most relevant. We want to ensure that people see the content that is important to them—whether that is posts from family and friends or news articles and videos from pages they follow.

We have built, and are continuing to build, new controls so that people can tell us directly what content they want to prioritize, take a break from, or get rid of in their News Feed. If our users want to make sure they see everything from a certain person, they can use the "See First" feature to put that person's posts at the top of their Feed (for more information, see *https://www.facebook.com/help/1188278037864643).* If they have heard too much from someone, users can

"Unfollow" that person (for more information, see *https://www.facebook.com/help/190078864497547).* If users just want to take a break from someone, the "Snooze" feature removes that person from their News Feed for 30 days (for more information, see *https://www.facebook.com/help/538433456491590).*

Question 2. Do you have recommendations for ways to more effectively address the extremist content found on many of the off-shore, fringe social media sites?

Answer. One idea is for third-party bodies to set standards governing the distribution of harmful content and measure companies against those standards. Regulation could set baselines for what is prohibited and require companies to build systems for keeping harmful content to a bare minimum.

Facebook already publishes transparency reports on how effectively we are removing harmful content. We believe every major internet service should do this quarterly, because it is just as important as financial reporting. Once we understand the prevalence of harmful content, we can see which companies are improving and where we should set the baselines.

We are also a founding member of the Global Internet Forum to Counter Terrorism (GIFCT), through which we partner with others in the tech industry to combat terrorism and violent extremism on-line. Our work is focused on four key areas, one of which is sharing knowledge with smaller tech companies and bringing other sectors' expertise to the table.

In this vein, we have partnered with Tech Against Terrorism to host 11 workshops in 9 countries on 4 continents. As a result, we have engaged with over 120 tech companies, over 25 NGO's, and 12 government bodies. And just recently, we rolled out a cross-platform counter-violent extremist toolkit that we jointly developed with the Institute for Strategic Dialogue. The toolkit will assist civil society organizations in developing on-line campaigns to challenge extremist ideologies, while prioritizing their safety, and will be available on-line soon. We know that the technology industry is not the best or most appropriate messenger when it comes to pushing back on violent extremists, which is why it is so important to support civil society organizations that have the credibility and knowledge to combat, respond to, and counter the promotion of violent extremism on-line.

Question 3. Can you describe the appeals process within your platform for users to challenge content removal decisions? How quickly does this process occur and how do you incorporate lessons learned when your company reverses a removal decision?

Answer. In April 2018, we introduced the option to request re-review of individual pieces of content that were removed for adult nudity or sexual activity, hate speech, or graphic violence. We have since extended this option so that re-review is now available for additional content areas, including dangerous organizations and individuals (a content area that includes our policies on terrorist propaganda), bullying and harassment, and regulated goods. We are also making this option available to individuals who have reported content that was not removed.

In order to request re-review of a content decision we made, in most instances you are given the option to "Request Review." We try to make the opportunity to request this review clear, either via a notification or interstitial, but we are always working to improve. Typically, re-review takes place within 24 hours.

Transparency in our appeals process is important, so in our May 2019 Community Standards Enforcement Report, we began including how much content people appealed and how much content was restored after we initially took action. Gathering and publishing those statistics keeps us accountable to the broader community and enables us to continue improving our content moderation. For more information, see *https://transparency.facebook.com/community-standards-enforcement.*

Question 4. Moderating terror and extremist content on social media platforms is a complex issue with no perfect solution. One consistent recommendation the committee has received from a variety of outside experts is the value of greater transparency in your respective content removal policies. What is your response to calls for you to open up your platforms for academics for research purposes, particularly allowing them to review the content you remove?

Answer. We are committed to transparency at Facebook. That is why we decided to publish our internal guidelines. Facebook's Community Standards are available at *https://www.facebook.com/communitystandards/.*

We publish these internal guidelines for two reasons. First, the guidelines help people understand where we draw the line on nuanced issues. Second, providing these details makes it easier for everyone, including experts in different fields, to give us feedback so that we can improve the guidelines—and the decisions we make—over time. The content policy team at Facebook, which is responsible for developing our Community Standards, seeks input 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.

To track our progress and demonstrate our continued commitment to making Facebook safe and inclusive, we regularly release a Community Standards Enforcement Report, which includes metrics on how Facebook is performing in preventing and removing content that violates our Community Standards. In total, we are now including metrics across 9 policies within our Community Standards: Adult nudity and sexual activity, bullying and harassment, child nudity and sexual exploitation of children, fake accounts, hate speech, regulated goods, spam, global terrorist propaganda, and violence and graphic content. For more information, see *https://transparency.facebook.com/community-standards-enforcement.*

We are also moving forward with plans to establish an independent Oversight Board so people in the community can appeal our content decisions. We know that our systems can feel opaque and people should have a way to hold us accountable and make sure that we are enforcing our standards fairly. This independent Oversight Board will look at some of our hardest cases, and the decisions it makes will be binding. We have spent the first half of this year working with experts on speech and safety, running workshops around the world, and asking for public input on how this could work. We published a report with all the feedback we have gotten so far at the end of June. For more information, see *https://newsroom.fb.com/news/2019/06/global-feedback-on-oversight-board/.*

And with regard to the call from academics that we open up our platform for research purposes, we launched the Content Policy Research Initiative (CPRI), which invites experts and researchers to help inform development of our content policies and assess possible product solutions to countering hateful and harmful content. At present, CPRI is focused on:

- Hate speech and harmful speech
- Preventing off-line harm
- Bullying and harassment
- Fairness in global enforcement.

CPRI is comprised of both funding opportunities to support external research on content policy issues as well as workshops where we openly share internal research methodology, discuss how we measure violations on the platform, explain policy making and processes, and work to identify areas that are ripe for collaboration with the research community. For more information about CPRI, see *https://research.fb.com/programs/content-policy-research/#About.*

Questions From Chairman Bennie G. Thompson for Nick Pickles

Question 1. Does your company currently make data on on-line terror content and misinformation—including the amount and types of content you remove—available for academics and other stakeholders to research? In what ways or what partnerships? Will you commit to making such data available to researchers?

Answer. Twitter is a uniquely open service. The overwhelming majority of content posted is publicly available and made available through our free public and commercial application programming interfaces ("APIs"). We make public Tweet data available to Twitter users, developers, researchers, and other third parties. We encourage developers and others to create products using this public data for purposes that serve the public interest and the general Twitter community. Such uses have included saving lives during flooding in Jakarta, helping the U.S. Geological Survey track earthquakes, and working with the United Nations to achieve its Sustainable Development Goals. This service is a hallmark of our commitment to transparency, collaboration, and innovation.

Moreover, in October 2018, we published the first comprehensive archive of Tweets and media associated with suspected state-backed information operations on Twitter and since then we have provided two further updates covering a range of actors. Thousands of researchers from across the globe have now made use of these datasets, which contain more than 30 million Tweets and more than 1 terabyte of media, using our archive to conduct their own investigations and to share their insights and independent analysis with the world.

By making this data open and accessible, we seek to empower researchers, journalists, governments, and members of the public to deepen their understanding of critical issues impacting the integrity of public conversation on-line, particularly around elections. This transparency is core to our mission.

Additionally, for the past 7 years, our biannual Twitter Transparency Report *(transparency.twitter.com)* has highlighted trends in requests made to Twitter from around the globe. Over time, we have significantly expanded the information we disclose adding metrics on platform manipulation, Twitter Rules enforcement, and our proactive efforts to eradicate terrorist content, violent extremism, and child sexual exploitation from our service.

We have now suspended more than 1.5 million accounts for violations related to the promotion of terrorism between August 1, 2015, and December 31, 2018. According to our most recent Twitter Transparency Report, in 2018, a total of 371,669 accounts were suspended for violations related to promotion of terrorism. We continue to see more than 90 percent of these accounts suspended through proactive measures.

The trend we are observing year-over-year is a steady decrease in terrorist organizations attempting to use our service. This is due to zero-tolerance policy enforcement that has allowed us to take swift action on ban evaders and other identified forms of behavior used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage—before the account even Tweets. We are encouraged by these metrics but will remain vigilant. Our goal is to stay one step ahead of emergent behaviors and new attempts to circumvent our robust approach.

Finally, the Global Internet Forum to Counter Terrorism (GIFCT) facilitates, among other things, information sharing; technical cooperation; and research collaboration, including with academic institutions. GIFCT's partnership with the Royal United Services Insitute (RUSI) to establish the Global Research Network on Terror and Terrorism highlights the industry's commitment to working closely with academics and researchers. The Network is a consortium of academic institutions and think tanks that conducts research and shares views on on-line terrorist content; recruiting tactics terrorists use on-line; the ethics and laws surrounding terrorist content moderation; public-private partnerships to address the issue; and the resources tech companies need to adequately and responsibly remove terrorist content from their platforms.

This network is providing valuable insights and feedback. For example, one recent paper from the network, used Twitter's open API to evaluate attempts by Daesh (also known as the Islamic State of Iraq and Syria, ISIS) to use Twitter to disseminate its on-line magazine, *Rumiyah.* The researchers found: "Twitter was effective in its response to Daesh's attempts to use its platform as a gateway to Rumiyah . . . a high proportion of the user accounts that posted outlinks to PDFs of *Rumiyah* were suspended and the tweets that these accounts posted received relatively few retweets." See Stuart Macdonald, Daniel Grinnell, Anina Kinzel, and Nuria Lorenzo-Dus, Global Research Network on Terrorism and Technology: Paper No. 2, *A Study of Outlinks Contained in Tweets Mentioning Rumiyah* (2019) (on-line at *https://rusi.org/sites/default/files/20190628_grntt_paper_2_0.pdf).*

Question 2. Private posts containing objectionable content pose a unique challenge for moderation. How does your company reconcile data privacy with the challenges of moderating content that violates your terms of service, including terrorist content and misinformation?

Answer. Unlike many other internet companies and social media platforms, Twitter is public by its nature. People come to Twitter to speak publicly, and public Tweets are viewable and searchable by anyone. We are committed to providing a service that fosters and facilitates free and open democratic debate, and we do so by making it possible for people to react to, comment on, and engage with content that they or other accounts choose to share, in accordance with the Twitter Rules.

Twitter employs extensive content detection technology to identify and police harmful and abusive content embedded in various forms of media on the platform. We use PhotoDNA and hash matching technology in the context of child sexual exploitation, and we use proprietary internal technology to identify terrorist accounts, including URL analyses. We use these technologies to identify previously identified content in order to surface it for agent review, and we continually expand our databases of known violative content.

Question 3. Does your company currently share AI training data related to counterterrorism with other social media companies? If not, is there a plan to share such data in the future?

Answer. Machine learning plays an important role across a multitude of our product surface areas. We strive to give our users control and transparency over these, by allowing them to opt out of the algorithmic time line and safe search, for example. Making Twitter more healthy requires making the way we practice machine learning more fair, accountable, and transparent.

To continually advance the state of machine learning, inside and outside Twitter, we are building out a research group at Twitter to focus on a few key strategic areas such as natural language processing, reinforcement learning, machine learning ethics, recommendation systems, and graph deep learning.

Additionally, studying the societal impact of machine learning is a growing area of research in which Twitter has been participating. We are partnering with researchers at the University of California Berkeley to establish a new research initiative focused on studying and improving the performance of ML in social systems, such as Twitter. The team at UC Berkeley will closely collaborate with a corresponding team inside Twitter. As a company, Twitter is able to bring data and real-world insights to the table, but by partnering with UC Berkeley we can create a research program that has the right mix of fundamental and applied research components to make a real practical impact across industry.

Today, the consequences of exposing algorithmic decisions and machine learning models to hundreds of millions of people are poorly understood. Even less is known about how these algorithms might interact with social dynamics: People might change their behaviour in response to what the algorithms recommend to them, and as a result of this shift in behaviour the algorithm itself might change, creating a potentially self-reinforcing feedback loop. We also know that individuals or groups will seek to game or exploit our algorithms and safeguarding against this is essential.

By bringing together the academic expertise of UC Berkeley with our industry perspective, we are looking to do fundamental work in this nascent space and apply it to improve Twitter.

We welcome efforts to increase collaboration in this area, both with industry and governments. The work of the GIFCT to foster technical collaboration will enable us to build on work already done, and policy makers could support these efforts with greater legal protections for companies sharing content of this nature.

Question 4. How is your company working together with other companies to share technology, information, and resources to combat misinformation? Is there an analogue to the Global Internet Forum to Counter Terrorism (GIFCT) for misinformation on your platforms?

Answer. The challenges posed by misinformation are serious and wide-ranging. We are carefully considering how our approach should evolve to respond to the growing range of threats the public conversation faces in this regard. In particular, the public conversation occurring on Twitter is never more important than during elections, the cornerstone of our democracy. Any attempts to undermine the integrity of our service is antithetical to our fundamental rights and undermines the core tenets of freedom of expression.

We remain vigilant about malicious foreign efforts to manipulate and divide people in the United States and throughout the world, including through the use of foreign disinformation campaigns that rely upon the use of deepfakes. In April 2019, we issued a new policy regarding election integrity governing 3 categories of manipulative behavior and content related to elections. First, an individual cannot share false or misleading information about how to participate in an election. This includes but is not limited to misleading information about how to vote or register to vote, requirements for voting, including identification requirements, and the official, announced date, or time of an election. Second, an individual cannot share false or misleading information intended to intimidate or dissuade voters from participating in an election. This includes but is not limited to misleading claims that polling places are closed, that polling has ended, or other misleading information relating to votes not being counted.

We also do not allow misleading claims about police or law enforcement activity related to polling places or elections, long lines, equipment problems, voting procedures or techniques which could dissuade voters from participating in an election, and threats regarding voting locations. Finally, we also do not allow the creation of fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation to a candidate, elected official, political party, electoral authority, or Government entity.

If we see the use of any manipulated content to spread misinformation in violation of our policies governing election integrity, we will remove that content.

The solutions, which will require both product and policy interventions, will need to protect the rights of people to engage in parody, satire, and political commentary while protecting the integrity of the public conversation. As Mr. Pickles testified at the hearing: “We are continuing to explore how we may take action—through both policy and product—on these types of issues in the future. We continue to critically examine additional safeguards we can implement to protect the conversation occurring on Twitter.”

Our existing efforts to make available a comprehensive archive of Tweets and media associated with suspected state-backed information operations we remove from Twitter is also a valuable tool. Our industry peers can leverage the range of signals we publish including links, media, and account indicators. The datatsets we have published include more than 30 million Tweets and more than 1 terabyte of media.

Twitter cannot address these issues alone. The challenges we face as a society are complex, varied, and constantly evolving. Every entity has a role to play—including how the media chooses to cover examples of manipulated media. A whole-of-society approach includes educators and media literacy groups to promote better understanding of these issues. This is a long-term problem requiring a long-term response, not just the removal of content.

Question 5. What is your company doing to ensure that your human content moderators are provided with all the resources they need in order to carry out their jobs, including mental health resources and adequate pay?

Answer. In addition to an increased investment in machine learning, our efforts to improve the health of the public conversation do include global content review teams made up of agency partners. These teams are sometimes exposed to material that is sensitive and potentially distressing in nature.

The well-being of those who review content is a primary concern for our teams and our highest priority is to ensure our staff and contractors are treated with compassion, care, and respect. We are continually evaluating our partners' standards and remain committed to protecting the well-being of the teams tasked with this important and challenging role. We have a full suite of support services available for our employees, including content moderators. As part of our work with third parties, we require in our contracts the provision of support services, including a period of time after an individual changes roles.

In the long term, one of the most valuable investments we can make is in technology and tooling. The more we can leverage these to minimise the exposure to content, the less frequently our employees and contractors will come into contact with it. We will continue to support those engaged in these roles.

Question 6. Prior to introducing new policies or products on your platforms, what processes does your company have in place to anticipate and plan for unintended consequences, harmful side effects, or exploitation by bad actors, including terrorists and those seeking to spread misinformation?

Answer. We draft and enforce the Twitter Rules to keep people safe on our service, and to protect the health of the public conversation. The Twitter Rules apply to everyone. In general, we create our rules with a rigorous policy development process; it involves in-depth research, analysis of the behavior of individuals on Twitter and historical violation patterns, and immersion in academic material.

We appreciate these issues are complex, and we value the input of external voices in developing our approach. As part of the internal development process, we consult with a wide range of stakeholders and we focus consideration regarding the risk of gaming, subverting, or otherwise abusing our policies and product changes. We supplement this work with conversations with outside experts and organizations where appropriate.

For example, many scholars have examined the relationship between dehumanization and violence. In September 2018, we tried something new by asking the public for feedback on a policy before it became part of the Twitter Rules. Our goal is to test a new format for policy development whereby the individuals who use Twitter have a role in directly shaping our efforts to protect them. We wanted to expand our hateful conduct policy to include content that dehumanizes others based on their membership in an identifiable group, even when the material does not include a direct target.

We asked for feedback to ensure we considered a wide range of perspectives and to hear directly from the different communities and cultures who use Twitter around the globe. In 2 weeks, we received more than 8,000 responses from people located in more than 30 countries.

Following our review of public comments, in July 2019, we expanded our rules against hateful conduct to include language that dehumanizes others on the basis of religion.

Questions From Ranking Member Mike Rogers for Nick Pickles

Question 1. During the hearing, Professor Strossen raised a number of concerns about the dangers of moderating speech. Rather than censorship, she offered a number of suggestions to empower users. What, if anything, are your companies doing to provide more filtering tools to enhance the ability of users to control the content they see?

Answer. Twitter provides a variety of tools to individuals on our service to enable them to control the content they see. At the most basic level, individuals on Twitter control the content they see by choosing which accounts to follow. They can unfollow an account at any time, or choose to receive a notification for every Tweet an account sends.

We also enable individuals on the service to control their individual experience through tools such as the ability to block and mute. If an individual has reported a Tweet, we will hide it behind a notice and give the individual the choice on whether or not he or she want to view the content again. We will also hide content behind an interstitial if an individual has muted or blocked an account and their Tweets are shared by someone else.

Individuals may also mute a conversation they do not wish to be a part of, or mute a specific word, hashtag, or phrase. The individual can control how long this stays in place.

We also offer a range of advanced filters for notifications that individuals on Twitter can customize. This includes the option to hide notifications from accounts that an individual does not follow or who do not follow the individual, from those that have not confirmed an email address or phone number, those who have not set a profile photograph, or from all new accounts.

We also give people control over what they see in search results through a "Safe Search" option. This option excludes potentially sensitive content from search results, such as spam, adult content, and the accounts an individual has muted or blocked. Individual accounts may mark their own posts as sensitive as well. Twitter's safe search mode excludes potentially sensitive content, along with accounts an individual may have muted or blocked, from search results. Safe Search is enabled by default, and people have the option to turn safe search off, or back on, at any time.

In December 2018, Twitter introduced a sparkle icon located at the top of individuals' time lines to more easily switch on and off reverse chronological time line, allowing them to view tweets without algorithmic ranking. As described above, the algorithms we employ are designed to help people see the most relevant Tweets. The icon now allows individuals using Twitter to easily switch to chronological order ranking of the Tweets from only those accounts they follow. This improvement allows individuals on Twitter to see how algorithms affect what they see, and enables greater transparency into the technology we use to rank Tweets.

We additionally empower individuals on the service to control their experiences through notices, or instititals. Our systems and teams may add notices on Tweets to give individuals on the service more context or notice before an individual on Twitter clicks on the Tweet. Twitter may add a notice to an account or Tweet to provide more context on the actions our systems or teams may take. In some instances, this is because the behavior violates the Twitter Rules. We may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through the filter. This allows us to identify potentially sensitive content that some people may not wish to see.

Question 2. Do you have recommendations for ways to more effectively address the extremist content found on many of the off-shore, fringe social media sites?

Answer. Although Twitter strictly enforces our policies removing terrorist and extremist content that violates our Rules, it does not eliminate the ideology underpinning them. Quite often, it moves these views into darker corners of the internet where they cannot be challenged and held to account. As Twitter and our peer companies improve in our efforts, this content continues to migrate to less public and more private platforms and messaging services. We are committed to learning and improving, but every part of the on-line ecosystem has a part to play.

There are a range of approaches policy makers could consider. Broadening the range of companies who are part of the discussion is essential if we are to form a robust view on how to tackle these issues. Widening the discussion will also bring important perspectives to the fore about the nature and challenges of content moderation at scale. The role of financial incentives is also useful to consider. For example, a recent report from the Global Disinformation Index project focused on the ways the on-line ecosystem is being abused by bad actors to monetize disinformation. The same may be true of the monetization of terrorist content on some parts of the on-line ecosystem. See Global Disinformation Project, *Cutting the Funding of Disinformation: The Ad-Tech Solution* (May 2019) (on-line at *https://disinformationindex.org/wp-content/uploads/2019/05/GDI_Report_Screen_AW-2.pdf).*

We acknowledge that we have a role to play and acknowledge that we will never reach a point where we are finished tackling these issues. Tech companies and content removal on-line cannot alone, however, solve these issues. They are systemic and societal and as such they require an whole-of-society approach. We welcome the opportunity to continue to work with our industry peers, Government, academics, and civil society to find the right solutions.

Question 3. Can you describe the appeals process within your platform for users to challenge content removal decisions? How quickly does this process occur and how do you incorporate lessons learned when your company reverses a removal decision?

Answer. Content moderation on a global scale is a new challenge not only for our company, but also across our industry. When an action is taken in error, we act promptly to correct them. We now offer people who use Twitter the ability to more easily file an appeal from within the Twitter app when we tell them which Tweet has broken our rules. This makes the appeal process quicker and easier for users. We anticipate this new process will enable us to respond 60 percent faster to appeals.

We also allow individuals to file a report through a web form that can be accessed at *http://help.twitter.com/appeals.* We also continue to improve our transparency around the actions we take, including better in-app notices where we have removed Tweets for breaking our rules. We also communicate with both the account who reports a Tweet and the account which posted it with additional detail on our actions. These steps are all a part of our continued commitment to transparency, and we will continue to better inform individuals who use Twitter on our work in these areas.

If an account was suspended or locked in error, an individual can appeal. First, the individual must log in to the account that is suspended and file an appeal. The individual must describe the nature of the appeal and provide an explanation of why the account is not in violation of the Twitter Rules. Twitter employees will engage with the account holder via email to resolve the suspension.

Question 4. Moderating terror and extremist content on social media platforms is a complex issue with no perfect solution. One consistent recommendation the committee has received from a variety of outside experts is the value of greater transparency in your respective content removal policies. What is your response to calls for you to open up your platforms for academics for research purposes, particularly allowing them to review the content you remove?

Answer. In regard to the removal of accounts, our biannual Twitter Transparency Report highlights trends in enforcement of our Rules, legal requests, intellectual property-related requests, and email privacy best practices. The report also provides insight into whether or not we take action on these requests. The Transparency Report includes information requests from governments world-wide and non-government legal requests we have received for account information. Removal requests are also included in the Transparency Report and include world-wide legal demands from governments and other authorized reporters, as well as reports based on local laws from trusted reporters and non-governmental organizations, to remove or withhold content.

As part of our commitment to educate users about our rules and to further prohibit the promotion of terrorism or violent extremist groups, we have updated our rules and associated materials to be clearer on where these policies apply. We agree that our rules should be clear and understandable. Recently we completed a process to refresh our rules and ensure that they are easier to understand. This includes each specific rule being short enough to Tweet.

In addition, we have improved the supporting information in our help center, which adds context and examples to the Rules. With regard to terrorism and violent extremism, there is a dedicated page in our help center accessed at *https://help.twitter.com/en/rules-and-policies/violent-groups.*

We have now suspended more than 1.5 million accounts for violations related to the promotion of terrorism between August 1, 2015, and December 31, 2018. In 2018, a total of 371,669 accounts were suspended for violations related to promotion of terrorism. We continue to see more than 90 percent of these accounts suspended through proactive measures.

With regard to academic access, Twitter is a uniquely open service. The overwhelming majority of content posted is publicly available and made available through our free public and commercial application programming interfaces ("APIs"). We make public Tweet data available to Twitter users, developers, researchers, and other third parties. We encourage developers and others to create products using this public data for purposes that serve the public interest and the general Twitter community.

QUESTIONS FROM CHAIRMAN BENNIE G. THOMPSON FOR DEREK SLATER

Question 1. Does your company currently make data on on-line terror content and misinformation—including the amount and types of content you remove—available for academics and other stakeholders to research? In what ways or what partnerships? Will you commit to making such data available to researchers?

Answer. Response was not received at the time of publication.

Question 2. Private posts containing objectionable content pose a unique challenge for moderation. How does your company reconcile data privacy with the challenges of moderating content that violates your terms of service, including terrorist content and misinformation?

Answer. Response was not received at the time of publication.

Question 3. Does your company currently share AI training data related to counterterrorism with other social media companies? If not, is there a plan to share such data in the future?

Answer. Response was not received at the time of publication.

Question 4. How is your company working together with other companies to share technology, information, and resources to combat misinformation? Is there an analogue to the Global Internet Forum to Counter Terrorism (GIFCT) for misinformation on your platforms?

Answer. Response was not received at the time of publication.

Question 5. What is your company doing to ensure that your human content moderators are provided with all the resources they need in order to carry out their jobs, including mental health resources and adequate pay?

Answer. Response was not received at the time of publication.

Question 6. Prior to introducing new policies or products on your platforms, what processes does your company have in place to anticipate and plan for unintended consequences, harmful side effects, or exploitation by bad actors, including terrorists and those seeking to spread misinformation?

Answer. Response was not received at the time of publication.

QUESTIONS FROM HONORABLE LAUREN UNDERWOOD FOR DEREK SLATER

Question 1a. During the hearing, you committed to providing in writing information on what percentage of users who view YouTube videos with information cues actually click on the link for more information.

Please provide this information, broken down by each month since YouTube's CEO announced the policy in March 2018.

Answer. Response was not received at the time of publication.

Question 1b. During the hearing, you stated that information cues link to an online encyclopedia, in addition to Wikipedia. Please provide a complete list, as well as a breakdown by percentage, of the websites that YouTube's information cues link to.

Answer. Response was not received at the time of publication.

Question 2a. Does YouTube vet the Wikipedia articles that it links to in "information cues" to ensure their accuracy, or work with Wikipedia to ensure that the articles are locked against malicious edits?

Answer. Response was not received at the time of publication.

Question 2b. During the hearing, you stated that YouTube "has a process to make sure that [YouTube is] displaying accurate information" in information cues. Please provide a detailed description of that process.

Answer. Response was not received at the time of publication.

QUESTIONS FROM RANKING MEMBER MIKE ROGERS FOR DEREK SLATER

Question 1. During the hearing, Professor Strossen raised a number of concerns about the dangers of moderating speech. Rather than censorship, she offered a number of suggestions to empower users. What, if anything, are your companies doing to provide more filtering tools to enhance the ability of users to control the content they see?

Answer. Response was not received at the time of publication.

Question 2. Do you have recommendations for ways to more effectively address the extremist content found on many of the off-shore, fringe social media sites?

Answer. Response was not received at the time of publication.

Question 3. Can you describe the appeals process within your platform for users to challenge content removal decisions? How quickly does this process occur and how do you incorporate lessons learned when your company reverses a removal decision?

Answer. Response was not received at the time of publication.

Question 4. Moderating terror and extremist content on social media platforms is a complex issue with no perfect solution. One consistent recommendation the committee has received from a variety of outside experts is the value of greater transparency in your respective content removal policies. What is your response to calls for you to open up your platforms for academics for research purposes, particularly allowing them to review the content you remove?

Answer. Response was not received at the time of publication.

Question From Ranking Member Mike Rogers for Nadine Strossen

Question. During the hearing, you highlighted the importance of transparency by mainstream social media companies. Can you provide more: (1) Background on the importance of transparency and (2) recommendations for how transparency measures could be implemented by these companies? Additionally, do you have (3) recommendations for transparency measures for the Global Internet Forum to Counter Terrorism (GIFCT)?

Answer.

(1) Background on the importance of transparency

In enforcing their "content moderation" policies, determining what speakers and speech will be allowed on their platforms—and which will not be allowed—the dominant social media companies ("companies") exercise vast censorial power that even exceeds the scope of censorial power that in the past only Government has wielded. However, as private-sector entities, these companies are not subject to the Constitutional constraints that limit Government power, including requirements to respect not only freedom of speech and press, but also privacy, due process/fair procedures, and equality. This means that the companies may also exercise their vast power in ways that undermine our democratic self-government. For example, they could discriminatorily suppress certain speakers or ideas based on partisan political preferences, or they could promote disinformation about political candidates and public policy issues.

For these reasons, it is essential that steps be taken to curb the companies' powers to suppress users' freedoms. However, it is also essential that any such steps are respectful of the companies' own rights and freedoms. For example, the companies' own free speech rights would be infringed by Government regulations dictating what speech they could or could not permit on their platforms—to the same extent that such Government regulations would infringe on the free speech rights of more traditional media, such as newspapers and broadcasters.

Recognizing these countervailing free speech concerns, many experts who have studied these issues have concurred that, as a baseline minimum approach for protecting the rights of platform users—and associated democratic concerns—the companies should design and implement their content moderation policies in a manner that complies with certain basic process standards, including transparency. Concerning transparency in particular, the companies should disclose what their content moderation policies are, and how they are enforced, both in the aggregate and in particular cases. The companies should not only provide information about how their content moderation policies are enforced in general, or in particular categories (as explained further in the next section), but they should also provide information to individual users whose content is removed. (This kind of individualized disclosure/transparency is often referred to as "notice," invoking a fundamental due process/fairness concept.)

Government officials could and should use their "bully pulpit" to encourage companies to provide aggregate and individualized information consistent with the goal of transparency; civil society organizations and the companies' customers should do likewise. In fact, many Government officials, civil society organizations, and individual experts—not only in the United States, but also in other countries and in international agencies—have advocated such transparency. A number of the companies have endorsed at least the general concept of enhanced transparency, and have undertaken some compliance efforts.

As New America's Open Technology Institute has commented, the companies' disclosure of data about from whom they get takedown requests, and how they respond to such requests, "is a vital first step toward enabling the public to hold [them] accountable in their roles as arbiters of speech, and also hold accountable the governments and private parties that seek to censor on-line speech."[1] Moreover, this re-

[1] *https://www.newamerica.org/oti/blog/announcing-otis-new-transparency-reporting-toolkit-focused-on-content-takedowns/.*

porting "helps the public identify where they think the [c]ompanies are doing too much—or not enough—to address content issues," and also benefits the Companies, including by "helping to build trust with their users and policy makers."[2]

(2) Recommendations for how transparency measures could be implemented

Transparency about overall content takedown practices

The Santa Clara Principles (2018)

In May 2018, a group of expert organizations and individuals issued "the Santa Clara Principles," which they described "as initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights."[3] Since then, the Santa Clara principles have been endorsed—at least in spirit—by diverse experts, including the United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, and some of the Companies.

In terms of transparency, these Principles declared that "companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines." The Principles further specified that, at a minimum, "this information should be broken down along each of these dimensions":

- Total number of discrete posts and accounts flagged
- Total number of discrete posts removed and accounts suspended
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by category of rule violated
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by format of content at issue (e.g., text, audio, image, video, live stream)
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by source of flag (e.g., governments, trusted flaggers, users, different types of automated detection); and
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent).

Finally, the Santa Clara Principles called for this data to "be provided in a regular report, ideally quarterly, in an openly licensed, machine-readable format."

The Transparency Reporting Toolkit: Content Takedown Reporting (2018)

Especially detailed recommendations for transparency reporting have been provided by the Open Technology Institute and Harvard University's Berkman Klein Center for Internet & Society. In 2016, both organizations released the "Transparency Reporting Toolkit," which "aimed to make it easier for companies to create and improve their transparency reports around government demands for user data and to make transparency reporting more consistent, easier to understand and more effective."[4] In October 2018, the Open Technology Institute issued a new transparency reporting toolkit that focused expressly on content takedowns ("2018 Toolkit"). Based on extensive consultations with a broad, diverse array of companies and civil society experts, the 2018 Toolkit identified general "best practices" for content takedown reporting regarding any kind of content, as well as additional, more specific best practices for reporting about certain types of content takedown (copyright-related, network shutdowns and service interruptions, and "right to be forgotten" delistings).

Below I will list the general best practices that the 2018 Toolkit recommends, all of which it explains in more detail:

- Issuing reports for clearly and consistently delineated reporting periods
- Issuing reports specific to the type of demand
- Reporting on types of demands using specific numbers
- Breaking down demands by country
- Reporting on categories of objectionable content targeted by demands
- Reporting on products targeted by demands
- Reporting on specific government agencies/parties that submitted demands
- Specifying which laws pertain to specific demands

[2] Spandana Singh & Kevin Bankston, "The Transparency Reporting Toolkit: Content Takedown Reporting," *New America,* Oct. 2018, at 6. *https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/*

[3] *https://santaclaraprinciples.org/.*

[4] *https://www.newamerica.org/oti/policy-papers/transparency-reporting-toolkit-reporting-guide-and-template/.*

- Reporting on the number of accounts and items specified in demands
- Reporting on the number of accounts and items impacted by demands; and
- Reporting on how the company responded to demands.

The foregoing best practices focus on quantitative transparency. The 2018 Toolkit also discussed some additional best practices that seek to improve the qualitative transparency surrounding content takedowns. These include:

- Defining terms clearly
- Providing meaningful explanations of internal policies
- Offering case studies to illustrate the company's practices and the issues it faces
- Reporting on specific notices where reasonable and permitted by law
- Providing meaningful numbers that reflect how many pieces of content or accounts were taken down, blocked or otherwise restricted based on automated flagging or review
- Linking relevant reports to one another
- Publishing reports at static and functioning URLs
- Publishing data in a structured data format
- Publishing reports using a non-restrictive Creative Commons license; and
- Offering a Frequently Asked Questions section for the report.

Transparency about takedown of particular content: notice

The Santa Clara Principles also set out basic recommendations for implementing this fundamental, individualized facet of transparency. First:

"In general companies should provide detailed guidance to the community about what content is prohibited, including examples of permissible and impermissible content and the guidelines used by reviewers. Companies should also provide an explanation of how automated detection is used across each category of content."

Additionally, "[c]ompanies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension." This required notice to individual users must include the following details, at a minimum:

- URL, content excerpt, and/or other information sufficient to allow identification of the content removed
- The specific clause of the guidelines that the content was found to violate
- How the content was detected and removed (flagged by other users, governments, trusted flaggers, automated detection, or external legal or other complaint); and
- Explanation of the process through which the user can appeal the decision.

Moreover, the Principles provide that "[t]he identity of individual flaggers should generally not be revealed, however, content flagged by government should be identified as such, unless prohibited by law." Finally, they specify:

"Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. Users who flag content should also be presented with a log of content they have reported and the outcomes of moderation processes."

[3] Recommendations for transparency measures for the Global Internet Forum to Counter Terrorism (GIFCT)

Before laying out the recommendations for transparency measures for GIFCT, which have been urgently called for by a wide array of experts, I will briefly summarize GIFCT's operations. GIFCT was formed by Facebook, Microsoft, Twitter, and YouTube, and publicly announced in 2016. It created a hash database ("the database") that contains digital hash "fingerprints" of images and videos that the participants have identified as "extreme terrorist material," based on their own internal content moderation standards—not based on any legal definition. The participating companies then use automated filtering tools to identify and remove duplicates of the hashed images or videos.

GIFCT raises serious human rights problems, as well as serious questions about its efficacy in countering terrorism; none of these important issues can be definitively evaluated because of the lack of transparency about GIFCT's operations. Accordingly, it is imperative that the companies disclose sufficient information to facilitate assessment of GIFCT's costs and benefits in terms of both countering terrorism and respecting freedom of speech and other human rights.

The foregoing critiques of GIFCT's lack of transparency have been made by multiple, diverse observers, including a large group of expert organizations and individuals from many different countries, in a joint February 4, 2019 letter to the European Parliament. Underscoring their shared commitment to the goal of countering terrorist violence, and not questioning GIFCT operators' positive intent to promote that goal, these experts stressed that "lawmakers and the public have no meaningful information about how well" the database actually "serves this goal, and at what cost to democratic values and individual human rights."

I will quote here some of this letter's key points about needed transparency for meaningful evaluation and accountability:

> "Almost nothing is publicly known about the specific content that platforms block using the database, or about companies' internal processes or error rates, and there is insufficient clarity around the participating companies' definitions of 'terrorist content.' Furthermore, there are no reports about how many legal processes or investigations were opened after the content was blocked. This data would be crucial to understand to what extent the measures are effective and necessary in a democratic society, which are some of the sine qua non requisites for restriction of fundamental rights."

This letter noted some well-publicized failures of algorithmic removals of alleged "terrorist content" that actually constituted important material "for news reporting, combating terrorist recruitment on-line, or scholarship," because algorithmic filters "are blind to . . . contextual differences" between material with otherwise similar "terrorist content." However, among the information that has not yet been disclosed is "whether major platforms like YouTube or Facebook adequately correct for" such problems "through employees' review of filtering decisions."

The letter urged the European Parliament not to adopt any regulation that would incorporate GIFCT precisely because of the absence of transparency:

> "The European public is being asked to rely on claims by platforms or vendors about the efficacy of the database . . . or else to assume that any current problems will be solved by hypothetical future technologies or untested, post-removal appeal mechanisms. Such optimistic assumptions cannot be justified given the serious problems researchers have found with the few filtering tools available for independent review."

CONCLUSION

Members of Congress and others cannot meaningfully assess the impact of the companies' efforts to counter on-line terrorist content (including through GIFCT), misinformation, or any other controversial, potentially problematic content, in the absence of detailed information about the companies' content moderation policies. In particular, policy makers and the public cannot assess either: (i) How effective such efforts are in reducing the targeted content, or (ii) how much legitimate, even valuable content is also removed in the process. In short, no meaningful cost-benefit analysis can be done of the aggregate results of content moderation policies without much more information, of the sort I have outlined. Likewise, individual users whose expression has been suppressed cannot exercise the important right to appeal such suppression without detailed information of the sort also laid out. Members of Congress, as well as other public officials, NGO's, and the companies' customers should all continue to advocate for such increased transparency.