ARTIFICIAL INTELLIGENCE AND COUNTERTERRORISM: POSSIBILITIES AND LIMITATIONS

HEARING

BEFORE THE

SUBCOMMITTEE ON INTELLIGENCE AND COUNTERTERRORISM

OF THE

COMMITTEE ON HOMELAND SECURITY HOUSE OF REPRESENTATIVES

ONE HUNDRED SIXTEENTH CONGRESS

FIRST SESSION

JUNE 25, 2019

Serial No. 116–28

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
38–781 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*

SUBCOMMITTEE ON INTELLIGENCE AND COUNTERTERRORISM

MAX ROSE, New York, *Chairman*

SHEILA JACKSON LEE, Texas
JAMES R. LANGEVIN, Rhode Island
ELISSA SLOTKIN, Michigan
BENNIE G. THOMPSON, Mississippi *(ex officio)*

MARK WALKER, North Carolina, *Ranking Member*
PETER T. KING, New York
MARK GREEN, Tennessee
MIKE ROGERS, Alabama *(ex officio)*

VACANCY, *Subcommittee Staff Director*
MANDY BOWERS, *Minority Subcommittee Staff Director*

CONTENTS

ARTIFICIAL INTELLIGENCE AND COUNTERTERRORISM: POSSIBILITIES AND LIMITATIONS

Tuesday, June 25, 2019

U.S. HOUSE OF REPRESENTATIVES,
COMMITTEE ON HOMELAND SECURITY,
SUBCOMMITTEE ON INTELLIGENCE
AND COUNTERTERRORISM,
Washington, DC.

The subcommittee met, pursuant to notice, at 10 a.m., in room 310, Cannon House Office Building, Hon. Max Rose (Chairman of the subcommittee) presiding.

Present: Representatives Rose, Jackson Lee, Langevin, Thompson (ex officio), Walker, and Green.

Mr. ROSE. The Subcommittee on Intelligence and Counterterrorism will come to order.

Good morning, everyone. Thank you so much for being here. Today, the Subcommittee on Intelligence and Counterterrorism is meeting to examine the role of artificial intelligence, or AI, in addressing counterterrorism content on social media platforms.

We all know that AI can perform a myriad of tasks, complex and otherwise. It is everywhere in the news. The issue, though, that we are looking to address today, though, the question that we are looking to address today is very simple, and that is, what can AI do and what can AI not do as it pertains to counterterrorist screening? Because we are hearing the same thing from social media companies, and that is, AI's got this. It is only going to get better.

We take down 99 percent of content, hundreds of thousands, millions of pieces of content due to our superior AI platforms. But nonetheless, though, we have seen egregious problems with counterterrorist screening on social media platforms.

On March 15, a white supremacist extremist opened up fire at two mosques in Christchurch, New Zealand, killing 51 people and wounding 49 more. Shockingly, the terrorist was able to live stream the attack on Facebook because Facebook's AI did not deem the footage gruesome enough. They had never seen it before.

The video was then uploaded to Facebook by other users, and 300,000 of these attempts made it through, proving that their technology is not yet up to the task. In fact, instead of preventing terrorist content from spreading, the Associated Press recently reported that Facebook's AI was making videos of and promoting the terrorist content it should have been removing.

I hope our witnesses today will help us better understand the current state of AI and its limitations, capabilities, and future promise, especially as it relates to countering on-line terrorist content.

We are receiving wake-up calls left and right about this problem. Two years ago, the big tech companies, led by Facebook, Twitter, Google, Microsoft, got together and they formed the Global Internet Forum to Counterterrorism, or the GIFCT, to share best practices and certain best technologies to combat the spread of on-line terrorist content.

While the GIFCT touts impressive numbers, recent reporting and persistent lack of transparency from the tech companies has raised fundamental questions about the effectiveness of AI and other technologies at identifying terror content. Let's be—in the plainest of language, the GIFCT, from everything I have seen to date, is a joke.

There is no full-time employee. There is no brick and mortar. They call it an association, but there are much smaller associations that have dedicated far more resources, and this is a classic collective action problem.

So we have been looking at this, myself and Ranking Member Walker, as well as the rest of the committee, we have been looking at this problem for months now, and we have been approached by the social media companies with this libertarian, technocratic elitism that is highly, highly disturbing, and it centers around them claiming that AI can accomplish everything. And as a consequence, when we ask them, how many people have you hired, what kind of resources have you dedicated to this problem, they will not give us a straight answer because, again, they refer to AI. They say you can hire 50,000 people, but why do that when we have AI.

So today, we want to get to the root of whether that is a legitimate response or not or, to the contrary, whether tech firms and social media platforms are putting us all at risk via their wanton disregard for their National security obligations.

There has been a frustrating lack of transparency amongst these social media companies. They have to do better, and we as a Congress must do more to hold them accountable. As I said, our National security is at stake.

I thank the witnesses and Members for being here, and I look forward to making progress on this important issue.

[The statement of Chairman Rose follows:]

STATEMENT OF CHAIRMAN MAX ROSE

JUNE 25, 2019

AI can perform a myriad of complex tasks that formerly required a human being. Social media companies use AI to help identify and remove terrorist content and materials that violate their terms of service, so far with mixed results at best. But we've seen in gruesome detail the failures which serve as a critical reminder that AI is not up to the task. On March 15, a white supremacist extremist opened fire at 2 mosques in Christchurch, New Zealand, killing 51 people and wounding 49 more. Shockingly, the terrorist was able to live-stream the attack on Facebook because its Artificial Intelligence, or AI, did not deem the footage gruesome enough. The video was then uploaded to Facebook by other users and 300,000 of these attempts made it through—proving that their technology is not yet up to the task.

In fact, instead of preventing terrorist content from spreading, the Associated Press recently reported that Facebook's AI was making videos of and promoting the

terrorist content it should have been removing. I hope our witnesses will help us better understand the current state of AI and its limitations, capabilities, and future promise, especially as it relates to countering on-line terrorist content. This incident is a wake-up that not enough is being done either through technology or human moderators to protect us from terrorist threats on social media—including terrorists using these platforms to recruit, plan, and broadcast their attacks. Two years ago, the big tech companies—led by Facebook, Twitter, Google, and Microsoft—got together to form the Global Internet Forum to Counter Terrorism, or GIFCT, to share best practices and certain basic technologies to combat the spread of on-line terrorist content.

While the GIFCT touts impressive numbers in removing terrorist content automatically, recent reporting and persistent lack of transparency from the tech companies have raised fundamental questions about the effectiveness of AI and other technologies to at identifying terror content. I come to this hearing with an open mind and a willingness to work with social media companies to do what is right, but I have been disappointed so far. There has been a frustrating lack of transparency from the social media companies about their efforts to address terror content on their platforms.

Weeks ago, I wrote asking about their personnel and resources committed to this important effort, and I have yet to receive satisfactory answers. They must do better. We, as Congress, must do more to hold them accountable. Our National security is at stake.

Mr. ROSE. I now recognize the Ranking Member of the subcommittee, Mr. Walker, for an opening statement.

Mr. WALKER. I want to thank Chairman Rose for holding this hearing today. I look forward to hearing from our distinguished panel. I believe we have got Georgetown and Stanford, at least, represented today. Certainly want to talk about the limitations in utilizing artificial intelligence to monitor on-line extremist content.

The ingenuity and superiority of the United States' private sector continues to drive the development of new technologies, products, and services that really revolutionize the world. The development of AI is another example, and we have only seen the beginning of what this technology can do.

AI has the potential to address a variety of major global issues, and research and developmental is happening across all sectors. U.S. educational institutions, including those in my home State of North Carolina, are leading cutting-edge researches into health care, pharmaceuticals, transportation, data science, and many more fields.

Today, we are primarily reviewing how the technology is used by social media companies to operate their platforms and identify the exact content that needs to be removed. It is clear that technology is not a silver bullet for identifying and removing extremist content, given the volume of content uploaded every second on social media platforms. AI technology in its current form is limited and cannot currently evaluate context when reviewing content.

For example, there have been a number of notable examples, one just mentioned in the past few years, where AI has flagged portions of even, get this, the Declaration of Independence and removed historical images from media reports. We must also be mindful that algorithms and content moderation policies are ultimately subjective as they are developed and operated by humans who possess sometimes their own bias.

As legislators, we must proceed with caution on the appropriate role for Congress in this situation, understanding the potential to stymie free speech. We must also recognize that social media companies themselves have a First Amendment right to host, to de-

velop, and modify their terms of service and content moderation policies to foster an open and free space for expression. We have come to a crossroads in the debate on what content should be prohibited from social media platforms and the appropriate mechanisms to identify and remove such content.

Today's hearing will help us to further our understanding of the current capabilities of AI technology and receive recommendations on what more the social media companies could be doing regarding the application of AI relating to the content moderation.

So at a minimum, we need to discuss the continually-changing terms of service implemented by many of the companies and the need for greater transparency in how they are making content removal decisions, not only to the individual users, but also to the community as a whole and at large.

I look forward to the testimony, and I want to thank the witnesses for appearing here today.

I yield back the balance of my time, Mr. Chairman.

[The statement of Ranking Member Walker follows:]

STATEMENT OF RANKING MEMBER MARK WALKER

JUNE 25, 2019

I want to thank Chairman Rose for holding this hearing today. I look forward to hearing from our distinguished panel on the capabilities and limitations in utilizing artificial intelligence, or AI, to monitor on-line extremist content.

The ingenuity and superiority of the U.S. private sector continues to drive the development of new technologies, products, and services that have revolutionized the world. The development of AI is another example and we have only seen the beginning of what this technology can do.

AI has the potential to address a variety of major global issues, and research and development is happening across all sectors. U.S. educational institutions, including those in my home State of North Carolina, are leading cutting-edge research into health care, pharmaceuticals, transportation, data science, and many more fields.

Today, we are primarily reviewing how the technology is used by social media companies to operate their platforms and identify content that may need to be removed.

It is clear that technology is not a silver bullet for identifying and removing extremist content, given the volume of content uploaded every second on social media platforms.

AI technology, in its current form, is limited and cannot currently evaluate context when reviewing content. For example, there have been a number of notable examples in the past few years where AI has flagged portions of the Declaration of Independence and removed historical images from media reports. We must also be mindful that algorithms and content moderation policies are ultimately subjective, as they are developed and operated by humans who possess their own bias.

As legislators, we must proceed with caution on the appropriate role for Congress in this situation, understanding the potential to stymie free speech.

We also must recognize that the social media companies themselves have a First Amendment right to host, develop, and modify their terms of service and content moderation policies to foster an open and free space for expression.

We have come to a crossroads in the debate on what content should be prohibited from social media platforms and the appropriate mechanisms to identify and remove such content. Today's hearing will help us to further our understanding of the current capabilities of AI technology and receive recommendations on what more the social media companies could be doing regarding the application of AI relating to content moderation.

At a minimum, we need to discuss the continually-changing terms of service implemented by many of the companies and the need for greater transparency in how they are making content removal decisions, not only to the individual users, but also to the community as a whole.

I look forward to the testimony and I want to thank the witnesses for appearing here today. I yield back the balance of my time.

Mr. ROSE. Thank you, Ranking Member.

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

[The statement of Chairman Thompson follows:]

STATEMENT OF CHAIRMAN BENNIE G. THOMPSON

JUNE 25, 2019

In March, a white supremacist terrorist in Christchurch, New Zealand, exploited social media platforms to live-stream violent images across the world—millions of times over. Technology such as Artificial Intelligence is one tool social media companies use to help identify, monitor, and remove such terrorist content. We are witnessing a new technological age with Artificial Intelligence or "AI." Computer systems are increasingly able to perform tasks that previously required human intelligence.

Over time, this capability will only be refined and as the technology improves. We are here today to understand the technological possibilities and current limitations of AI when it comes to countering terrorism on-line. Congress has a responsibility not only to track the progression of these emerging technological breakthroughs, but to understand how they affect our security. The individuals here today represent some of the brightest minds in the AI field. I hope to hear from our witnesses about whether the technology can accurately flag terror content or other material that violates terms of service without unduly impeding the flow of legitimate content on-line.

I also hope to hear about where the technology is still lacking, and whether the social media companies are working to improve its effectiveness on their platforms. Today's hearing lays an important foundation for our full committee hearing tomorrow, where we will engage social media companies about the challenges they face in addressing terror content and misinformation on their platforms.

Mr. ROSE. I welcome our panel of witnesses. Our first witness is Mr. Alex Stamos, adjunct professor at the Freeman Spogli Institute. Prior to this position, Mr. Stamos served as the chief security officer at Facebook. In this role, he led a team of engineers, researchers, investigators, and analysts charged with understanding and mitigating information security risk to the company and safety risk to the 2.5 billion people on Facebook, Instagram, and WhatsApp.

Next, we are joined by Mr. Ben Buchanan, an assistant teaching professor at Georgetown University and senior faculty fellow with the Center for Security and Emerging Technology. Previously, he has written journal articles and peer-reviewed papers on artificial intelligence, attributing cyber attacks, deterrence in cyber operations, cryptography, elections, cybersecurity, and the spread of malicious code between nations and non-state actors.

Finally, we have Mr. Julian Sanchez, a senior fellow with the Cato Institute, where he studies issues at the intersection of technology, privacy, and civil liberties, with a particular focus on National security and intelligence surveillance. Previously, he served as the Washington editor for the technology news site Ars Technica, where he covered surveillance, intellectual property, and telecom policy.

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

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

STATEMENT OF ALEX STAMOS, ADJUNCT PROFESSOR, FREEMAN SPOGLI INSTITUTE, PROGRAM DIRECTOR, STANFORD INTERNET OBSERVATORY, ENCINA HALL

Mr. STAMOS. Good morning, Chairman Rose, Ranking Member Walker. Thank you very much for this opportunity to discuss the potential uses and limitations of artificial intelligence and on-line counterterrorism enforcement. My name is Alex Stamos. I am currently the director of Stanford Internet Observatory, which is a program of the Stanford University Cyber Policy Center. Our group is performing cross-disciplinary research into the misuse of the internet, with a goal of providing actionable solutions for tech companies and governments.

I am also a William J. Perry fellow at the Center for International Security and Cooperation, a visiting scholar at the Hoover Institution, a member of the NATO Cybersecurity Center, and a member of the Annan Commission on Elections and Democracy.

As you said, before joining Stanford, I was the chief security officer at Facebook from June 2015 to August 2018. During that time, I witnessed the company's battle against on-line terrorism, built and supervised a counterterrorism investigations unit, and oversaw the company's research into Russian attacks against Democratic elections in the United States. Previously, I was the chief information security officer at Yahoo and the co-founder of iSEC Partners, which is a technical security consultancy.

I am honored to be here today, and I hope that my experience will help clarify some of the misunderstandings, confusion, and hype about the potential of artificial intelligence that is currently circulating in media reports and policy discussions, particularly as it relates to the issue of counterterrorism and on-line safety.

I have submitted written testimony that goes into these issues in much greater detail than I can cover in 5 minutes, and I thank you for submitting that into the record.

As someone who has seen some of the world's best machine learning experts try to apply these techniques to real-world problems, I am convinced that both the promise and the peril are often exaggerated, making it difficult to have an honest and accurate discussion about the policy implications. The capabilities of these techniques are overstated by tech executives looking for easy answers to difficult problems, start-up founders who need venture capital, and media outlets who lack the adequate technical expertise to properly kick the tires on wild claims.

If we want to accurately assess the impact of artificial intelligence and the policy regime that should accompany it, we need to be disciplined in defining both its challenges and its capability. Today, I will use the more appropriate term "machine learning" as much as possible instead of artificial intelligence, because as we are going to discuss, there is a lot more that is artificial than intelligent about even the state-of-the-art today.

The world's best machine learning resembles a crowd of millions of preschoolers. There are certainly problems for which having a humongous group of children could be taught to solve. Imagine having to take a mountain of Skittles and sort them into five mountains based upon color. That is something that millions of preschoolers could help you with, but adding more preschoolers to

that group would speed up the task but would not allow them to do more complicated tasks.

No number of preschoolers could get together to build the Taj Mahal or explain to you the plot of Ulysses. Similarly, modern machine learning can be incredibly powerful for accomplishing routine tasks at amazing speed and scale, but these technologies are primitive and very fragile. Decision making based upon societal values and cultural context is completely beyond current capabilities.

One important thing to understand about modern machine learning is that most of the practical techniques in use today can't be told what they are supposed to look for; they have to be shown. Most of the algorithms relevant to our discussion today are known as classifiers.

Classifiers are systems that sort digital information into various categories. A classifier is generally trained by feeding the data that has already been labeled by humans, preferably large data sets that represent the diversity of all potential input. To use our skills example, to train a machine learning algorithm to sort our mountain of candy, you can't tell it to sort out something green. You have to feed it hundreds of examples of Skittles labeled with the correct colors.

The quality of this training set is key. If you fail to include examples of a slightly different color, such as sour apple, a human being would recognize that as green, but a machine learning algorithm wouldn't.

Machine learning excels at identifying subtle patterns in old data and applying it to new data. It fails when those patterns are not completely relevant to the new situation, and it cannot consider any other context than within which it has been trained.

Despite these weaknesses, I do believe there are still many potential uses of machine learning to keep people safe on-line. In my written testimony, I specifically discuss the example of the terrible mosque shooting in Christchurch as an example where several improvements could have been made in response by tech platforms. As I wrote in detail, the existence of potential solutions are less of a problem in some cases than the existence of social media sites that intentionally aim to host extremist content. This is sometimes less a question of the existence of the tools than the willingness to use them.

There are 7 additional steps I recommended in my written testimony that I believe the social platform should undertake to address critical safety issues. The first is for them to embrace transparent and proportional responses to content violations. The second is to make moderated content available for academic study. The third is to establish better coordinating bodies for multiple different kinds of abuse. The fourth was for the responsible platforms to reduce the movement of users to the sites that are intentionally hosting radicalized content. The fifth is to establish new and separate standards for manipulating in synthetic media. The sixth is to create robust perceptual fingerprinting algorithms that would allow for better and faster sharing between the companies. The seventh is to work on client-side machine learning for a number of different safety purposes.

There are many difficult decisions our country has to make when balancing individual privacy, speech rights, and collective safety. New technologies such as end-to-end encryption are creating a whole new set of balances between legitimate equities. We will actually be convening a workshop on this at Stanford in September in which we will bring together civil society, law enforcement, tech companies, and academics to discuss a new way forward.

Thank you very much for the opportunity to speak today. I am looking forward to your questions.

[The prepared statement of Mr. Stamos follows:]

PREPARED STATEMENT OF ALEXANDER STAMOS

JUNE 25, 2019

I. INTRODUCTION

Chairman Rose, Ranking Member Walker, and committee Members: Thank you for this opportunity to discuss the potential uses and limitations of artificial intelligence in on-line counterterrorism enforcement. My name is Alex Stamos. I am currently the director of the Stanford Internet Observatory, a program of the Stanford University Cyber Policy Center. Our group is performing cross-disciplinary research into misuse of the internet with the goal of providing actionable solutions for tech companies and governments. I am also the William J. Perry Fellow at the Center for International Security and Cooperation, a visiting scholar at the Hoover Institution, a member of the NATO Cybersecurity Center of Excellence advisory council, and a member of the Annan Commission on Elections and Democracy. Before joining Stanford, I was the chief security officer at Facebook from June 2015 until August 2018. During that time, I witnessed the company's battle against on-line terrorism, built and supervised a counter-terrorism investigations unit, and oversaw the company's research into Russian attacks against democratic elections in the United States. Previously, I was the chief information security officer at Yahoo and the co-founder of iSEC Partners, a technical cybersecurity consultancy.

I am honored to be here today and hope that my experience will help clarify some of the misunderstandings, confusion, and hype about the potential of artificial intelligence that is currently circulating media reports and policy discussions, particularly as it relates to the issue of counterterrorism and on-line safety.

As someone who has seen some of the world's best machine learning experts try to apply these techniques to real-world problems, I'm convinced that both the promise and the peril are often exaggerated, making it difficult to have an honest and accurate discussion about the policy implications. The capabilities of these techniques are overstated by tech executives looking for easy answers to difficult problems, start-up founders who need venture capital investments, and media outlets lacking adequate technical expertise to properly kick the tires on wild claims. If we want to accurately assess the impact of artificial intelligence—and the policy regime that should accompany it—we need to be disciplined in defining both its challenges and its capabilities. Today, I will use the more appropriate term "machine learning" as much as possible instead of artificial intelligence because, as we will discuss, there is much more that is artificial than intelligent even with the current state-of-the-art.

II. THE POWER AND LIMITATIONS OF MACHINE LEARNING

The world's best machine learning resembles a crowd of millions of preschoolers. There are certainly problems which a humongous group of children could be taught to solve, such as sorting a mountain of Skittles into 5 smaller mountains based on color. Adding more students to help with tasks like this can improve the speed of their work but won't allow them to perform more complicated individual tasks. No number of small children could work together to build the Taj Mahal or explain the plot of Ulysses. Similarly, modern machine learning can be incredibly powerful for accomplishing routine tasks at amazing scale and speed. However, these technologies are also primitive and often very fragile, in that any deviation from foreseen conditions, including evaluating the impact of individual decisions on the system as a whole, stymie today's best machine learning. Decision making based on societal values and cultural context is completely beyond its capabilities. We still rely on humans for this cognitive ability.

One important thing to understand about modern machine learning is that most of the practical techniques in use today cannot be told what they are supposed to do; they must be shown. Many of the algorithms relevant to our discussion today are known as "classifiers." These are systems that sort digital information into various categories. A classifier is generally trained by feeding it data that has already been labeled by humans, preferably large datasets that represent the diversity of potential inputs. To use our Skittles example, to train a machine learning algorithm to sort our mountain of candy we would start by giving it hundreds of examples of Skittles labeled with the correct colors. The quality of this training set is key; failing to include examples of slightly different sour apple pieces in the set, which humans still perceive as "green", would mean the system would be unprepared for something like a collection of Crazy Sours,[1] not to mention future colors that don't yet exist. Machine learning excels at identifying subtle patterns in old data and applying it to new data. It fails when those patterns are not completely relevant to a new situation and it cannot consider any other context other than in which it has been trained.

In the counterterrorism and more general content moderation context, humans and machines at large tech platforms already work together to understand and make millions of moderation decisions each day. The scale of this work is difficult to fathom. According to Facebook's most recent enforcement report,[2] over 4 billion enforcement actions were taken in the first quarter of this year. This is roughly 500 enforcements per second, 24 hours a day. This only reflects the number of decisions where Facebook decided to act; the overall number of decisions considered, including those where no action was taken, is much higher.

I will point out two interesting conclusions to draw from this data. First, the design of the charts obfuscates the fact that some types of enforcement are around 1,000 times more common than others. For example, Facebook reports taking down approximately 1.76 billion pieces of spam and 4 million pieces of hate speech in 1Q2019. This means that hate speech is 0.2 percent the volume of spam.

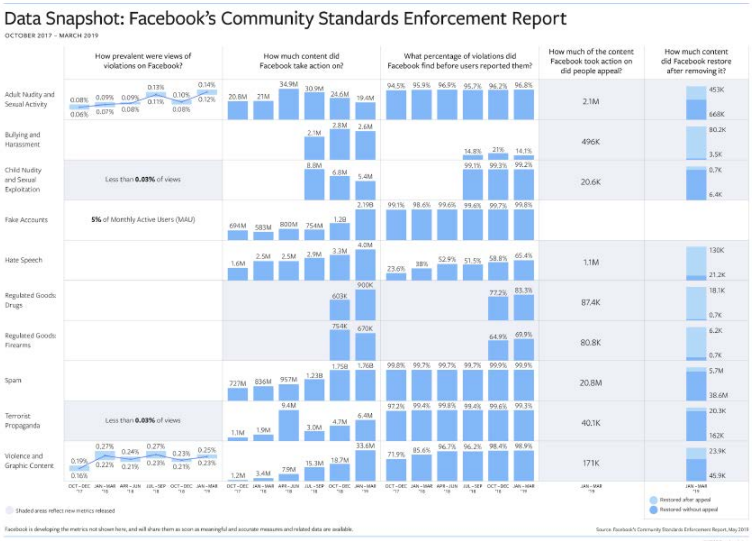
Second, there is a significant difference in the volume of actions taken proactively versus after a user report based on the category of violation. Only 14.1 percent of "Bullying and Harassment" actions were proactive, compared to 99.3 percent for "Terrorist Propaganda."

---

[1] I have perhaps stretched this example too thin, but the unexpected diversity of Skittles colors makes for an interesting example of incomplete or biased training of a machine learning classifier. *https://en.wikipedia.org/wiki/List_of_Skittles_products*.

[2] *https://newsroom.fb.com/news/2019/05/enforcing-our-community-standards-3/*.

FIGURE 1.—SUMMARY DATA FROM FACEBOOK'S COMMUNITY STANDARDS ENFORCEMENT REPORT, PUBLISHED IN MAY 2019 [3]



These disparities reflect the strengths and weaknesses of Facebook's current machine learning systems, but the lessons apply to other uses. Machine learning is much more effective in situations where there are massive sets of both good and bad content available to train classifier models, such as with spam. It is also effective in stopping content for which there are known signatures and general consensus, such as child sexual abuse material [4] (CSAM). It is not good at making decisions when challenging ideas like satire and context come into play.[5] Our political discourse is rife with these modes of speech. These weaknesses have led some groups to caution against too aggressive use of machine learning in content moderation regimes.[6]

III. APPLYING MACHINE LEARNING TO TERRORISM

The March 2019 terrorist attack against the Al Noor Mosque in Christchurch, New Zealand, is a recent example of violence that was undoubtedly influenced, and likely even inspired, by the perpetrator's on-line interactions. The attacker's manifesto and video can only be fully understood in the context of on-line video game, meme, and white supremacist subcultures. Many words have been spent assigning blame for this attack to social media, but the conversation has created more heat than light for platforms and policy makers due to the lack of specificity in how this attacker and others leveraged the internet to fulfill their ultimate goal of spreading hate and terror.

While at Facebook, I worked with Brian Fishman, a Counterterrorism Research Fellow with the International Security Program at New America and a Fellow with the Combating Terrorism Center at West Point. He has spent his career studying terrorists and their on-line activities. He recently published an analysis in the *Texas National Security Review* outlining 7 top-level functions [7] that the internet can serve

[3] Hi-resolution chart available here: *https://fbnewsroomus.files.wordpress.com/2019/05/cser-data-snapshot-052219-final-hires.png*.

[4] This is the preferred term of art for "child pornography" among child safety specialists.

[5] Here is a crude but informative example of a content moderation decision (perhaps automated) that was not aware of sarcasm: *https://twitter.com/thetweetofgod/status/1138461712871436288?s=21*.

[6] *https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf*.

National Security Review outlining 7 top-level functions [7] that the internet can serve for terrorism.[8]

Several of these functions, such as "Financing," are mostly relevant to organized groups such as the Islamic State. However, it is important for today's discussion to understand how social media served a few of these functions for the Christchurch shooter and his allies as well as how machine learning can realistically be applied to each.

Please note that I am extrapolating based on publicly-available information on the Christchurch attacker's on-line activities. The lack of data to inform detailed public discussion of the radicalization path leading to recent terrorist attacks limits the ability for academics, product managers, engineers, and policy makers alike to formulate effective technical responses.

Audience Development

While this term initially seems more relevant to an organized terrorist group with formalized recruitment strategies, the white supremacist terrorism context of the Christchurch attack demonstrates how the internet also provides key capabilities for less structured hate groups to attract new adherents to their cause. The early stages of radicalization do not require exposure to content that calls explicitly for violence. Content that simplifies legitimate grievances and assigns blame to specific categories of people can create the conditions necessary for self-radicalization. As a National Institute of Justice survey of radicalization research put it, " . . . frame crystallization (i.e., identifying and agreeing on who is to blame for a situation and what needs to be done to address it) is a facilitator of terrorism."[9] Content of this type is often found on large social media sites and often does not violate those sites' policies unless explicitly calling for dehumanization or violence.

Based on my experience, the best use of machine learning to interrupt this step is perhaps in blunting the damage caused by other machine learning algorithms, namely recommendation engines. The role of recommendation engines varies widely between social networks, but in many cases, they can be the primary determinant of what content a user consumes. Much has been written on the danger of such systems, although data is scarce and peer-reviewed academic studies remain rare. Nevertheless, it has been shown that recommendation engines can be influenced to push radicalizing content to large audiences. The ML used by recommendation engines can be updated to identify such abuses and limit the audience of non-violating, yet radical content.

Community Maintenance

At this point, there is no evidence that the murderous actions of the Christchurch shooting involved direct participation from anyone but the suspect in custody. However, the propaganda campaign that followed the shooting, conducted while the suspect was already in custody, included thousands of individuals with only the flimsiest of on-line ties to the shooter himself.

This collective action was made possible by the existence of radical, anonymous on-line communities in which racist, anti-Semitic, anti-immigrant, misogynist, and white supremacist thought is not only tolerated but affirmed and normalized. In the case of the Christchurch shooter, the community of choice was 8chan, a message board explicitly created as a response to hate speech restrictions on other sites. It is currently owned and operated by an American living in the Philippines [10] and hosted by two U.S. technology providers headquartered in San Francisco.[11]

The Christchurch shooter posted links to his livestream and multiple copies of his manifesto on 8chan just minutes before beginning his attack. The 8chan thread lasted for hours afterward, filled with supportive comments from other members, including discussion of how to spread the message of the shooter. Once the original thread was taken down, dozens more were created with links to the shooting video and advice on how to defeat the site's content filters. Today, it is still easy to find

---

[7] The functions Fishman names are: Content Hosting, Audience Development, Brand Control, Secure Communication, Community Maintenance, Financing and Information Collection, and Curation.

[8] Fishman, B. (2019, May 24). Crossroads: Counterterrorism and the Internet. Retrieved from *https://tnsr.org/2019/02/crossroads-counter-terrorism-and-the-internet*.

[9] Smith, A. (2018, June). How Radicalization to Terrorism Occurs in the United States: What Research Sponsored by the National Institute of Justice Tells Us. Retrieved from: *https://www.ncjrs.gov/pdffiles1/nij/250171.pdf*.

[10] Jim Watkins, as discussed here: *https://splinternews.com/meet-the-man-keeping-8chan-the-worlds-most-vile-websit-1793856249*.

[11] NT Technology *(https://nttec.com)* and Cloudflare *(https://www.cloudflare.com)* are the major hosting providers for 8chan.

entire discussion threads on 8chan dedicated to celebrating the attacker and discussions of "continuing his work".

There is some potential application of machine learning techniques to address this issue. To the extent that these communities operate in private spaces hosted on large platforms, machine learning can be used to detect and shut down these groups at scale. There are difficult privacy issues to balance here, as any such activity will require humans to enter spaces that might be considered private by the participants. Detailed investigations should be based upon a strong internal predicate, such as a high-confidence classification by machine learning. The technical challenges, however, are minimal.

A much more pressing issue than the existence of machine learning techniques is the willingness of the worst actors to deploy them. I am often asked why the major tech companies were more successful in eliminating Islamic State content from their platforms versus white supremacists. This is a complicated issue, with multiple factors including the tendency of ISIS members to self-identify, quite visibly and with prominent iconography that machine learning can easily detect, as well as the success of Western law enforcement in infiltrating ISIS support channels and arresting adherents before their planned attacks. A major factor, however, was that very few organizations were willing to intentionally host forums that could serve the need for "community maintenance" for international terrorist organizations. This is a significant departure from the multiple options available to white supremacists, as sites like 8chan happily cultivate them as cherished users.

Figure 2: An example of the Christchurch shooter enlisting his online community to help spread his message using 8chan.

Content Hosting

There were two phases to the Christchurch shooter's content strategy. The first phase was to get the content in the hands of supporters. His manifesto was pregenerated, relatively small and easy to host. Getting a video into the hands of supporters was inherently more difficult because it needed to be streamed in real-time, as the shooter could not be confident of having time after the attack to upload. The shooter chose to use Facebook Live to stream his attack, but his supporters on 8chan recognized that this would not be a sustainable hosting location for the content and made their own copies before Facebook removed it.

The second phase was a coordinated campaign to defeat human and machine learning moderation, executed by the external supporters who modified and reuploaded the video and manifesto millions of times.[12]

[12] Facebook estimated 1.5M re-upload attempts in the first 24 hours. No data is available for YouTube or other large platforms. *https://www.washingtonpost.com/technology/2019/03/21/*

There is a vast difference in the capability of current machine learning techniques to address these two problems. With current techniques, training a system to detect content like the shooter's manifesto or video with no human intervention is extremely difficult. I do not have direct knowledge of why Facebook did not catch the video in real time, but it is worth noting that machine learning systems need to be trained on examples of good and bad content in order to work. The Christchurch video was unlike malicious videos Facebook's systems had seen before. It is much less bloody than ISIS beheading videos, and it is very different than the kind of suicide videos Facebook has dealt with in the past. The first-person perspective is reminiscent of a video game, and the soundtrack is punctured by gunshots, but not in a way easily distinguishable from movies or game footage. This was the first example of this kind of violent video on Facebook, and while an adult human could very quickly recognize the grave context, without a large corpus of similar training videos, it is unlikely that machine learning would have done so alone.

Addressing the second problem—the mass uploads of slightly varied videos—is more tractable, because once a human determines that a piece of content should be banned machine learning can help do so at incredible speed and scale. To this end, the platforms were partially successful. Facebook estimated that around 80 percent of re-upload attempts were caught via automated means. The other 20 percent were likely videos that had been modified in ways specifically meant to trick machine learning. There are several simple techniques to do so and such techniques are well understood as they have been honed for years by content pirates looking to defeat the copyright scanning mechanisms of major video sites. Still, there is serious room for improvement in this area, which I will outline.

IV. WHAT CAN THE TECH COMPANIES DO?

There are several steps I recommend social platforms undertake to address the critical safety issues we are discussing today.

(1) Embrace transparent and proportional responses to content violations.—The major tech platforms face an impossible problem when various segments of societies around the world demand incompatible solutions to complex international issues. The news media regularly bemoans the power of these companies while calling for them to regulate the political speech of millions. Policy makers demand that the companies collect as little identifiable data as possible on users while also expecting them to be able to discern professional spies from billions of legitimate users. Political parties around the world ask for the platforms to censor their opponents and appeal any equivalent moderation of their own content.

The platforms have partially created this problem for themselves by not being transparent about the tradeoffs that must be considered when solving such problems. This includes many issues around content moderation, where the companies have been unwillingly forced into the position of referee over legally protected political speech in developed democracies like the United States. While there is no single answer that will keep all parties happy, the platforms must do a much better job of elucidating their thinking processes and developing public criteria that bind them to the precedents they create with every decision.

The other focus should be on expanding the public discussion on content moderation beyond just deleting posts and banning users—the standard response for extreme content. There remain many kinds of speech that are objectionable to some in society, but not to the point where huge, democratically unaccountable corporations should completely prohibit such speech. The decisions made in these gray areas create precedents that aim to serve public safety and democratic freedoms but can also imperil both.

In my view, each of these major platforms can be divided into several different components, each with unique capabilities to amplify speech and, by extension, the potential benefits and harm from that speech. I have included a basic decomposition of a social media platform, with the greatest reach (and least private) components on top and the components with the highest expectation of privacy and the least reach on the bottom.

The two most important components are the advertising and recommendation engines, partially due to the huge amplification either can provide, but also because these are the two components that put content in front of people who did not ask to see it. A member of a private anti-vaccination group has affirmatively chosen to expose themselves to those views, and it would be reasonable for large platforms to let private groups like that exist. It is also reasonable for them not to allow anti-

---

facebook-reexamine-how-recently-live-videos-are-flagged-after-christchurch-shooting/?utm__term=.c745ad62d020.

vaccination campaigns to trade money for amplification via advertising, thereby pushing their content onto millions who had demonstrated no desire to see it.

A public embrace of transparent mechanisms for content moderation by the companies, combined with more nuanced discussion by policy makers and the media, would go a long way toward creating an environment where these issues can be productively debated and better understood.

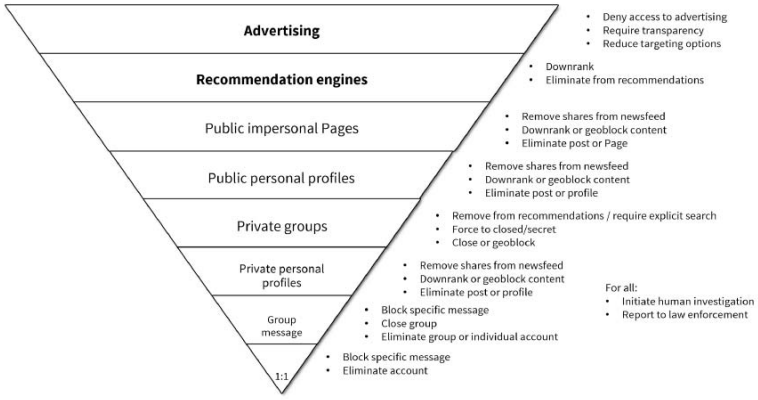


Figure 3: Levels of amplification effect and potential moderation tools.

(2) Make moderated content available for academic study.—Part of bringing new transparency to tech platforms' decision-making process should be the creation of archives of moderated content that could be provided to academic researchers under privacy-preserving terms. While the companies have been aggressively pushing back against claims of political bias it is difficult for outside observers to verify their claims without access to data. The deletion of moderated content also has negative impacts on groups studying war crimes [13] and academics who would like to better understand foreign influence campaigns. In both cases, a time-limited archive of moderated content could enable useful research while also protecting user privacy. Something like this already exists for copyright-related takedowns.[14]

This is an area for more study by Congress, as I have heard multiple attorneys from the large companies remark that such as archive would likely not be compatible with current U.S. or E.U. privacy laws. The creation of a safe harbor for academic study should be part of the consideration of any U.S. privacy legislation.

(3) Establish better coordinating bodies for multiple abuse types.—During the height of the struggle against the Islamic State's on-line propaganda efforts, the major tech companies created a new coordinating body: The Global Internet Forum to Counter Terrorism.[15] This group has been somewhat successful in building capabilities in smaller members while creating a forum for collaboration among the larger members. It is time to follow this initial foray with a much more ambitious coordinating body between tech companies focused on adversarial use of their technologies.

Several weeks ago, my colleagues at Stanford and I released a report [16] with 45 recommendations on securing the U.S. election system from attack. One of our recommendations was the creation of a coordinating body in the model of the Financial Services ISAC [17] and other successful examples. Such a body would need its own budget, staff with security clearances to receive threat briefings, technical tools, and the power to facilitate sharing and consensus building among the membership. Counterterrorism is one of several missions along with protecting against advanced cyber attack, election security, and combating fraud that could be handled by working groups inside such an organization without the need for separate, specialized organizations.

---

[13] *https://phys.org/news/2018-09-crucial-video-evidence-war-crimes.html*.
[14] *https://www.lumendatabase.org/*.
[15] *https://www.gifct.org/*.
[16] *http://electionreport.stanford.edu*.
[17] *https://www.fsisac.com/*.

Congress can assist with this effort by creating privacy and antitrust safe harbors around the sharing of information and banning of proscribed content.

(4) Reduce the movement of users to intentionally radicalizing sites.—One of the issues a new coordinating body could tackle is how to handle the message boards and social media sites that intentionally host radical groups that support violent acts. Such websites seem to be legal under U.S. law, although legal action in other countries could still bring pressure on their operators. The large tech firms do not (and should not) have the ability to ban such sites from existing. What they can do, however, is reduce the chance that content on their platforms can be used as a jumping-off point for these communities.

A coordinating body could decide to maintain a list of sites that could then be voluntarily banned from the major social media platforms. As of today, Facebook, Google, and Twitter are deciding on a per-page basis of whether to allow links to these sites. A global ban on these domains would be consistent with steps they have taken against malware-spreading, phishing, or spam domains and would allow those sites to exist while denying their supporters the capability to recruit new followers on the large platforms.

(5) Establish separate standards for manipulated media.—While not directly related to today's focus on counterterrorism, the rise of synthetic or manipulated media (such as Deep Fakes) is another serious challenge for the major social media platforms. Several recent hearings [18] have focused on recent video of Speaker Pelosi that was edited to slur her words. While distasteful, this video falls within the traditional bounds of allowed political criticism and was demonstrated to have been uploaded by an individual U.S. citizen and not as part of an organized disinformation campaign.[19] Personally, I believe this kind of distasteful political speech should not be centrally censored, either by Government action (which would almost certainly be Constitutionally precluded) or by the platforms.

This is a great example of an issue that deserves a more nuanced approach. In this case, I believe the tech platforms need a new set of policies defining manipulated and synthetic media that is not tied to any fact-checking processes. While the companies do not want to set themselves up as the Ministry of Truth, they should be able to label misleading videos based solely upon technical evidence and remove them from recommendation systems. Such labels should be applied much more aggressively than they are now, including to comedy clips [20] and other uses that are not intended to mislead.

(6) Create robust perceptual fingerprinting algorithms.—The most common standard for creating digital fingerprints of images is PhotoDNA. This technology, invented by Microsoft over a decade ago, has had a huge impact on the ability of technology providers to work with law enforcement and the National Center on Missing and Exploited Children (NCMEC) to fight the spread of child sexual abuse materials. While incredibly successful, PhotoDNA is showing its age and is not up to the current needs of our industry.

The first issue is the lack of robustness against intentional attempts to distort images to defeat the algorithm. Microsoft understands the potential weakness of PhotoDNA, which is why it carefully guards the secret of its operation using intellectual property laws and restrictive contracts with their partners. While Microsoft has allowed several other large companies to use the algorithm in their own data centers, it has never been embedded in client-side software and is no longer available in the source code form to smaller companies. PhotoDNA was also built specifically for still images and attempts to apply it to video have been computationally inefficient.

There are video hashing algorithms available inside of the big platforms, and these have been shared with other members of Global Internet Forum to Counter Terrorism (GIFCT), but this is a toolset that can still be expanded publicly.

This is also an area where academic computer science can directly contribute. There has been a great deal of academic work on machine vision over the last decade, and there is no reason why there cannot be a new revolution in perceptual algorithms that are robust enough against attack to be publicly published and deployed in many more circumstances.

My recommendation to industry is to encourage the creation of replacements for PhotoDNA via a large public competition, similar to those run by NIST to choose

---

[18] *https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=657*.

[19] *https://www.thedailybeast.com/we-found-shawn-brooks-the-guy-behind-the-viral-drunk-pelosi-video*.

[20] An example of a comedy clip that should be allowed to exist but labeled as edited: *https://www.facebook.com/JimmyKimmelLive/videos/drunk-donald-trump-i-dont-know-what-the-hell-hes-talking-about-edition/686503181736071/*.

encryption algorithms but backed with cash prizes. For a reasonable investment, a consortium of large companies could fund multiple rounds of research, development, testing and qualification of robust fingerprinting algorithms for various uses. The winning algorithms could then be licensed freely and deployed much more widely than PhotoDNA is currently.

(7) Develop client-side machine learning for safety purposes.—Another area of potential technical advancement is in the use of machine learning on our ever-more-powerful handheld devices. The deployment of end-to-end encryption technologies in billion-user platforms has led to huge improvements to the privacy of law-abiding individuals but has also posed serious challenges for law enforcement. At Stanford, we are looking into ways to solve this issue without reducing privacy and security.

One possible model is to deploy some of the machine learning techniques that have been used to look for malicious content into the end devices. Such an architectural shift would allow the platforms to provide mathematically proven privacy while also looking for potentially harmful content and prompting the user to decrypt the connection and ask for assistance. This would not be a valid approach to conspiratorial use of communication platforms among willing participants, but it could provide other mitigations as more platforms move to encrypting more data.

There are many difficult decisions our country has made when balancing individual privacy with collective safety. End-to-end encryption has created a whole new set of balances between legitimate equities, and we will be convening a workshop at Stanford in September to bring together civil society, law enforcement, tech companies, and academics to discuss ways forward.

Thank you again for the opportunity to speak with you today. I look forward to your questions.

Mr. ROSE. Thank you for your testimony.

I now recognize Mr. Buchanan for his statement for 5 minutes.

STATEMENT OF BEN BUCHANAN, ASSISTANT TEACHING PROFESSOR, GEORGETOWN UNIVERSITY, SENIOR FACULTY FELLOW, CENTER FOR SECURITY AND EMERGING TECHNOLOGY, MORTARA CENTER

Mr. BUCHANAN. Thank you, Chairman Rose and Ranking Member Walker. My name is Ben Buchanan. I am an assistant teaching professor at the School of Foreign Service and a senior faculty fellow at the Center for Security and Emerging Technology, both at Georgetown University. I am also a global fellow at the Woodrow Wilson Center.

My research specialty is examining how cybersecurity and machine learning shape international security. To help structure our discussion, I would like to offer a few thoughts.

First, AI offers some promise as a tool for moderation on-line. Social media platforms operate at a gigantic scale, sometimes including several billion users. It is deeply unrealistic to think that any team of humans will be able to monitor communications at that scale without automated tools. Social media moderation remains a thankless and grueling job, but machine learning is already useful in lessening the burden at least somewhat.

Optimists, as you said, Mr. Chairman, envision a future in which AI quickly and effectively takes on the majority of this moderation task. I am deeply skeptical that this is possible. While machine learning is powerful, the moderation problem is extremely difficult, and I would like to tell you why that is the case.

First, context is vitally important, and context can often be hard for algorithms to grasp. The same video of a terrorist attack might be propaganda in one setting but legitimate news reporting in another. The same video of soldiers on patrol may in one format be meant to instill patriotic pride but in other context serve as a threat to those soldiers.

My understanding of the technology is that it is a long way from being able to identify this context and respond appropriately.

Second, machine learning systems, by definition, rely on distilling patterns. Some objectionable content differs from what came before or as in a language that machine learning systems cannot parse well.

As you said, Mr. Chairman, Facebook says that the systems are 99 percent effective against propaganda from the Islamic State and al-Qaeda. This stat seems overly optimistic to me given the pattern limitations of machine learning systems. Both terrorist groups exhibit consistent patterns in their messages, making them easier to identify.

More generally, looking beyond just those two terrorist groups, the AP found that much objectionable content slipped through automated filtering, including an execution video, images of severed heads, and propaganda honoring martyred militants.

The third reason I am skeptical that machine learning can solve the moderation problem is that adversaries will adapt and systems will not have time to respond. Several of you have mentioned the terrible massacre in Christchurch, New Zealand, which was streamed live all over the world. Such a thing had never been done before. The video went viral, in part, due to how users slightly edited the footage to evade detection of automated systems.

Unfortunately, we must assume that our adversaries will continue to innovate and improve.

Fourth, and related, is that partial success with content moderation is often not sufficient. Though Facebook and YouTube were able to take down some copies of the Christchurch video, many other copies evaded their detection. The copies that did escape the filters were more than sufficient to ensure that the video attracted wide-spread attention.

When my students in class get 90 percent of the questions right, they get an A or an A-minus. Unfortunately, even a very good percentage in the moderation problem is not enough to ensure success.

In sum, to solve the moderation problem, an AI system would have to not just—not just identify content that might be objectionable, but also grasp context, discover new patterns, respond quickly to an adversary's changing tactics, and work correctly a very large percentage of the time without a large number of false positives. I am skeptical whether such a system will exist in the near future.

I would encourage you to ask social media companies whether they believe such a system is possible and why, or if they do not think it is attainable, then what their plan is to scale content moderation to billions of users.

There is one other point to this discussion that I believe deserves attention: Recommendation systems. Such systems will suggest content to users on a social media platform based on what they and others have already viewed. This creates a loop designed to keep users on the platform.

Some research suggests that these recommendation systems, such as the recommended videos feature on YouTube, are built to push users toward ever more extreme content. The research raises a very alarming possibility, that not only are automated moderation systems insufficient for removing objectionable content, but

that other automated systems, recommendation algorithms, in fact, are driving users to objectionable content that they otherwise would not find, making them a tool for radicalization.

The public data on this subject is limited and not yet conclusive, but should cause concern. I encourage you to ask technology companies about them and to make more data available.

In conclusion, we ought not to lose sight of why we are here today. Technology is not the reason for this hearing. We are here because humans abuse a system that other humans have created. Human moderation as well as current and near-future technology seem insufficient to stop that abuse. Social media platforms could likely do more, such as reducing how quickly messages go viral, expanding their teams focused on this issue, making more data available, and changing recommendation systems. That said, my best guess is these steps might mitigate the problem but are deeply unlikely to solve it.

I thank you again for holding this hearing and look forward to your questions.

[The prepared statement of Mr. Buchanan follows:]

PREPARED STATEMENT OF BEN BUCHANAN

Thank you, Chairman Rose and Ranking Member Walker, for holding this important hearing and for inviting me to testify.

My name is Ben Buchanan. I am an assistant teaching professor at the School of Foreign Service and a senior faculty fellow at the Center for Security and Emerging Technology, both at Georgetown University. I am also a global fellow at the Woodrow Wilson International Center for Scholars, where I teach introductory classes on AI and cybersecurity for Congressional staff. My research specialty is examining how cybersecurity and AI shape international security. I co-authored a paper entitled "Machine Learning for Policymakers."[1]

The title of todays' hearing rightly alludes to the possibilities and limitations of AI as it applies to counterterrorism. To help structure our examination of both, I'd like to offer some thoughts to conceptualize the potential areas of contribution and concern.

AI AS A TOOL OF MODERATION

AI offers some promise as a tool of moderation. Social media platforms operate at a gigantic scale, sometimes including several billion users. It is deeply unrealistic to think that any team of humans will be able to monitor communications at that scale without automated tools. Social media moderation remains a thankless and grueling job for those individuals who do it, but AI is already useful in lessening the burden at least somewhat. Perhaps most importantly, AI sometimes helps platforms respond more quickly to objectionable content, swiftly preventing it from spreading. Optimists envision a platform in which AI quickly and effectively takes on the vast, or the entire, share of the difficult job of moderation, leaving users to enjoy an on-line experience that meets their expectations.

I am deeply skeptical that this is possible. In general, policy makers underestimate the power of machine learning systems and the rapid rate of change, but I think the moderation problem is one of the most fiendishly difficult ones—so difficult, in fact, that technology companies struggle to come up with enforceable and clear standards that their human moderators can consistently enforce, much less standards that machines can apply.

There are at least four reasons why this problem is hard. First is that context is vitally important, and context can often be hard for algorithms to grasp. The same video of a terrorist attack might be propaganda in one setting but legitimate news reporting in another. The same video of soldiers on patrol may in one format be meant to instill patriotic pride but in another context serve as a threat to those

[1] Buchanan, Ben and Taylor Miller. "Machine Learning for Policymakers." Belfer Center for Science and International Affairs (2017), *https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.*

soldiers. My understanding of the technology is that it is a long way from being able to identify this context and respond appropriately.

Second is that machine learning-based systems by definition rely on distilling patterns, and objectionable content does not always fit into neatly observable patterns. For example, content moderation systems are vastly less effective in unfamiliar languages. In addition, they are less effective at catching objectionable content that takes on unfamiliar forms. For example, one 5-month study and whistleblower complaint obtained by the AP contends that Facebook, using both its automated and human moderation capabilities, removed only 38 percent of content posted by terrorist organizations. Facebook claims that its systems are much more effective, citing a 99 percent success rate, but it seems that the firm's denominator in calculating that percentage is only a subset of the content that is prohibited. The AP found that much objectionable content slipped through algorithmic filtering, including "an execution video, images of severed heads, propaganda honoring martyred militants."[2]

The third reason I am skeptical that AI can solve the moderation problem is that, sometimes, there is not sufficient time to train machine learning systems with new data. Consider the gun massacre in Christchurch, New Zealand, in which the objectionable content was streamed live all over the world. Such a thing had never been done before, and social media companies' automated systems were not nearly sufficient to keep the video from going viral, in part due to how users slightly edited the video to evade the detection of those systems. Unfortunately, we must assume that our adversaries will innovate and improve, finding weaknesses and exploiting them before companies have a chance to respond.

Fourth, and related, is that partial success with content moderation is often not sufficient. Consider the video of the terrible Christchurch shooting once more. As I said, though Facebook and YouTube were able to take down some copies, many other copies evaded their detection. The copies that did escape the filter were more than sufficient to ensure that the video still was able to go viral and attract widespread attention.[3]

In sum, to solve the moderation problem an AI system would have to not just identify content that might be objectionable, but also grasp context, be able to identify objectionable content that is distinct from what came before and in unfamiliar languages, respond quickly to an adversary's changing tactics, and work correctly a very large percentage of the time without a large number of false positives. I am skeptical whether such a system exists or will exist in the very near future. In short, from my vantage point, I see more limitations here than possibilities. I would encourage you to ask representatives of social media companies whether they think such a system is achievable or, if they do not think such a system is achievable, then what their plan is to scale content moderation to billions of users.

AI AS A TOOL OF RADICALIZATION

There is one other point to this discussion that I believe deserves significant attention: Research has recently come out suggesting that automated recommendation systems can contribute to the radicalization of individuals. Such systems will recommend videos, articles, or other content to users on a social media platform based on what they have consumed on it already and what others users have viewed and liked, creating a loop of content designed to keep users on the platform.

Some academics argue that it is an effect by design for these recommendation systems, such as the Recommended Videos feature on YouTube, to push users toward even more extreme videos.[4] In this sense, then, AI is not a force for moderation online but in fact a force for radicalization.

My assessment of this research is that raises significant concerns, though I do not think the data is yet definitive. The research does, however, raise a very alarming possibility: That not only are automated moderation systems insufficient for remov-

---

[2] Butler, Desmond, and Barbara Ortutay, "Facebook Auto-Generates Videos Celebrating Extremist Images", *Wall Street Journal,* 9 May 2019, *https://www.apnews.com/f97c24dab4f34bd0b48b36f2988952a4*.

[3] Timberg, Craig, Drew Harwell, Hamza Shaban, and Andrew Ba Tran, "The New Zealand Shooting Shows How YouTube and Facebook Spread Hate and Violent Images—Yet Again", *Washington Post,* 15 March 2019, *https://www.washingtonpost.com/technology/2019/03/15/facebook-youtube-twitter-amplified-video-christchurch-mosque-shooting/?utm_term=.b37e96-04a2da*.

[4] Nicas, Jack, "How YouTube Drives People to the Internet's Darkest Corners", *Wall Street Journal,* 7 February 2018, *https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478* Tufekci, Zeynep, "YouTube, the Great Radicalizer", *New York Times,* 10 March 2018, *https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html*.

ing objectionable content but that other automated systems—technology companies' recommendation algorithms—in fact are driving users to objectionable content that they otherwise would not find. If this is the case, then the technical limitations of AI work against the interests of National security, while its vast possibilities work for those who benefit from radicalization on-line. Again, the public data is limited and not yet conclusive, but it seems to me that the net effects of recommendation systems that steer users to content generated by others users need substantial additional study. I encourage you to ask technology companies about them.

Worse still is that automated algorithms on social media platforms cannot just drive users to objectionable content but help make that content more appealing and visible. The AP and academic researchers found that Facebook's algorithms automatically generate slick videos of some of the extremist content that has evaded its filtering. These algorithmically generated videos take images and videos that extremists have uploaded and package it to make it more neatly edited and synthesized—in essence, unintentionally doing the work of propaganda.[5]

CONCLUSION

In conclusion, we ought not to lose sight of a vital broader point: Technology is not the reason we are here today. We are here because humans abuse a system that other humans have created. Human moderation as well as current and near-future technology are—in my view—insufficient to stop that abuse. My sense is that there is likely more that social media platforms could do to better manage the problem, such as reducing how quickly messages go viral, expanding their AI research teams focused on this issue, and adjusting their recommendation and generation algorithms, even if it comes at the expense of their business. That said, my best guess is that these steps might mitigate the problem but are unlikely to solve it.

I thank you again for holding this hearing and I look forward to your questions.

Mr. ROSE. Thank you for your testimony.

I now recognize Mr. Sanchez to summarize his statement for 5 minutes.

STATEMENT OF JULIAN SANCHEZ, SENIOR FELLOW, CATO INSTITUTE

Mr. SANCHEZ. Thank you, Chairman Rose, Ranking Member Walker, and the committee in general, for the opportunity to address you today. My name is Julian Sanchez. I am a senior fellow at the Cato Institute. As a firm believer in comparative advantage, given the technical expertise represented on the panel, I think it makes more sense for me to focus on some of the broader policy considerations implicated by automated content filtering.

I think the Christchurch massacre and the attempts to halt spread of the video of that brutal attack are perhaps a good place to start because they illustrate some of the policy and value-based tradeoffs involved.

We have two pretty good public accounts by legal scholar Kate Klonick in *The New Yorker* and reporters Craig Timberg and Elizabeth Dwoskin in *The Washington Post* of the efforts by Facebook and YouTube, respectively, to limit spread of that video. If those platforms attracted a fair amount of criticism for their failure to halt its spread sufficiently, rapidly, or effectively, it was certainly not for lack of trying, as is clear from their portraits. Both companies had large moderation teams that worked around the clock in an effort to halt the spread of the video.

One issue they faced was, as mentioned previously, modifications by uploaders. Small changes to the video, reversing it left-to-right,

[5] Butler, Desmond, and Barbara Ortutay, "Facebook Auto-Generates Videos Celebrating Extremist Images", *Wall Street Journal,* 9 May 2019, *https://www.apnews.com/f97c24dab4f34bd0b48b36f2988952a4.*

adding a filter, adding animations, like Instagram filters do, atop it would make it difficult for an automated system to recognize that video as the same content. So there was constant updating, but it couldn't match the speed which the new changes were made.

There is the additional problem of judging context and purpose. Many and perhaps even most uploaders of that video were not uploading it with the purpose of glorifying terrorism. Many were doing so because they believed it newsworthy, many were doing so in order to condemn terrorism and illustrate the dangers that hateful ideologies can lead to. Some were perhaps misguidedly doing so to invoke sympathy for the victims.

It is a complicated question whether those purposes justify it in the eyes of the platforms, but in that case, the platforms made a decision to balance the social equities weighed in favor of broad prohibition, irrespective of the purpose of the upload, though, that is a difficult content sensitive decision to make. Indeed, in the interest of halting the spread of the video, many platforms were forced to implement broad restrictions on searching and sharing of new content in order to finally get it, to some extent, under control.

So I think, you know, this illustrates a few things. One is the effectiveness of AI has been very much exaggerated. We see in the popular press reports to the effect that, for example, Facebook's AI systems pull down 99 percent of the content—terrorist content that is uploaded, which is a slight misrepresentation of the actual report which found that, of the content they ultimately removed, 99 percent was identified by automated filtering systems, which is unsurprising.

Software works a lot faster than people do as a rule but doesn't tell you what percentage of content on the site that they didn't identify. All of it they didn't identify remained up. Doesn't tell you how much of that there was, and it doesn't tell you how much of the content it did flag and pull down was a false positive, either because it misidentified benign content as extremist or because it misidentified the context of the upload.

This is a common problem, and it is one reason that a series of U.N. special rapporteurs on human rights, as well as a broad array of civil society groups and human rights groups, have opposed a proposal in the European Union to mandate automated filtering by platforms to remove terrorist content.

Journalists and activists reported that there is a real social cost to false positives. For example, videos attempting to document human rights abuses and atrocities have been mistakenly pulled down, effectively deleting evidence of war crimes. More recently, a crackdown on white supremacists content by YouTube resulted in the removal of content hosted by a number of educational associations. Again, it is the same content; the purpose is different.

Again, you know, AI is not a panacea, and mandates of the kind the European Union proposed, as the U.N. special rapporteurs argued, would have incentivized a sort of err on the side of take-down mentality rather than an attempt to weigh values in a context-specific way.

Finally, I would just suggest that, given that all social media platforms engage in monitoring as private entities and filtering of content that goes far beyond what the Government would permit

it to engage in under the First Amendment, it would be unwise for Congress to intervene legislatively in a way that would call into legal question whether those private actions are truly private and open to court challenge, the policies that filter so much of the objectionable content that creates the public fora we now enjoy.

I thank you for this invitation and look forward to your questions.

[The prepared statement of Mr. Sanchez follows:]

PREPARED STATEMENT OF JULIAN SANCHEZ

JUNE 25, 2019

My thanks to the Chair, Ranking Member, and all Members of this subcommittee for the opportunity to speak to you today.

As a firm believer in the principle of comparative advantage, I don't intend to delve too deeply into the technical details of automated content filtering, which my co-panelists are far better suited than I to address. Instead I want to focus on legal and policy considerations, and above all to urge Congress to resist the temptation to intervene in the highly complex—and admittedly highly imperfect—processes by which private on-line platforms seek to moderate both content related to terrorism and "hateful" or otherwise objectionable speech more broadly. (My colleague at the Cato Institute, John Samples, recently published a policy paper dealing still more broadly with issues surrounding regulation of content moderation policies, which I can enthusiastically recommend to the committee's attention.)[1]

The major social media platforms all engage, to varying degrees, in extensive monitoring of user-posted content via, a combination of human and automated review, with the aim of restricting a wide array of speech those platforms deem objectionable, typically including nudity, individual harassment, and—more germane to our subject today—the promotion of extremist violence and, more broadly, hateful speech directed at specific groups on the basis of race, gender, religion, or sexuality. In response to public criticism, these platforms have in recent years taken steps to crack down more aggressively on hateful and extremist speech, investing in larger teams of human moderators and more sophisticated algorithmic tools designed to automatically flag such content.[2]

Elected officials and users of these platforms are often dissatisfied with these efforts—both with the speed and efficacy of content removal and the scope of individual platforms' policies. Yet it is clear that all the major platforms' policies go far further in restricting speech than would be permissible under our Constitution via state action.

The First Amendment protects hate speech. The Supreme Court has ruled in favor of the Constitutional right of American neo-Nazis to march in public brandishing swastikas,[3] and of a hate group to picket outside the funerals of veterans displaying incredibly vile homophobic and anti-military slogans.[4]

While direct threats and speech that is both intended and likely to incite "imminent" violence fall outside the ambit of the First Amendment, Supreme Court precedent distinguishes such speech from "the mere abstract teaching . . . of the moral propriety or even moral necessity for a resort to force and violence,"[5] which remains protected. Unsurprisingly, in light of this case law, a recent Congressional Research Service report found that "laws that criminalize the dissemination of the pure advocacy of terrorism, without more, would likely be deemed unconstitutional."[6]

---

[1] John Samples, "Why the Government Should Not Regulate Content Moderation of Social Media" (Cato Institute) *https://www.cato.org/publications/policy-analysis/why-government-should-not-regulate-content-moderation-social-media#full*.

[2] See, e.g., Kent Walker "Four steps we're taking today to fight terrorism online" Google (June 18, 2017) *https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/*; Monika Bickert and Brian Fishman "Hard Questions: What Are We Doing to Stay Ahead of Terrorists?" Facebook (November 8, 2018) *https://newsroom.fb.com/news/2018/11/staying-ahead-of-terrorists/*; "Terrorism and violent extremism policy" Twitter (March 2019) *https://help.twitter.com/en/rules-and-policies/violent-groups*.

[3] *National Socialist Party of America* v. *Village of Skokie,* 432 U.S. 43 (1977).

[4] *Snyder* v. *Phelps,* 562 U.S. 443 (2011).

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

[6] Kathleen Anne Ruane, "The Advocacy of Terrorism on the Internet: Freedom of Speech Issues and the Material Support Statutes" Congressional Research Service Report T44646 (September 8, 2016) *https://fas.org/sgp/crs/terror/R44626.pdf*.

Happily—at least, as far as most users of social media are concerned—the First Amendment does not bind private firms like YouTube, Twitter, or Facebook, leaving them with a much freer hand to restrict offensive content that our Constitution forbids the law from reaching. The Supreme Court reaffirmed that principle just this month, in a case involving a public access cable channel in New York. Yet as the Court noted in that decision, this applies only when private determinations to restrict content are truly private. They may be subject to First Amendment challenge if the private entity in question is functioning as a "state actor"—which can occur "when the Government compels the private entity to take a particular action" or "when the Government acts jointly with the private entity."[7]

Perversely, then, legislative efforts to compel more aggressive removal of hateful or extremist content risk producing the opposite of the intended result. Content moderation decisions that are clearly lawful as an exercise of purely private discretion could be recast as government censorship, opening the door to legal challenge. Should the courts determine that legislative mandates had rendered First Amendment standards applicable to on-line platforms, the ultimate result would almost certainly be more hateful and extremist speech on those platforms.

Bracketing legal considerations for the moment, it is also important to recognize that the ability of algorithmic tools to accurately identify hateful or extremist content is not as great as is commonly supposed. Last year, Facebook boasted that its automated filter detected 99.5 percent of the terrorist-related content the company removed before it was posted, with the remainder flagged by users.[8] Many press reports subtly misconstrued this claim. The *New York Times,* for example, wrote that Facebook's "A.I. found 99.5 percent of terrorist content on the site."[9] That, of course, is a very different proposition: Facebook's claim concerned the ratio of content removed after being flagged as terror-related by automated tools versus human reporting, which should be unsurprising given that software can process vast amounts of content far more quickly than human brains. It is not the claim that software filters successfully detected 99.5 percent of all terror-related content uploaded to the site—which would be impossible since, by definition, content not detected by either mechanism is omitted from the calculus. Nor does it tell us much about the false-positive ratio: How much content was misidentified as terror-related, or how often such content appeared in the context of posts either reporting on or condemning terrorist activities.

There is ample reason to believe that such false positives impose genuine social cost. Algorithms may be able to determine that a post contains images of extremist content, but they are far less adept at reading contextual cues to determine whether the purpose of the post is to glorify violence, condemn it, or merely document it—something that may in certain cases even be ambiguous to a human observer. Journalists and human rights activists, for example, have complained that tech company crackdowns on violent extremist videos have inadvertently frustrated efforts to document human rights violations,[10] and erased evidence of war crimes in Syria.[11] Just this month, a YouTube crackdown on white supremacist content resulted in the removal of a large number of historical videos posted by educational institutions, and by anti-racist activist groups dedicated to documenting and condemning hate speech.[12]

Of course, such errors are often reversed by human reviewers—at least when the groups affected have enough know-how and public prestige to compel a reconsideration. Government mandates, however, alter the calculus. As three United Nations special rapporteurs wrote, objecting to a proposal in the European Union to require automated filtering, the threat of legal penalties were "likely to incentivize platforms to err on the side of caution and remove content that is legitimate or law-

---

[7] *Manhattan Community Access Corp.* v. *Halleck,* 17–1702 (2019).

[8] Alex Schultz and Guy Rosen "Understanding the Facebook Community Standards Enforcement Report" *https://fbnewsroomus.files.wordpress.com/2018/05/understanding__the__community__standards__enforcement__report.pdf.*

[9] Sheera Frenkel, "Facebook Says It Deleted 865 Million Posts, Mostly Spam" *New York Times* (May 15, 2018). Facebook Says It Deleted 865 Million Posts, Mostly Spam *https://www.nytimes.com/2018/05/15/technology/facebook-removal-posts-fake-accounts.html.*

[10] Dia Kayyali and Raja Althaibani, "Vital Human Rights Evidence in Syria is Disappearing from YouTube" *https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/.*

[11] Bernhard Warner, "Tech Companies Are Deleting Evidence of War Crimes" *The Atlantic* (May 8, 2019). *https://www.theatlantic.com/ideas/archive/2019/05/facebook-algorithms-are-making-it-harder/588931/.*

[12] Elizabeth Dwoskin, "How YouTube erased history in its battle against white supremacy" *Washington Post* (June 13, 2019). *https://www.washingtonpost.com/technology/2019/06/13/how-youtube-erased-history-its-battle-against-white-supremacy/?utm__term=.e5391be45aa2.*

ful."[13] If the failure to filter to the Government's satisfaction risks stiff fines, any cost-benefit analysis for platforms will favor significant overfiltering: Better to pull down ten benign posts than risk leaving up one that might expose them to penalties. For precisely this reason, the E.U. proposal has been roundly condemned by human rights activists[14] and fiercely opposed by a wide array of civil society groups.[15]

A recent high-profile case illustrates the challenges platforms face: The efforts by platforms to restrict circulation of video depicting the brutal mass shooting of worshippers at a mosque in Christchurch, New Zealand. Legal scholar Kate Klonick documented the efforts of Facebook's content moderation team for *The New Yorker,*[16] while reporters Elizabeth Dwoskin and Craig Timberg wrote about the parallel struggles of YouTube's team for *The Washington Post*[17]—both accounts are illuminating and well worth reading.

Though both companies were subject to vigorous condemnation by elected officials for failing to limit the video quickly or comprehensively enough, the published accounts make clear this was not for want of trying. Teams of engineers and moderators at both platforms worked around the clock to stop the spread of the video, by increasingly aggressive means. Automated detection tools, however, were often frustrated by countermeasures employed by uploaders, who continuously modified the video until it could pass through the filters. This serves as a reminder that even if automated detection proves relatively effective at any given time, they are in a perennial arms race with determined humans probing for algorithmic blind spots.[18] There was also the problem of users who had—perhaps misguidedly—uploaded parts of the video in order to condemn the savagery of the attack and evoke sympathy for the victims. Here, the platforms made a difficult real-time value judgment that, in this case, the balance of equities favored an aggressive posture: Categorical prohibition of the content regardless of context or intent, coupled with tight restrictions on searching and sharing of recently uploaded video.

Both the decisions the firms made and the speed and adequacy with which they implemented them in a difficult circumstance will be—and should be—subject to debate and criticism. But it would be a grave error to imagine that broad legislative mandates are likely to produce better results than such context-sensitive judgments, or that smart software will somehow obviate the need for a difficult and delicate balancing of competing values.

I thank the committee again for the opportunity to testify, and look forward to your questions.

Mr. ROSE. Thank you to all the witnesses.

I would like to now recognize Chairman Thompson if he would like to make an opening statement as well.

Mr. THOMPSON. Thank you very much, Mr. Chairman.

In the interest of time to get to our witnesses, I will just include my written statement for the record and we can go into questions.

I yield back.

---

[13] David Kaye, Joseph Cannataci, and Fionnuala Ní Aoláin "Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; the Special Rapporteur on the right to privacy and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism" *https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234*.

[14] Faiza Patel, "EU 'Terrorist Content' Proposal Sets Dire Example for Free Speech Online" (Just Security) *https://www.justsecurity.org/62857/eu-terrorist-content-proposal-sets-dire-free-speech-online/*.

[15] "Letter to Ministers of Justice and Home Affairs on the Proposed Regulation on Terrorist Content Online" *https://cdt.org/files/2018/12/4-Dec-2018-CDT-Joint-Letter-Terrorist-Content-Regulation.pdf*.

[16] Kate Klonick, "Inside the Team at Facebook That Dealt With the Christchurch Shooting" *The New Yorker* (April 25, 2019) *https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting*.

[17] Elizabeth Dwoskin and Craig Timberg "Inside YouTube's struggles to shut down video of the New Zealand shooting—and the humans who outsmarted its systems" *Washington Post* (March 18, 2019) *https://www.washingtonpost.com/technology/2019/03/18/inside-youtubes-struggles-shut-down-video-new-zealand-shooting-humans-who-outsmarted-its-systems/?utm_term=.6a5916ba26c1*.

[18] See, e.g., Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran "Deceiving Google's Perspective API Built for Detecting Toxic Comments" Arvix (February 2017) *https://arxiv.org/abs/1702.08138*.

Mr. ROSE. All right. I want to respect your leadership as well and cede the floor to you for opening questions as well, Chairman.

Mr. THOMPSON. Thank you very much.

We will be holding a hearing with some of the companies tomorrow, and we are hoping to hear from them their position on some of the content.

If you had an opportunity to ask those questions of the companies—and I will start with Dr. Buchanan—what questions do you think our committee should ask those companies about the terrorist content and what they are doing to protect their systems?

Mr. BUCHANAN. Thank you, Congressman. I think there are a couple categories of inquiry you might pursue. One is how they plan to scale moderation to platforms that stretch to several billions of users. They will, I think, be optimistic about AI. As I said in my opening statement, I am less optimistic, and I think it is worth probing on if they do believe such a system can scale and function, how machine learning will overcome the problems of context and pattern matching that all three witnesses identified.

As I said as well, another topic of discussion might be recommendation, algorithms, and their potential forces a tool of radicalization. I think companies could share more data about that and open that up to further study.

Mr. THOMPSON. Mr. Stamos.

Mr. STAMOS. So I think Ben had some excellent questions. I think the two areas I would focus on is, first, I think a general problem the companies are facing on a bunch of different safety and content moderation issues is that their decision-making process is extremely opaque and the outcome of those decisions are very opaque. So I would be interested in hearing what plans they have to document the standards under which they are making these decisions and whether or not they would be willing to build capabilities for academics, such as Dr. Buchanan and myself, to have teams that have access to information that has been moderated.

One of the challenges for outside groups studying these issues is that under kind-of current privacy laws and the terms of service of the companies, it is very difficult to get access to data once it has been deleted. So if they are taking down pro-terrorist content, it is very difficult to understand—for us to understand who is posting that, what does the content look like, what is the possible effectiveness, because once it is deleted, it is completely gone. I think asking them about building mechanisms so that other people can look over their shoulder and provide both useful feedback to them but then also information for Congress and the general public I think is really important.

Mr. THOMPSON. Thank you.

Mr. Sanchez.

Mr. SANCHEZ. I think I would want to explore values questions. It is fairly easy in the absence of countermeasures to take down, you know, a particular kind of video of violence or of a hateful speech. So I would add to that, not just how you identify the video, but how do you judge the context, how do you determine the balance of interests when content might be uploaded as part of a critique, and how to deal with ambiguous cases where, you know,

someone may have a complicated attitude toward an extremist group.

Finally, I would ask them to explore the question of false positives. Again, as I highlighted previously, the issue of potential removal of evidence of war crimes and other socially valuable content that may be easily mistaken for harmful extremist content.

Mr. THOMPSON. Mr. Sanchez, you mentioned Congress needed to kind-of step lightly in terms of some of what we are doing. You know, Germany kind-of moved forward in doing some things in this area that was kind-of limiting for companies, but the companies followed.

What is your response to something like what was done in Germany?

Mr. SANCHEZ. Well, Germany has a very different legal context. They prohibit images of the swastika and the Holocaust denial, a wide range of speech related to Nazism, whereas our Supreme Court has upheld the Constitutional right of Neo-Nazis to march brandishing swastikas in public speeches.

What everyone's view on that is, that is the context here, and the ability of Facebook or YouTube to be more restrictive and say, we don't want Nazis on our platform depends on there being a private entity. As soon as they are viewable as a state actor as acting in coordination with the state or at the mandate of the state, the First Amendment applies. So we would risk, you know, either just the striking down——

Mr. THOMPSON. I understand. So even though the companies complied, they didn't look at it as, you know, this is a major income producer for our company and we should follow the laws of the country if we want to continue to do business in this country?

Mr. SANCHEZ. They clearly have decided to do that, and they are, to some extent, able to do that; although, again, there are, I think, value-based questions about whether you want to automatically remove, you know, all, for example, historical content related to Nazis.

Mr. THOMPSON. Mr. Stamos, you want to comment on that?

Mr. STAMOS. Yes, just real quick, Mr. Chairman. I just want to point out, NetzDG, the law we are talking about in Germany, explicitly tied the enforcement policies to existing German law. So the rules that they wanted Facebook and Twitter and such to enforce are rules that the Germans would enforce themselves. The difference is they moved the action out of the courts into those companies but under existing German precedent.

Almost all the speech we are talking about is legal in the United States. So as horrible as it is, the Christchurch shooter's video, the manifesto, are Constitutionally-protected speech in the United States, and the companies are acting on their own volition to take it down. So I think there is a significant difference here of the context in which the law was passed.

Mr. THOMPSON. Doctor.

Mr. BUCHANAN. I think I will defer to my colleagues on the legal questions here. I stick to international relations, and I am not a lawyer.

Mr. THOMPSON. Thank you. I yield back.

Mr. WALKER. Are you bragging? No.

Mr. ROSE. The Chair now recognizes Ranking Member Mr. Walker for questions.

Mr. WALKER. Thank you, Mr. Chairman and Mr. Chairman.

Mr. Sanchez, you note in your testimony that despite criticisms leveled at social media companies regarding content on their platforms, the companies are going much further in moderating content than the Government could and should do with a very clear warning against attempted Government intervention in this process.

That being said, do you think that tech companies should be doing more to self-police for terror and other violent content because, despite the company's efforts, the public concerns continue to grow?

Mr. SANCHEZ. You know, I think that is—that is hard to do in a generic way. I think there are case-by-case determinations to be made about what kind of speech is evolving around, you know, a particular piece of content or a particular group. So saying they should do more, should do less is hard to divorce from specific considerations at issue.

Mr. WALKER. OK. Let me dig in a little bit more. Let me go to Dr. Buchanan. Do you have an estimate of what the error rate is for extremist content flagged by AI? This could include news, educational materials, even counternarratives that are flagged as extremist content, but clearly are not. Do you have any percentage on that?

Mr. BUCHANAN. No. I don't think there is any publicly-available percentage. We have seen numbers from different academic studies that vary widely of how much content gets missed as a false negative, how much gets missed as a false positive. I think, as Mr. Stamos said, we would do better in studying this topic if there were more data available—made available by the companies, and right now, that is not the case.

Mr. WALKER. Mr. Stamos, you were the chief security officer for Facebook. Can you tell me those dates again, please?

Mr. STAMOS. June 2015 to August 2018.

Mr. WALKER. OK. You said there is a lot of misunderstanding and confusion and hype. From whose perspective? Who has the misunderstanding, who is confused, and who is hyping this?

Mr. STAMOS. So as I said, the companies themselves are hyping it. There is a number of independent companies who are trying to sell products who like to hype the capability of AI. There is a kind of a media, moral panic around some of these technologies that I think increases the coverage.

Mr. WALKER. OK. I believe you said, regarding the Christchurch shooting, quote—I believe you said some things could have been made or done differently with that?

Mr. STAMOS. Yes.

Mr. WALKER. You listed 7 things. My question is, were these things that should have been implemented during your time as chief security officer there at Facebook? This shooting only happened months after your tenure was completed. Do you know why or were you forbidden from implementing these things that could have made a difference?

Mr. STAMOS. I was never forbidden from implementing these things. So when I started at Facebook, the No. 1 content issue was

the Islamic State, and that was the No. 1 area of focus, both in building the investigative teams, which worked for me, and then the engineering teams that were elsewhere in the org.

This is an adversarial problem, so all of these issues are dynamic issues where you have intelligent adversaries who are changing their game based upon your moves.

So I do think there are things that I would have done differently if I knew what was going to happen for sure, but there is this kind of theory—I heard the term from the Chairman of like wanton disregard from the companies, and that is not, like, my perspective from somebody who worked for it. There is more Ph.D.s working on the link between violence and on-line content at some of these companies than in most of academia. There is a lot of people working on it, but it is actually not an easy problem.

Mr. WALKER. How do you suggest social media companies incorporate the issue of wrongful flagging into their AI policies?

Mr. STAMOS. So, as I said both in written and oral testimony, I think the transparency is key. The companies are operating as quasi-governments, right. They are making decisions on a global scale that would normally be reserved for governments, governments that, in our expectation, should be democratically accountable. I think the lack of transparency about what decisions they are making and how they make it is a critical problem that they have.

Mr. WALKER. With all that said, and I am going to get a quick answer because we have about a minute left, from all the panel members, is the consensus of the panel that AI in its current state is not a place to be the primary technology used to detect and block the extremist and other violent content? Dr. Buchanan.

Mr. BUCHANAN. Yes. I think there is substantial limitations to the technology as it applies to the moderation problem.

Mr. WALKER. OK. Mr. Stamos, would you——

Mr. STAMOS. I think AI is a critical part of doing it at scale, but most of the hardened decisions have to be made by humans.

Mr. WALKER. OK. Mr. Sanchez, would you expound? Got about 30 seconds left.

Mr. SANCHEZ. It is best for the easy cases. It has a large role in rapidly dealing with easy cases.

Mr. WALKER. I just want to make sure that this is something that is going to be long-term and some of the implementations that need to take place. I do appreciate your testimonies.

I am running out of time without another question, so with that, I will yield back to the Chairman.

Mr. ROSE. Thank you, Ranking Member Walker.

On that, I will recognize myself for 5 minutes.

I want to focus, Mr. Stamos, on your time at Facebook. Obviously, an enormous amount of content posted on Facebook every minute. Facebook, as well as Twitter, YouTube, all have agreed that they do not want terror content on their platforms.

How does Facebook manage this? How did they do it during your time? What is your reading of how they do it now, your understanding of personnel allocations, resource allocations? What—what is not working right now at Facebook? What are they not doing well enough?

Mr. STAMOS. So that is a comprehensive question. So the experience of starting, you know, my—starting my tenure was right around, like I said, the growth of the Islamic State's use on-line. Islamic State was unique among in the history of Islamic terrorist groups in it being staffed by digital natives, that you had millennials who worked for the Islamic State living in Western countries who were very smart about the manipulation of social media.

One of the failures at the company going into that is that our policies did not really capture that kind of organized use of the platform to cause terror and to recruit people for terrorist organizations.

A lot—this is a problem that we had in a couple of cases. We had the same issue with the Russian interference in 2016, where going into the 2015–2016 time line, Facebook's policies were very focused on individual pieces of content, not looking at the big picture of organized campaigns that were trying to manipulate the company.

So, coming out of that, we ended up doing a couple things. So we ended up hiring a bunch of experts in this. So I actually cited a paper from a friend of mine, Brian Fishman, who we hired from the West Point Counterterrorism Center. We hired a bunch of specialists in Islamic and, eventually, in like white nationalist terrorism to help build up policies to catch those folks.

We built an investigations team to investigate the worst cases, and then built out content moderation policies to catch them as well as the AI that come with them.

As far as what the current staffing is——

Mr. ROSE. But I just want to—I think we agree that the policies right now is not the issue. There is a commitment, theoretically. We agree on that.

Mr. STAMOS. Right.

Mr. ROSE. What my question to you is, is that it is really after the policy gets established. Because you all agree collectively that AI doesn't work well enough to just rely on it.

So this is a numbers game. Once you have the policy in place, this is a numbers game. So after that point, you get the experts, you get the Ph.D.s, everyone understands what you are looking for. Did you feel during your time there that Facebook had enough people on staff that were dedicating enough resources to actually catch all of this stuff?

Mr. STAMOS. Well, you are never going to catch all of it.

Mr. ROSE. Catch enough of it.

Mr. STAMOS. Catch enough of it. I think they could always invest more, the truth is, and that—during my time there, there was——

Mr. ROSE. Did you at any time ask to invest more?

Mr. STAMOS. I did and we did.

Mr. ROSE. What did Facebook tell you when you asked to invest more in counterterrorist screening?

Mr. STAMOS. So I received the staffing request I got to recruit people to do this kind of work, so recruit people from the FBI, from NSA and such, to build up a team to investigate these issues.

I think the more important staffing was not really on my team as much as in the content moderation teams, right? So, you know, my team was doing the kind of—the investigations of the worst-

case scenarios, you know, the investigation after the Paris attack, the investigations of people who might be planning terrorist attacks.

The content moderation is where they have really grown the size of the team, and that is where you need it. Because one of the issues in international terrorism is diversity of languages you have to deal with is actually incredibly broad.

Mr. ROSE. Absolutely.

Mr. STAMOS. So that is where there has been a bunch of growth. Whether it is properly sized or not, I am not really sure. I mean, I think they can always put more resources.

My recommendations are more on some big-picture changes. I think one thing I would like to see the companies do is the difference with Christchurch versus the Islamic State is that we are now dealing with a different kind of problem where these white supremacist groups have on-line hosts who are happy to host them. That was not true for Islamic State. People did not want to be the official IT people for the Islamic State, but people are happy to be the official IT people for people like the Christchurch shooter. I think they would be more aggressive dealing with that problem.

Mr. ROSE. With my limited time, this is also a collective action problem. So what was your interaction with the GIFCT, and what do you think the future of the GIFCT should look like?

Mr. STAMOS. That is a great question, sir. GIFCT was set up during my time there. I was a big supporter of it, but I think on all of these issues on interference in elections, on terrorism, on a bunch of different safety issues, the time has come for the companies to put together a stand-alone staffed organization that has its own people studying these issues and can really serve to push the coordination problem.

The one I really like is called the FS–ISAC, the financial services industry. All of those companies hate each other, right? The big iBanks in New York. They all hate each other, but they realize that all of their boats rise and fall on the same tide, which is the trustworthiness that people place in the banking system. So they have come together with a well-staffed, well-resourced organization that helps force them to share information, and then operates as a bridge between them and the Government. I think it is time for the companies to build something like that, and GIFCT could become, basically, a working group, but a group like that could work on all of these issues in a much better way.

Mr. ROSE. Thank you.

I will now recognize the good Congressman from Rhode Island, Congressman Langevin.

I am sorry. I passed over you, buddy. Congressman Green, my friend.

Mr. GREEN. Thank you, Mr. Chairman, and thank you for teeing all this up.

I want to thank the witnesses for coming today. My colleague actually asked the question for Mr. Buchanan that I wanted to ask about capabilities, so I will move on to Mr. Stamos.

Sir, you make many interesting recommendations in your written testimony that may help significantly improve companies' ability to detect extremism, specifically you say, create a robust perceptual

fingerprinting algorithm and develop client-side machine learning for safety purposes.

Can you elaborate on what is holding companies back from doing just those things? Have you discussed with the companies these suggestions in detail, and what is their feedback?

Mr. STAMOS. Thank you, Congressman. So to do a little bit of background, probably the most—the best story in all of social media of a problem that they have handled the best is child sexual abuse material, which is the term that we use for child pornography. The reason for that is a couple. No. 1, the law is clear, the definition of these things are clear, people are very motivated to work hard on it. Also, there has been enough motivation to create standards by which the companies can help each other. So the key technology here was one that was invented at Microsoft, something between 10 and 15 years ago, called PhotoDNA, that now all the big companies utilize.

My recommendation is that I would like to see further improvements on that technology. There are significant technical issues with the PhotoDNA. Effectively, it has to be kept secret, because if it was ever publicly known, the child pornography people would be able to defeat it and be able to upload their images. So it is not robust is what we call it.

So I would love to see the companies come up with a new set of algorithms for a variety of different purposes, including video, which PhotoDNA does very poorly, that they could then share with many, many companies.

One of the issues here is that there is a lot of companies that host this kind of content and they are not getting help from the big guys. So I would like to see the big guys pay for the creation of that kind of stuff, and then that would let them share all of these hashes with each other. So, like, in the Christchurch shooter video, if one of those companies said this video is bad, just like with C–SAM, they would be able to ban it very, very quickly, and then create better algorithms that can deal with people intentionally manipulating it. That was a big problem with the Christchurch videos. People were trying to defeat the algorithm.

Mr. GREEN. Making the comparison, though, between, let's say, child pornography and, let's say, Christchurch, though, I think is a little bit problematic. It is always bad when there is a little child in an image, right, like that. It is always bad. But if the Peshmerga is trying to show how ISIS has beheaded Christians, and so the algorithm is gore or whatever, that is different and more difficult to differentiate, because that may be showing something that is very horrible, but you want it out there so that people can be aware of the fact that ISIS is doing this.

So that kind-of leads me to my question for Mr. Sanchez. You mentioned in your written testimony how many times content is taken down by good groups who are standing up and fighting evil such as those documenting Bashar al-Assad's war crimes. Are you hopeful that technology can actually do the differentiation that I alluded to in my example, and what is standing in the way of them getting good enough to make that differentiation?

Mr. SANCHEZ [Off mic]. As our waitlisting progress, this particular channel, this particular group, we understand is engaged in

an educational or a sort-of civil—civil rights or human rights mission, and so we will sort-of exempt some of their content from the algorithm. In the near term, that is what you can do.

If you are talking about a situation where you may have random citizens uploading content that they capture on their phone because they were present at an atrocity, that may not be a person you have any reason to treat specially, and, indeed, there may be nothing about the content in itself that distinguishes it. The same video that someone posts to say, look at the atrocities these groups are committing and how important it is to continue fighting them and keep the world's attention, might be the same video posted by another group to say, look at our glorious triumph in killing infidels.

Mr. GREEN. Right.

Mr. SANCHEZ. You know, I don't think a computer or any other algorithm can tell you the difference between those two cases because it is the same video.

Mr. GREEN. So, basically, you are saying there is no chance of fixing that with AI?

Mr. SANCHEZ. I think it is—there are ways AI can do it, maybe by looking at context and seeing words, but human communication is complicated and bad actors adapt with countermeasures. They change the way they talk, they change their patterns. So even if it succeeds in the short term, it is always a temporary success.

Mr. GREEN. Thank you.

Mr. Chairman, I yield.

Mr. ROSE. Thank you, Ranger.

I would like to next recognize Congressman Langevin from Rhode Island.

Mr. LANGEVIN. Thank you, Mr. Chairman. I want to thank our witnesses for being here today.

Mr. Stamos, Mr. Buchanan, I certainly appreciate your sober assessment of the current state of machine learning technology. I know this is an evolving technology that we are still kind-of coming to grips with in understanding and realizing its potential and also its limitations.

While I believe there is a lot of promise for machine learning going forward, I think many of the use cases that we see today are much more discrete with respect to terrorist messaging and misinformation, for example. So much of the attention you would say has been placed on dealing with the content of the speech itself, but do you believe that it makes more sense to focus on the actors themselves? What is the potential for machine learning to help companies better understand their users and move accounts that violate terms of service?

For Mr. Stamos, for Mr. Buchanan, whoever wants to start.

Mr. STAMOS. Thank you, Congressman. You are right. Actually, the use of machine learning from a metadata perspective that looks at users is actually probably one of the most effective uses. So during—if you look at all kinds of abuse on-line, a huge percentage of all kinds of abuse, ranging from bullying and harassment to uploading of child exploitation, is powered by fake accounts, which you expect, right?

People generally don't commit these crimes under accounts that they use for their real personas, and that is actually one of the most effective uses of machine learning at Facebook. Especially after the Russian issue in 2016, we built out a number of algorithms that were trained on fake accounts to look at do these accounts have the indication of being created specifically to violate policies. Do they look inauthentic?

I think—so in my written testimony, I cited a Facebook report, and they report that around 2 billion of those fake accounts are now taken down every quarter, which is a huge increase than when I was there. So I do agree that that is one of the good uses of ML for this.

Mr. LANGEVIN. Mr. Buchanan, do you have something to add?

Mr. BUCHANAN. Yes. I think I would agree with everything Mr. Stamos said. I think it is worth maybe spelling out slightly more as to why this is at least a partially effective tool, particularly in how it raises the cost for adversaries in their operations.

So much of social media, so much of this discussion is about scale, and by being able to interfere with the account-creation process and the posting process, requiring adversaries to go through more hoops, in effect, to post their content on-line and keep it on-line, we raise the cost for the adversaries.

In that respect, I think machine learning that is applied to accounts and to users certainly is an effective tool, both for propaganda operations and for extremist content more generally.

Mr. LANGEVIN. Mr. Stamos, you highlight two particular areas social media companies should focus on: Advertising and recommendation engines. So how much should machine learning play in the advertising business decisions, and how—and how is machine learning being used today in recommendation engines? What changes could you—and Dr. Buchanan, I would welcome, sir, your input as well—to suggest such algorithms to help stem radicalization?

Mr. STAMOS. Thank you, Congressman. You make a great point. The use of machine learning in advertising is a critical thing, I think, we have to think about. The reason on-line ads are at all effective and the reason these companies make any money is because of machine learning, because machine learning allows for people who advertise on their platforms to reach people who are more likely to click the ads.

It does create a number of problems. It creates an inherent biased problem that is not so relevant to our topic, but that is some of the issues around housing ads and employment ads and stuff come out of the machine learning algorithms kind-of reflecting human biases without them knowing that they are doing so.

The other issue that I have with the machine learning in advertising is its use in political campaigns. So we actually released a report a couple weeks ago that I cited in my testimony from Stanford around election security. One thing we would like to see is restrictions on the use of the most powerful machine learning targeting techniques in political campaigns, specifically around securing our elections from both domestic and foreign interference. I think that is less relevant in the terrorism context as we don't have a lot of examples of terrorists using advertising, right, giving actual

credit card numbers, but I think in other forms of abuse that is a big deal.

Mr. LANGEVIN. Thank you.

Mr. Stamos, if I could—well, my time is about to expire. Maybe you can talk about this in an email——

Mr. ROSE. We will do a second round.

Mr. LANGEVIN. OK. I will stop there, and I yield back.

Mr. ROSE. Great. I would love to now recognize Congresswoman Jackson Lee from Texas.

Ms. JACKSON LEE. Thank you very much, Mr. Chairman.

Let me just ask the general question, since this committee has tried to secure information regarding how much resources are utilized for finding bad actors, bad information on the social media and with respect to the tech family, if you will. So let me just ask each witness.

Dr. Buchanan, do you think the industry overall needs to invest more resources in interdicting bad information?

Mr. BUCHANAN. Yes. On balance, I think there probably is space for further investment. I think there is also space for significant policy changes regarding transparency. I am not sure which would give bigger impact in the long run, but both probably are useful.

Ms. JACKSON LEE. Dr. Stamos.

Thank you.

Mr. STAMOS. Yes, ma'am. I do think there needs to be investment, both by the big companies, and I think we need to figure out societally how we get support for the smaller companies as well because that is where I think there is actually kind-of a real blind spot that we have on some of these issues.

Ms. JACKSON LEE. Dr. Sanchez—Mr. Sanchez.

Mr. SANCHEZ. I lack the same internal perspective as Alex to judge sort of the adequacy of their current internal investment, but certainly, you know, there is room for improvement.

Ms. JACKSON LEE. As relates to AI, there are concerns about—well, let me just ask this question: How important is data when it comes to artificial intelligence, and is it more important than algorithms or computing power? What are the strengths of each of these machine learning components?

Dr. Buchanan.

Mr. BUCHANAN. That is a very broad question. I think as applied to counterterrorism, this is an instance in which the systems in play are classifiers, and those rely quite a bit on previous examples, so the data is incredibly important. I don't think for a company like Facebook or Google this is a problem of computing power or machine learning research talent.

Ms. JACKSON LEE. Dr. Stamos. I have a follow-up question for you.

Mr. STAMOS. You are absolutely right. The availability of data is actually the critical thing for most uses of machine learning. Anybody can rent a huge amount of computing power from the Microsofts and the Googles and the Amazons of the world, but getting training sets is extremely difficult.

Ms. JACKSON LEE. I would like to say that we are not picking on one company verse another, the fact that you are formerly with Facebook, but it has come to our attention that, although Facebook

indicates that they intercept about 83 percent of this terrorist data, a whistleblower indicated that they really only claim less than 30 percent of the profiles of friends of terrorist groups had been removed from the platform over a 5-month period.

So my question would be, are we getting numbers that are really not accurate? Is there a way to ensure that the likes of Twitter, Facebook, and others, Google owns YouTube, are really getting at the core of these terrorist information bases, if you will?

Mr. STAMOS. In my experience, the numbers the companies share are accurate but often incomplete, right? So when you look at the numbers in the transparency report, it doesn't include certain numbers such as, you know, multiple people have pointed out, some of these numbers on 99 percent caught. The denominator there is of the content that they ended up taking action on, it is not on the prevalent—the number we call prevalence, which is all the stuff you missed, right?

What I would like to see from a transparency perspective is I would like to see all the companies to set up either together or separately archives of all the content they are doing moderation on and the ability for academics to access that under NDA so that we can look over their shoulders and provide feedback to you of what we think is going on.

Ms. JACKSON LEE. Let me ask this question for all witnesses: Can bad data lead to bias, and how do you improve the quality of data that is being used for AI?

Mr. BUCHANAN. Yes, Congresswoman. I think it is very clear that, generally speaking, bad data or mislabeled data, biased data, can lead to bad outcomes for AI systems. It is very important, regardless of what kind of machine learning one is doing, to have a data set that is large and representative of the outcomes you want to achieve.

Ms. JACKSON LEE. Mr. Stamos.

Mr. STAMOS. I agree that bad data can lead to bias. Bias can also come out of good data where the AI pulls out human biases that we were not conscious of. So I think the other thing we need is we need—and this is an area of a lot of academic study, we need measurements of fairness and bias that apply even if the training sets are perfect, and I think that is kind-of the next generation of the issue here.

Ms. JACKSON LEE. Mr. Sanchez.

Mr. SANCHEZ. I think I would agree with that and say that, you know, I think a lot of—when people talk about algorithmic bias, often it is less—often it is a question of training sets, you know, white engineers training a set on a lot of white folks and then, you know, the recognition doesn't work for people who look different.

But a lot of the time, you know, algorithms are what they are learning from is human behavior and human preferences, and so to the extent we have cognitive bias, the algorithms we teach inherit those biases from us.

Ms. JACKSON LEE. Thank you.

Thank you, Mr. Chairman. I yield back.

Mr. ROSE. Thank you.

I would like to now represent for a second round of questions, Ranking Member Walker.

Mr. WALKER. Thank you, Mr. Chairman.

I don't think I will use the full 5 minutes, but I wanted to just to dial down a little bit of something. In fact, just about 30 seconds ago, Mr. Stamos, you talked about measurements of fairness. In your time at Facebook, did you feel like there were any biases toward those held conservative or religious viewpoints? Were you able to identify any kind of biases, bad biases or whatever the terminology might be?

Mr. STAMOS. No.

Mr. WALKER. OK. I have got about 4 pages of things that would—even some that Facebook later on issued an apology with—we have Brian Fisher, the president of Human Coalition has been back and forth.

Are you familiar with that name at all?

Mr. STAMOS. No, I am not, sir.

Mr. WALKER. Human Coalition?

How about the group Zion's Joy, familiar with that group?

Mr. STAMOS. I am sorry.

Mr. WALKER. Well, let me refresh your memory just a little bit there. Zion's Joy is a gospel music group, a multiracial gospel group, who engaged, during your time with Facebook, and the *New York Times* picked up on it. Eventually, Facebook offered an apology to that.

Do you have any recall of that situation that went back and forth?

Mr. STAMOS. So to be clear, I worked on adversary use of the platform to cause harm. I was not part of kind-of political content moderation. The overall issue here is—you know, according to this report, Facebook made about 4 billion content moderation decisions in the quarter.

If you had 4 billion stars you are looking at, you can draw any line or any constellation you want. The truth is, is basically, every group believes that Facebook is biased against them. We would hear this from every racial group, every political group. Everybody that believes they are being oppressed. The truth is the companies make lots of mistakes.

Mr. WALKER. I don't believe I have heard a lot of that from groups that would promote progressive or left causes that would suggest that you have Facebook biases against them. But the reason I ask you——

Mr. STAMOS. I will share my Twitter DM with you sometime.

Mr. WALKER. Well, we all—I mean, yes, we can exchange Twitter any time.

The point that I am making is, when the algorithms—those are human uploaded. Those are human—basically, broken down to a way that some kind of human input—there used to be a—when I was in middle school, we talked about COBOL and FORTRAN, garbage in, garbage out, right?

Mr. STAMOS. Yes.

Mr. WALKER. So when you talk about these bad biases, do you think it is an issue that some of these algorithms are loaded with a bias as they are put into the system to monitor such content?

Mr. STAMOS. I think the bias of individual moderators is absolutely a problem. That is true globally. At Facebook, we had all

kinds of issues with individual moderators being from a certain political group, a certain ethnic group, and having biases against others. I am sure that has existed in individual moderators in the United States. But the companies are pretty well aware of this. There is an internal QA process to try to work on it.

I must say, like, I have never been in a meeting where people have talked about let's shut down conservative thought. Like, that is not how——

Mr. WALKER. Right. I don't think anyone would be that blatant, even if that was their intent.

So you talked a lot about transparency. You don't have an issue of people looking into this to see if there are bad biases.

Mr. STAMOS. Right.

Mr. WALKER. Correct?

Mr. STAMOS. That is one of the problems the companies have created for themselves, is they can't disprove that there is no bias, because there is no record of everything they have moderated. So that is why I think if there was a massive database of here are all the things that they have made a decision on in the last 90 days, then you could have outside academics look and see whether the bias exists.

Mr. WALKER. Mr. Sanchez, do you have any evidence, or even if it is a hypothesis from what you have seen, any kind of bias slanted in big social media to the left, to the right? Have you guys been able to identify any of that?

Mr. SANCHEZ. I will say, in my experience, it does seem like everyone—everyone is convinced that they are on the wrong end of decisions. I think it is because people pay attention to, you know, cases similar to them. If you are a conservative, you are likely to hear about conservatives who are angry about moderation decisions that are affecting them. If you are progressive, you are likely to hear about, you know, the person who was arguing with a Nazi and they got banned and the Nazi didn't. So that creates an impression that the universe of questionable moderation decisions affects your group. I don't know how well that reflects reality.

Mr. WALKER. Very informative. Thanks for letting me drill down a little bit longer.

I yield back.

Mr. ROSE. Thank you, Ranking Member.

I will now yield to myself for 5 minutes.

Mr. Stamos, go back to you. Can we just focus on, for a second, the nitty gritty of examples where AI was deficient? We have heard things about watermarks. We have heard things about the absence of spacing. Without, obviously, giving away tips to people, what are some egregious examples, because I know they exist, of the failures of AI as it pertain to terrorist content?

Mr. STAMOS. So to go back to the Christchurch video. I tried, in my written statement, to break into three different categories where I think we could do better. One of the—the obvious ones where we can have the most improvement is on the last step of the propaganda campaign, which was the shooter engaged an on-line community off of the big sites on a small site called 8chan to then push his propaganda for him. They were then exchanging among themselves tips on, how do you defeat the filters of the companies?

One of the reasons they had a good understanding of this is there is a big subculture of people who upload copyrighted videos to places like YouTube. So they have been working for years on how to defeat these algorithms. The algorithms, while perhaps slightly different between copyright and counterterrorism, have the same kind of fundamental problem.

Mr. ROSE. So it does seem that we are always going to be chasing this.

Mr. STAMOS. Yes.

Mr. ROSE. AI will get better. People will get better.

Mr. STAMOS. Yes.

Mr. ROSE. Should there be a mandated ratio of screeners to users?

Mr. STAMOS. You know, that is a tough question. I am not sure how you would estimate that.

Mr. ROSE. That is why you get paid the big bucks. It is clear you all have the consensus, AI, as far as the eye can see right now, will never be good enough. Facebook, 2 billion, 3 billion, 4 billion users. We need to figure out a way to create some type of standard for what right looks like. It appears to me that the best option right now is, say, look, you got 2 billion users. We need—we want a 20-to-1 ratio.

What they often come back to us with, Twitter, YouTube, they all say the same thing is, well, you are oversimplifying this. You know, we could have better technology. Our response to that, based off what you are saying, is, no, you cannot. That this is a personnel problem, this is a resource problem. You need to have screeners. They cannot be underpaid, overworked contractors either.

Do you think that is an unfair statement?

Mr. STAMOS. No. It is crazy for me to be saying this, but I think the people who have one of the more thoughtful approaches of this is actually the French Government right now, in that they are proposing a model of a self-regulatory agency where Western governments, democracies are able to participate in. That holds the companies accountable to the rules that they set out for themselves.

So I think a first step is to make them standardized on these are the standard numbers that we are going to share in all these different areas, and then push them to be accountable to their current standards. I think you can do that without trying to mandate specific numbers, which I think the technology will get passed legislation pretty quickly there.

I mean, like, one of the founding laws that we have to work off of on all these issues is the Electronic Communications Privacy Act which was signed by Ronald Reagan, right? These laws last forever and are very difficult to apply sometimes by tech companies. So the French approach is actually the one that I think is the best right now to try to push the companies to do better.

Mr. ROSE. OK. Just to close out, then. We have also been struggling and exploring ways that we can establish this standard. Because I think the social media companies right now rightfully fear that we yell and scream and then we are not telling them what we want of them.

So do you think that the best route is for us to push and potentially even mandate that they develop an industry standard, that

we can then establish an independent body to hold them to? Should we push a standard to them?

I give this to the three of you. What recommendations do you have for how we can establish that standard and if we should?

Mr. STAMOS. From my perspective, I think—I take Mr. Sanchez' point well, which is I think it is very difficult for the U.S. Congress to establish standards, because almost all the speech we are talking about is First Amendment protected in other contexts. Holding them to the standards they create I think is a totally appropriate thing for Congress. The other thing I would push for is the transparency, so at least we know what is going on and we know what kind of decisions——

Mr. ROSE. Keep in mind, we are talking just about terror content. We are not moving to hate—and it is very purposeful here.

Mr. STAMOS. Yes.

Mr. ROSE. We are not moving to hate speech. We are not moving into things that rightfully—that have significant legal and Constitutional issues around that.

We are talking here about someone getting up and saying, I urge you all to join Hamas. I urge you all to join Hezbollah. I urge you all to join al-Qaeda, and any other organization listed by the State Department as an FTO.

You don't think that we can establish a standard around that?

Mr. STAMOS. I am not a lawyer. I am an engineer. My understanding, especially when we are talking about, like, in the Christchurch and the white supremacist content issue, is that most of this content is protected by the First Amendment. But that is my understanding.

Mr. ROSE. Anyone else.

Mr. SANCHEZ. I mean, I think there may be a narrow category of sort-of direct support for violent action that would be, you know, subject to legislation without raising First Amendment issues. But, in general, advocating violence saying I approve of al-Qaeda, I think what they do is wholly injustice. Vile, but it is protected speech. So I think the creation of standard—or requiring companies to create a standard, unless you are willing to accept literally any standard, including we do whatever we want, creates Constitutional problems.

Holding them to standards they publicly articulate I think can be done by the FTC. If you are not upholding standards you say you adhere to, you could pursue that as an unfair and deceptive trade practice.

Mr. BUCHANAN. I think that is right. Again, I will skip the legal side of this. But it does seem to me that it is very difficult, especially for those of us who are academics, to get a good sense of how the companies are doing on this. As some of the earlier questions mentioned, there is a very broad range of numbers of what percentage of content is being managed appropriately. Absent raw data from the companies, it is exceptionally hard to hold them accountable regardless of which standard is applied.

Mr. ROSE. Right. I would like to commend the committee staff for the first panel in the history of Congress with no lawyers on it. That is great.

Congressman Langevin from Rhode Island.

Mr. LANGEVIN. Thank you, Mr. Chairman.

I just want to go back to a couple of my earlier questions. So, Mr. Stamos, we talked about advertising. We really didn't get to the recommendation engines part of the question.

So how is, again, machine learning being used today in our recommendation earnings? What changes do you suggest in making such algorithms to help stem radicalization? Again, Mr. Buchanan, if you want to comment as well on that.

Mr. STAMOS. Thank you, Congressman. As I tried to put in my written testimony, I do believe that advertising and recommendation engines are the first things that we need to focus on here. The thing about recommendations, No. 1, they work in very different ways for different platforms.

So in some platforms, like a Twitter or a Facebook, the biggest determinant of the content you see is who you are connected with in the social graph. On sites like YouTube, it is the recommendation engine that is putting it in front of you.

Effectively, most recommendation engines are machine learning, and they are based upon the idea of statistically trying to understand what it is that you want to consume on their site based upon your previous actions. So it tries to show you stuff that the machine guesses that you will like.

The problem there is the machine doesn't understand whether it is actually good for you or not, right? It is using a metric of whether you are satisfied or you are happy with the content that does not figure out whether or not the content is good or whether it is good in general.

I think one of the key things that we need, our recommendation engines, is you need what are called counter metrics. You need metrics that are a little more complicated and nuanced to measure whether or not you are putting information in front of people that is radicalizing. Is it pushing the edge of the content policies? Does it create people to fight with one another? I think there are ways to do that in a content-neutral way that the companies need to consider.

Mr. LANGEVIN. OK. Thanks.

Mr. Buchanan.

Mr. BUCHANAN. I would only add that this is one place where I think you can, at least in the abstract, see some tension between what might be good more generally for a society and the business models of particular companies. The recommendation engines exist to keep people on the platform, to keep people engaged with the platform. To some degree, it probably is good business to have something like this.

I think, as Mr. Stamos said, we probably want some more nuance on those recommendations to balance the business desires of the companies with what appear to be, based on the limited data that is available, some significantly negative broader social effects.

At a minimum, I think we want more transparency onto the recommendation systems such that we can study how and if they are driving people toward more extreme content in any given political direction or direction related to extremism or terrorism.

Mr. LANGEVIN. How do we get to that? Is that regulation? Is it legislation?

Mr. BUCHANAN. I think this is a fair question to ask the companies tomorrow. It seems to me that there is the opportunity for companies to make this data available to researchers to study, not just on what content they take down, but also on, to some degree, the functioning of their recommendation systems. That would shine some light.

I would note, even short of that, we have see some good reporting and some good academic research on the subject which I think raises concerns. But given the limited data available, it is not yet conclusive.

Mr. LANGEVIN. OK. Mr. Stamos, Mr. Buchanan, I just want to also follow up on my earlier line of questioning regarding focus on the users and scale. So there are billions of moderation decisions made each year regarding hundreds of billions of posts. So in contrast, there are only around 4 billion internet users. So scale is certainly a big factor.

Beyond the focus on removing fake accounts, what else can social media companies do to stop repeat offenders? In terms of improving the ecosystem, what steps should major platforms take to stop or reduce linking to less responsible platforms?

Mr. STAMOS. So the recidivism issue is a huge deal of people who have been kicked off the platforms and come back. I think the continued application of deeper and deeper machine learning algorithms to look for those people is going to be critical.

You brought up something that I have been big on, is I think something the companies could do right now is they could recognize that the white supremacist terrorist problem is different structurally than the Islamic terrorism problem in that we have a set of 4 or 5 websites that are intentionally being the hosts for this content. The companies could privately decide to ban links to those hosts.

Those hosts will still exist. People can go to them. But we shouldn't be allowing the 8chans, the Gabs, the white supremacist sites of the world to utilize the amplification effects on a Facebook or YouTube to then bring people over. I think that is a decision that those companies could make either collectively or on an individual basis to ban those sites from linking out.

Mr. BUCHANAN. If I might, I think there are two points here that are worthy of mention. The first is I think this is a very fair question to ask social media companies, which is to say, as your platform scales to billions of users, if you agree that machine learning techniques are insufficient to moderation, what is your plan to manage this, given the negative external social costs? That is the broader point.

The more narrow point, I think, is as Mr. Stamos said, one function of this plan that I think is comparatively underexplored are ways to moderate and reduce the amplification provided by big tech companies. Whether that is banning links, whether that is determining how much something shows up in the recommendation system or in the news feed, there seems to be substantial room to improve there, some options available that I think are worth exploring.

Mr. LANGEVIN. OK.

Thank you, Mr. Chairman. I will yield back.

Mr. ROSE. Thank you, sir.

To close it out, I would like to recognize Ms. Jackson Lee from Texas.

Ms. JACKSON LEE. Chairman, thank you very much.

The focus on elections and both the impact of the major tech companies certainly came to a head in 2016, although we realize that it has certainly been part of our world for a longer period of time than that. Questions of election security, I think, are crucial for every American.

I would hope as we move into the 2020 season of elections, that we can be clearly focused in on the concept of election security. The tech companies are very much intimately involved in that. Frankly, the terminology of election security should really be the preservation of the democracy and the values of this Nation; that whoever votes and whoever they vote for are voting on informed information that is truthful information. I think that is the very premise, whether we were having poster boards, calling people on the telephone of ancient days or now today.

So I want to ask about a growing concern, particularly going into the 2020 elections, is the emergence of deep fakes. Those that were partially used or maybe dominantly used in 2016, are artificially generated videos that appear to be real. Deep fakes may be disastrous for the 2020 election. I, frankly, believe that we should draw together to find the best solution for that, because truth does make for a better democracy.

So I am asking each of you if you can take note of these three elements that I would like to refer you to. That is, what is the current state of deep fake technology? Should we be worried? Why or why not? How will AI technologies contribute to misinformation going into the next election?

Mr. Stamos, let me give you one question first, and the three of you ask the other questions.

I understand that you work to make security a more representative and diverse field, which we appreciate. Would you mind explaining why this would be positive for the intersection of work dealing with AI and counterterrorism?

Mr. STAMOS. Absolutely, Congresswoman.

Ms. JACKSON LEE. Then the other gentlemen will, along with you, Mr. Stamos, start with Mr. Buchanan, will answer that other deep fake question.

Thank you.

Mr. STAMOS. Yes, Congresswoman. I completely agree. I think security and safely overall is about the adversarial use of technology to cause harm. The harms different people face from the internet are very dependent on their background, where they are from, where they have grown up, the people who are around them. Unfortunately, the teams that work on this mostly look like me. They are mostly white suburban from—computer science-educated. The lack of diversity in security and safety teams is a huge issue with us not predicting the kinds of things that are actually going to affect people in the field. So that is something that I worked on in Facebook and we are working on a lot at Stanford.

As to the deep fakes question, so the current state of deep fakes—I think everybody has seen what the deep fakes look like.

We are currently at a point where detecting deep fakes afterwards when you are specifically looking is still doable. But this is an arm's race. The way deep fakes are generated is a thing called the generative adversarial network, which is actually you basically get two AI systems to fight one another to try to trick one another. So kind-of by definition this technology is moving in a direction where it is becoming harder and harder to technically detect whether something is a deep fake or not.

Should we be worried? My personal belief is that the use of deep fakes to harass and abuse individual people is actually a much larger issue right now than in the political world, because when a deep fake or a cheap fake like the Pelosi video comes out, it is very easy to have counter programming saying, look, this is fake. We can prove it is fake, and you can do so.

The use of deep fakes to create fake nonconsensual intimate imagery, which is the term for revenge porn that we use, is actually really horrible. It has real bad impact, mostly on women, young women who are targeted with it. There is nobody helping them out with that. Actually, if I was still going to—if I was still at the companies, I would put—while there is more political view of, like, politicians being hurt by deep fakes, honestly, you folks can kind-of take care of yourselves, and the media is taking care of that. I would put more focus on individuals who are being hurt through the NCII problem.

How will AI lead to election stuff? I am less worried about deep fakes. I am more worried about AI getting better at creating personas that seem human. If you look at the Russian internet research agency work that we did, we could tell that most of that content was created by non-English speaking people. If we get to the point of where computers can do that, then it means our adversaries are much less constrained by their ability to put a bunch of people into a building in St. Petersburg. Then that makes it possible for organizations that are much smaller than the Russian Federation to do that kind of work.

Ms. JACKSON LEE. Mr. Buchanan.

Thank you.

Mr. BUCHANAN. I think we agree on a lot of points here. First of all, in terms of the deep fake technology, to a human, it is virtually indistinguishable at this point. Machines can capture some of the nuance and do some detection. But as Mr. Stamos said, the technology here is, in effect, the result of an arm's race between a generator that generates the fake and the evaluator that determines whether or not it is real.

In terms of the near-term impact of deep fakes, I would agree with Mr. Stamos that there is tremendous already existing impact to individuals, particularly women, that doesn't get enough attention. More generally, its effect on election security. I think that we should look at it under the broader framework of misinformation and propaganda and foreign interference rather than just as a technical specimen.

If we would remove deep fake technology from the playing field, I would still be exceptionally concerned about the possibility of foreign interference in the 2020 election. This is one arrow in their quiver, but my suspicion is that they have many other arrows in

their quiver. We need to think more generally and more broadly about how one combats misinformation campaigns.

One thing that I think might be tangible in the near term for 2020 is better labeling of videos on a platform like Facebook when there is significant evidence that it is fake or doctored. We have seen some steps in that direction, but I think we could do a lot more.

Ms. JACKSON LEE. Mr. Sanchez.

Mr. SANCHEZ. So I think a lot of the same difficulties with applying antiterrorism cross apply here. There are difficulties of context. You will find cases where something that is initially a parity or something that is created for humor is then presented as authentic. So do you categorically remove it or take some other approach? There is a question of whether leaving something up with countermessaging is more effective or simply blocking and assuming that the lie travels faster than the truth.

I don't know if it is a categorically different problem from disinformation more generally, but I imagine the technology is getting to the point where we will very soon find out.

Ms. JACKSON LEE. Mr. Chairman, you have been enormously courteous. I thank you very much. Our work is cut out for us, and I look forward to working on these many issues.

Mr. ROSE. Absolutely. I think that there is—Ranking Member, correct me if I am wrong, there is an opportunity for some bipartisan work——

Mr. WALKER. 100 percent.

Mr. ROSE [continuing]. Around this issue, because this does represent a real National security threat. This problem is not going away.

So with that, I do thank the witnesses for their valuable testimony.

Ms. Jackson Lee, thank you for your kinds words as well, 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.

Pursuant to committee rule VII(D), any hearing record will be open for 10 days.

Without objection, the subcommittee stands adjourned.

[Whereupon, at 11:28 a.m., the subcommittee was adjourned.]