PREPARED WRITTEN TESTIMONY AND STATEMENT FOR THE RECORD

OF

Alexander Stamos
Director, Stanford Internet Observatory
Stanford University

BEFORE

U.S. House of Representatives
Committee on Homeland Security
Subcommittee on Intelligence and Counterterrorism

ON

"Artificial Intelligence and Counterterrorism: Possibilities and Limitations"

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 online 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 online 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 online 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, startup 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 five 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 1000 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% 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% of "Bullying and Harassment" actions were proactive, compared to 99.3% 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/

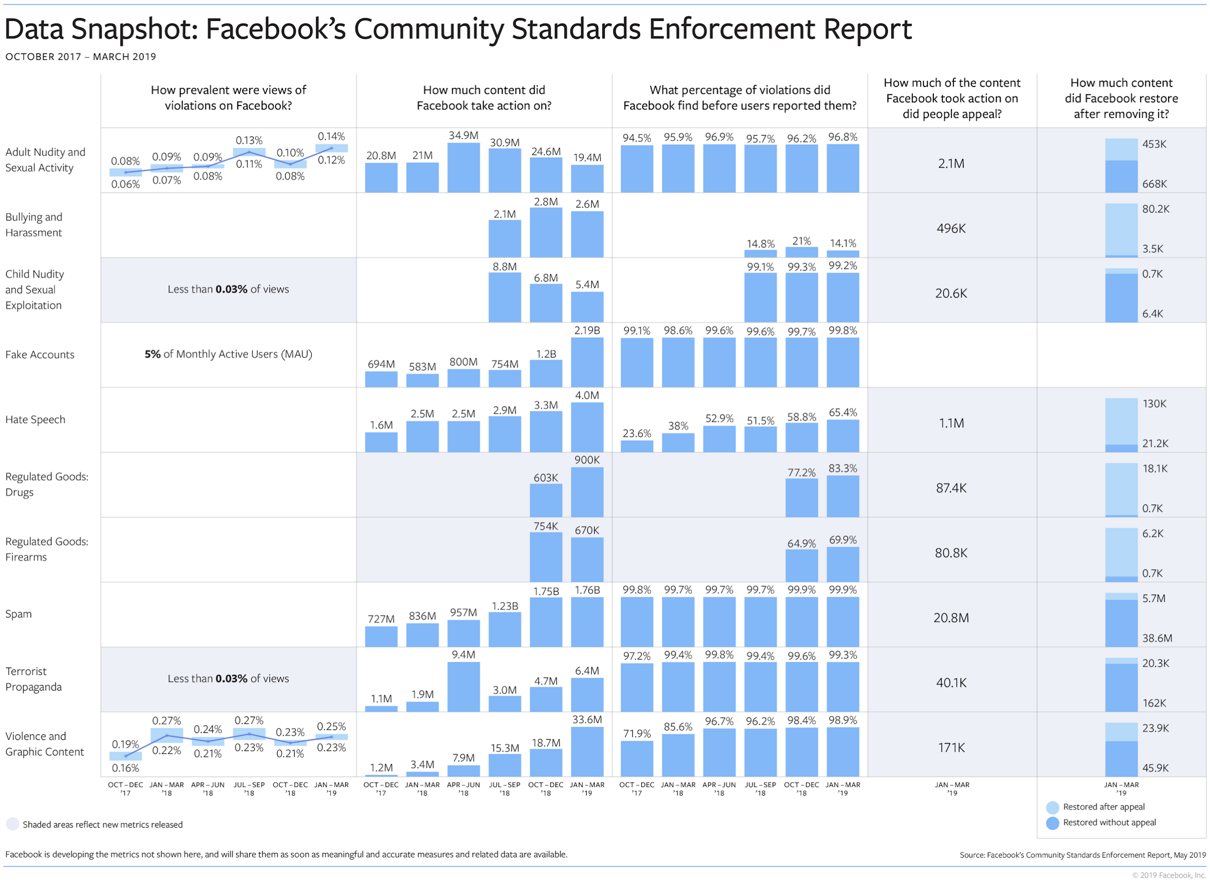


Figure 1: Summary data from Facebook's community standards enforcement report, published in May 2019[3].

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].

[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

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 online interactions. The attacker's manifesto and video can only be fully understood in the context of online 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 online activities. He recently published an analysis in the *Texas National Security Review* outlining seven 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 online 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.

[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

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 online ties to the shooter himself.

This collective action was made possible by the existence of radical, anonymous online 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].

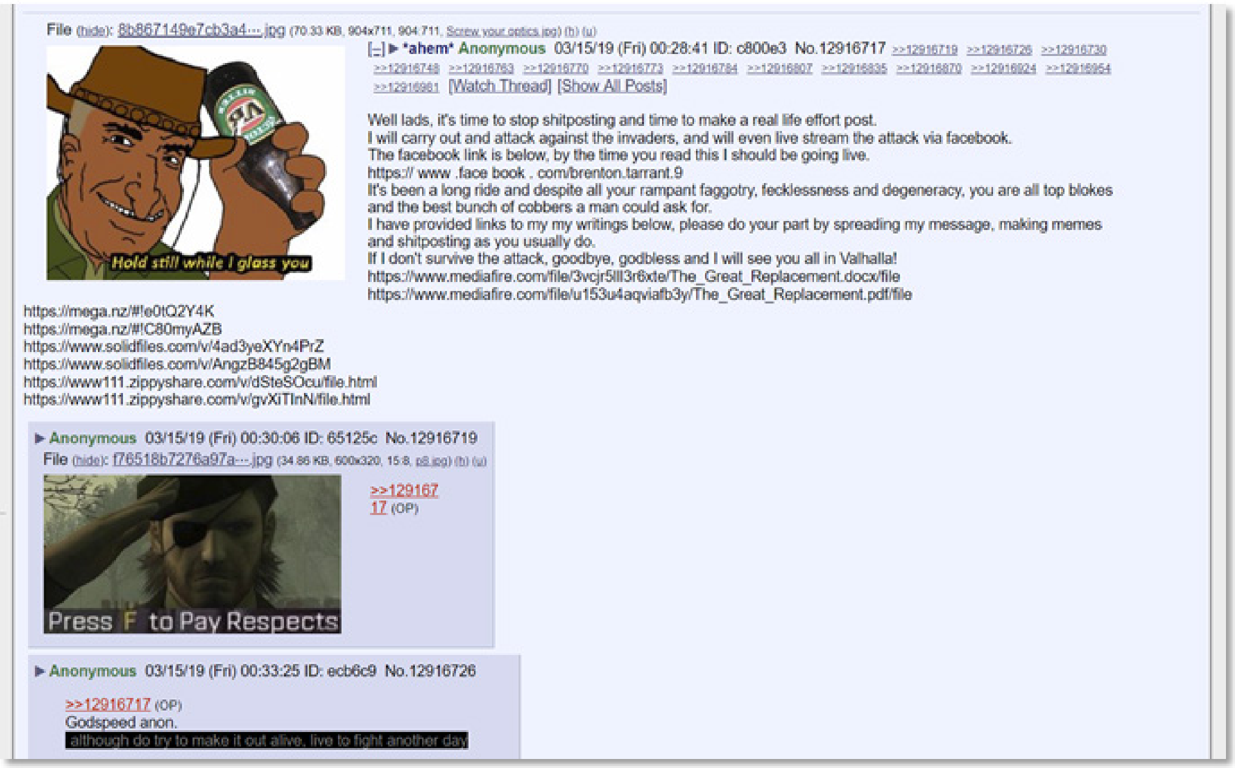
The Christchurch shooter posted links to his livestream and multiple copies of his manifesto on 8chan just minutes before beginning his attack. The 8chan thread lasted for hours afterward, filled with supportive comments from other members, including discussion of how to spread the message of the shooter. 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 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.

[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.



File (hide): 8b867149e7cb3a4⋯.jpg (70.33 KB, 904x711, 904:711, Screw your optics.jpg) (h) (u)

[–] ▶ *ahem* Anonymous 03/15/19 (Fri) 00:28:41 ID: c800e3 No.12916717 >>12916719 >>12916726 >>12916730 >>12916748 >>12916763 >>12916770 >>12916773 >>12916784 >>12916807 >>12916835 >>12916870 >>12916924 >>12916954 >>12916981 [Watch Thread] [Show All Posts]

Well lads, it's time to stop shitposting and time to make a real life effort post.
I will carry out and attack against the invaders, and will even live stream the attack via facebook.
The facebook link is below, by the time you read this I should be going live.
https:// www .face book . com/brenton.tarrant.9
It's been a long ride and despite all your rampant faggotry, fecklessness and degeneracy, you are all top blokes and the best bunch of cobbers a man could ask for.
I have provided links to my my writings below, please do your part by spreading my message, making memes and shitposting as you usually do.
If I don't survive the attack, goodbye, godbless and I will see you all in Valhalla!
https://www.mediafire.com/file/3vcjr5lll3r6xte/The_Great_Replacement.docx/file
https://www.mediafire.com/file/u153u4aqviafb3y/The_Great_Replacement.pdf/file
https://mega.nz/#!e0tQ2Y4K
https://mega.nz/#!C80myAZB
https://www.solidfiles.com/v/4ad3yeXYn4PrZ
https://www.solidfiles.com/v/AngzB845g2gBM
https://www111.zippyshare.com/v/dSteSOcu/file.html
https://www111.zippyshare.com/v/gvXiTInN/file.html

▶ Anonymous 03/15/19 (Fri) 00:30:06 ID: 65125c No.12916719
File (hide): f76518b7276a97a⋯.jpg (34.86 KB, 600x320, 15:8, p8.jpg) (h) (u)

>>129167
17 (OP)

▶ Anonymous 03/15/19 (Fri) 00:33:25 ID: ecb6c9 No.12916726

>>12916717 (OP)
Godspeed anon.
although do try to make it out alive, live to fight another day

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 pre-generated, 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 re-uploaded the video and manifesto millions of times[12].

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

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

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% of re-upload attempts were caught via automated means. The other 20% 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. Policymakers 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-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.

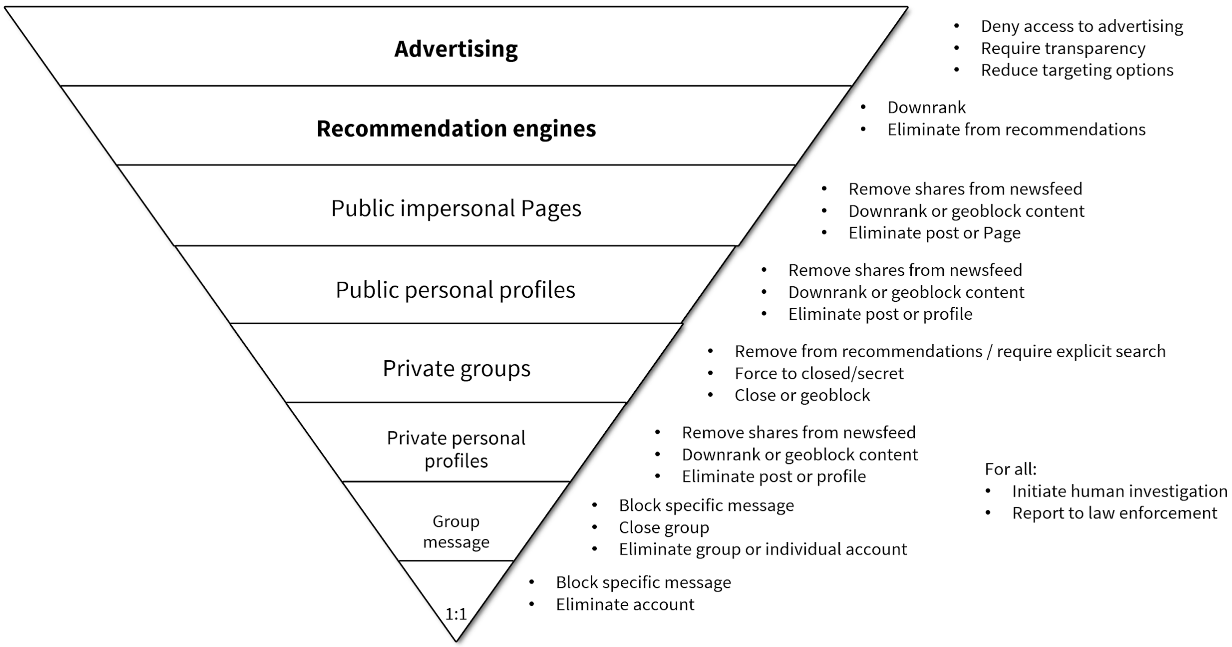

Advertising
Deny access to advertising
Require transparency
Reduce targeting options
Recommendation engines
Downrank
Eliminate from recommendations
Public impersonal Pages
Remove shares from newsfeed
Downrank or geoblock content
Eliminate post or Page
Public personal profiles
Remove shares from newsfeed
Downrank or geoblock content
Eliminate post or profile
Private groups
Remove from recommendations / require explicit search
Force to closed/secret
Close or geoblock
Private personal profiles
Remove shares from newsfeed
Downrank or geoblock content
Eliminate post or profile
Group message
Block specific message
Close group
Eliminate group or individual account
1:1
Block specific message
Eliminate account
For all:
Initiate human investigation
Report to law enforcement

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 US or EU privacy laws. The creation of a safe harbor for academic study should be part of the consideration of any US privacy legislation.

3) Establish better coordinating bodies for multiple abuse types. During the height of the struggle against the Islamic State's online 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.

[13] https://phys.org/news/2018-09-crucial-video-evidence-war-crimes.html

[14] https://www.lumendatabase.org/

[15] https://www.gifct.org/

Several weeks ago, my colleagues at Stanford and I released a report[16] with forty-five 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 cyberattack, election security, and combating fraud that could be handled by working groups inside such an organization without the need for separate, specialized organizations.

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 US 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 US 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.

[16] http://electionreport.stanford.edu

[17] https://www.fsisac.com/

[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/

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 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.

Complete Bio

Alex Stamos is a cybersecurity expert, business leader and entrepreneur working to improve the security and safety of the Internet through his teaching and research at Stanford University. Stamos is the Director of the Stanford Internet Observatory, a program dedicated to studying and mitigating the use of the internet to cause harm. He is also an Adjunct Professor at Stanford's Freeman-Spogli Institute, a William J. Perry Fellow at the Center for International Security and Cooperation, and a visiting scholar at the Hoover Institution.

Prior to joining Stanford, Alex served as the Chief Security Officer of Facebook. In this role, Stamos led a team of engineers, researchers, investigators and analysts charged with understanding and mitigating information security risks to the company and safety risks to the 2.5 billion people on Facebook, Instagram and WhatsApp. During his time at Facebook, he oversaw the company's investigation into manipulation of the 2016 US election and helped pioneer several successful protections against these new classes of abuse. As a senior executive, Alex represented Facebook and Silicon Valley to regulators, lawmakers and civil society on six continents, and has served as a bridge between the interests of the Internet policy community and the complicated reality of platforms operating at billion-user scale. In April 2017, he co-authored "Information Operations and Facebook", a highly cited examination of the influence campaign against the US election, which still stands as the most thorough description of the issue by a major technology company.

Before joining Facebook, Alex was the Chief Information Security Officer at Yahoo, rebuilding a storied security team while dealing with multiple assaults by nation-state actors. While at Yahoo, he led the company's response to the Snowden disclosures by implementing massive cryptographic improvements in his first months. He also represented the company in an open hearing of the US Senate's Permanent Subcommittee on Investigations.

In 2004, Alex co-founded iSEC Partners, an elite security consultancy known for groundbreaking work in secure software development, embedded and mobile security. As a trusted partner to the world's largest enterprises, Alex coordinated the response to the "Aurora" attacks by the People's Liberation Army at multiple Silicon Valley firms and led groundbreaking work securing the world's largest desktop and mobile platforms. During this time, he also served as an expert witness in several notable civil and criminal cases, such as the Google Street View incident and pro bono work for the defendants in Sony vs George Hotz and US vs Aaron Swartz. After the 2010 acquisition of iSEC Partners by NCC Group, Alex formed an experimental R&D division at the combined company, producing five patents.

A noted speaker and writer, he has appeared at the Munich Security Conference, NATO CyCon, Web Summit, DEF CON, CanSecWest and numerous other events. His 2017 keynote at Black Hat was noted for its call for a security industry more representative of the diverse people it serves and the actual risks they face. Throughout his career, Alex has worked toward making security a more representative field and has highlighted the work of diverse technologists as an organizer of the Trustworthy Technology Conference and OURSA.

Alex has been involved with securing elections globally as a member of the Kofi Annan Commission on Elections and Democracy in the Digital Age. He is a member of the Aspen Institute's Cyber Security Task Force, the Bay Area CSO Council and the Council on Foreign Relations. Alex also serves on the advisory board to NATO's Collective Cybersecurity Center of Excellence in Tallinn, Estonia.

Stamos worked under Prof. David Patterson while earning a BS in Electrical Engineering and Computer Science at the University of California, Berkeley. He lives in the Bay Area with his wife and three children.