S. Hrg. 117–514

SOCIAL MEDIA PLATFORMS AND THE AMPLIFICATION OF DOMESTIC EXTREMISM AND OTHER HARMFUL CONTENT

HEARING

BEFORE THE

COMMITTEE ON HOMELAND SECURITY AND GOVERNMENTAL AFFAIRS UNITED STATES SENATE

ONE HUNDRED SEVENTEENTH CONGRESS

FIRST SESSION

OCTOBER 28, 2021

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

Printed for the use of the
Committee on Homeland Security and Governmental Affairs

U.S. GOVERNMENT PUBLISHING OFFICE
47–981 PDF WASHINGTON : 2022

COMMITTEE ON HOMELAND SECURITY AND GOVERNMENTAL AFFAIRS

GARY C. PETERS, Michigan, *Chairman*

THOMAS R. CARPER, Delaware
MAGGIE HASSAN, New Hampshire
KYRSTEN SINEMA, Arizona
JACKY ROSEN, Nevada
ALEX PADILLA, California
JON OSSOFF, Georgia

ROB PORTMAN, Ohio
RON JOHNSON, Wisconsin
RAND PAUL, Kentucky
JAMES LANKFORD, Oklahoma
MITT ROMNEY, Utah
RICK SCOTT, Florida
JOSH HAWLEY, Missouri

DAVID M. WEINBERG, *Staff Director*
ZACHARY I. SCHRAM, *Chief Counsel*
CHRISTOPER J. MILKINS, *Director of Homeland Security*
MORAN BANAI, *Senior Professional Staff Member*
KELSEY N. SMITH, *Research Assistant*
PAMELA THIESSEN, *Minority Staff Director*
ANDREW DOCKHAM, *Minority Chief Counsel and Deputy Staff Director*
KIRSTEN D. MADISON, *Minority Director of Homeland Security*
MAGGIE FRANKEL, *Minority Senior Professional Staff Member*
SAM J. MULOPULOS, *Minority Professional Staff Member*
LAURA W. KILBRIDE, *Chief Clerk*
THOMAS J. SPINO, *Hearing Clerk*

CONTENTS

WITNESSES

THURSDAY, OCTOBER 28, 2021

ALPHABETICAL LIST OF WITNESSES

APPENDIX

SOCIAL MEDIA PLATFORMS AND THE AMPLIFICATION OF DOMESTIC EXTREMISM AND OTHER HARMFUL CONTENT

THURSDAY, OCTOBER 28, 2021

U.S. SENATE,
COMMITTEE ON HOMELAND SECURITY
AND GOVERNMENTAL AFFAIRS,
Washington, DC.

The Committee met, pursuant to notice, at 10:15 a.m., via Webex and in room 342, Dirksen Senate Office Building, Hon. Gary C. Peters, Chairman of the Committee, presiding.

Present: Senators Peters, Hassan, Sinema, Rosen, Padilla, Ossoff, Portman, Johnson, Lankford, Romney, Scott, and Hawley.

OPENING STATEMENT OF CHAIRMAN PETERS[1]

Chairman PETERS. The Committee will come to order. I would like to thank our witnesses for joining us here today.

Our Committee has held several hearings this year, examining the rise of domestic terrorism, and today's hearing is going to focus on the role of social media platforms and the role that they play in the amplification of domestic extremist content and how that content can translate into, unfortunately, real-world violence.

Yesterday marked three years since a white supremacist gunman opened fire in a Pittsburgh synagogue, killing 11 innocent people in the deadliest attack on the Jewish community in the United States. The attacker used the fringe social media platform Gab prior to the attack to connect with like-minded extremists and spread his own hateful and anti-Semitic views online. While bad, violent, and hateful ideology has long terrorized communities for many Americans, the shocking attack was the first glimpse at how quickly increased exposure to extremist content and users with similar beliefs can radicalize domestic terrorists and drive them to act on their violent intentions.

Less than a year after the Tree of Life attack, we saw a white nationalist open fire in an El Paso shopping center. This attacker was one of many who viewed video of the Christchurch mosque massacres that widely circulated on social media just a few months earlier, a video he reportedly cited as inspiration for his deadly attack in a 2,300-word racist manifesto that he also posted online.

On January 6, 2021, we saw a stark example of how individuals went beyond seeing and sharing extreme content across numerous social media platforms. They were spurred to action by what they

[1] The prepared statement of Senator Peters appear in the Appendix on page 39.

repeatedly saw online, and ultimately, a mob violently attacked Capitol Police and breached the Capitol Building.

In attack after attack, there are signs that social media platforms played a role in exposing people to increasingly extreme content and even amplifying dangerous content to even more users. Yet, there are still many unanswered questions about what role social media platforms play in amplifying extremist content. We need a better understanding of the algorithms that drive what users see on social media platforms, how companies target ads, and how these companies balance content moderation with generating revenue.

For the majority of social media users who want to connect with distant family and friends or stay up to date on their favorite topics, there is very little transparency about why they see the content, recommendations or ads that populate their feeds. While social media companies have promoted how they are providing more curated content for their users, we have seen how users can be shown increasingly polarizing content. In worst case scenarios, users are reportedly recommending more and more extreme content, nudging them down a dark and dangerous "rabbit hole."

Recent reporting and congressional testimony and revelations in the Facebook Papers have shed some light on business models that appear to have prioritized profits over safety and decisions that appear to disregard the platforms' effect on our homeland security. It is simply not enough for companies to pledge that they will get tougher on harmful content. Those pledges have gone largely unfulfilled for several years now. Americans deserve answers on how the platforms themselves are designed to funnel specific content to certain users and how that might distort users' views and shape their behavior, both online and offline.

As part of my efforts to investigate rising domestic terrorism, including the January 6th attack, I have requested information from major social media companies about their practices and policies to address extremist content so that we can better understand how they are working to tackle this serious threat. While we are continuing to work with companies to get answers and examine relevant data, I am looking forward to hearing from our experts today about how these platforms balance safety and business decisions and, specifically, how these decisions relate to rising domestic extremism.

Ranking Member Portman, you are welcome to start with your opening remarks.

OPENING STATEMENT OF SENATOR PORTMAN

Senator PORTMAN. Thank you, Mr. Chairman. I appreciate your holding this hearing. It is a very important topic. I look forward to hearing from our experts today and then, I think in a future hearing, hearing from some of the companies themselves.

The role that social media plays in directing content that can lead to online and offline harm has taken on new significance in the past several weeks as we have learned more about this from whistleblowers and others, and news has emerged about malfeasance by some of the largest internet platforms. This exploitation of social media, of course, is not new. In 2016, as Chair of the Per-

manent Subcommittee on Investigations (PSI) of this Committee, I held a hearing which examined the Islamic State of Iraq and Syria (ISISs), use of online platforms in furtherance of their violent goals. We learned from testimony that social media accelerates, in this case, ISIS's radicalization and recruitment efforts and can also speed up their mobilization to violence.

Today, foreign terrorist actors continue to try to weaponize social media to inspire radicalization and attacks against Americans and American interests. This use of social media for nefarious purposes is not limited to terrorists, of course. Drug traffickers, foreign adversaries, and a host of other threat actors exploit online platforms, particularly social media. China and Russia use these platforms to conduct influence campaigns targeting Americans, including being involved in our elections. Drug cartels and gangs use these platforms to traffic narcotics. Traffickers use these platforms to exploit children and other vulnerable people, and domestic violent extremists (DVE) across the ideological spectrum use social media to spread propaganda and recruit members to their cause. So it is a broad problem.

Social media platforms acknowledge that the threats exist, and they talk about what they are doing to prevent bad actors from exploiting their sites, including artificial intelligence (AI) to help moderate content, networks of cross-industry partnerships, and on-staff experts—some of you may have been in that position—all to prevent or remove dangerous content. However, there is still a persistent threat of harmful content despite what the platforms say they are doing.

These social media companies are businesses, so it is not surprising that Congress and others have trouble getting more information on their algorithms and how they operate, how they are designed to amplify content. That is proprietary information; I understand that. But Congress has heard from whistleblowers, like Frances Haugen recently, that Facebook has not addressed troubling aspects of its algorithms, which have promoted of variety of concerning and alarming posts.

This raises important questions about whether or not it is time to revisit the immunity provided by Section 230. In 2017, during my time as PSI Chair, I introduced legislation which would remove Section 230 immunity from platforms that knowingly facilitated sex trafficking. So we have dealt with this issue. That legislation, called the Stop Enabling Sex Traffickers Act (SESTA), was actually signed into law back in 2018. We have figured out how to deal successfully with Section 230 at least in this one narrow area but very important area.

I take advocates, researchers, and even platforms at their word when they call for regulation. A regulation can take many forms which puts a premium on having sound information and analysis as we consider legislation to solve these problems. In other words, we need to know more. We need to be able to look under the hood and figure out what the issues are to be able to regulate properly.

These experts are in front of us here today to help evaluate the extent of the problem and also discuss what we should be doing about it. So far, we have found out a lot about social media business models from third-party researchers and from whistleblowers.

The findings are largely based on anecdotal evidence. If they want to ensure that Congress pursues evidence-based policy solutions, I think it is incumbent upon the platforms to provide quality data.

I am already working with Senator Coons, who is Chair of the Senate Judiciary Subcommittee on Privacy, on legislation that would require the largest tech platforms to share data with legitimate researchers and scholars so that we can all work together on solutions to these problems that all of us have identified. Dr. Persily has been an important partner in this work. I look forward to his testimony today.

Importantly, as we look at these issues, we must take care that our efforts hold these platforms accountable, but it is done in a manner that balances First Amendment protections, which I understand Professor Franks is going to discuss in her testimony.

Mr. Chairman, again thanks for having this hearing, and I look forward to hearing from our witnesses, and I thank you for giving us a chance to have real experts in front of us.

Chairman PETERS. Thank you, Ranking Member Portman.

It is the practice of this Committee to swear in witnesses, so if each of you will please stand and raise your right hand, including those who are joining us by video.

Do you swear the testimony you will give before this Committee will be the truth, the whole truth, and nothing but the truth, so help you, God?

Ms. KORNBLUH. Yes.

Mr. SIFRY. Yes.

Ms. O'NEIL. Yes.

Mr. PERSILY. Yes.

Ms. FRANKS. Yes.

Chairman PETERS. You may be seated.

Our first witness today is the Honorable Karen Kornbluh, former Ambassador to the Organization for Economic Cooperation and Development (OECD) and who currently serves as a senior fellow at the German Marshall Fund of the United States, where she leads its Digital Innovation and Democracy Initiative to ensure technology supports democracies across the globe. Prior to her role with the German Marshall Fund (GMF), Ms. Kornbluh served in previous administrations as Chief of Staff (CoS) at the U.S. Treasury Department and of the Office of Legislative and Intergovernmental Affairs at the Federal Communication Commission (FCC), where she negotiated early internet policies.

Welcome to the Committee. You are now recognized for your 5-minute opening remarks.

TESTIMONY OF THE HONORABLE KAREN KORNBLUH,[1] DIRECTOR, DIGITAL INNOVATION AND DEMOCRACY INITIATIVE, AND SENIOR FELLOW, THE GERMAN MARSHALL FUND OF THE UNITED STATES

Ms. KORNBLUH. Thank you, Chairman Peters, Ranking Member Portman, and Committee Members for the opportunity to testify on this critical issue.

[1] The prepared statement of Ms. Kornbluh appears in the Appendix on page 41.

To underscore the points that you both made, the National Strategy for Countering Domestic Terror States clearly that the widespread availability of domestic terrorist recruitment material online is a national security threat. I would like to stress three points today. First, the design of the platform and its algorithms can promote radicalization. Second, this cannot be addressed by after-the-fact, whack-a-mole content moderation. Third, we need to change the platforms' incentives so that they fix these dangerous design elements.

As part of a test, Facebook researchers created an account for a fictional Carol Smith, a 41-year-old conservative mother from North Carolina. Within days, Carol was recommended pages related to QAnon, and within only three weeks the platform showed her an account associated with the militia group, Three Percenters. She did not ask to be shown this content, she had no idea why she got it, and she had no idea who was paying for it.

Facebook groups can be manipulative as well. Internal research found that a full 70 percent of Facebook political groups in the United States were rife with hate, bullying, harassment, and misinformation. Facebook's own algorithms recommend extremist groups to users. For instance, Facebook directs users who like certain militia pages toward other militia groups.

The platforms also provide tools that help extremists to organize. So-called "super inviters" can create invitation links to groups that can be shared on or off Facebook. The platform provides inviters recommendations of specific friends to invite, allowing them to recruit from other conspiracy and militia groups. As an example, Stop the Steal inviters sent these kinds of automated invitations to members of other groups, resulting in high membership overlap with Proud Boy and militia groups and fueling Stop the Steal group's meteoric growth rates.

This is a national security vulnerability. It was recently revealed that 140 million Americans were targeted by troll farms operating out of Eastern Europe.

Similar algorithm radicalization is evidence on other platforms. TikTok's algorithm, for instance, also promotes content from QAnon, the Patriot Party, Oath Keepers, and Three Percenters.

YouTube has 290 extremist channels. When researchers showed an interest in militant movements, YouTube suggested videos to them with titles like "Five Steps to Organizing a Successful Militia." The platform also recommended videos about weapons, ammunition, and tactical gear.

Extremists find it all too easy to work across platforms. Diehards can organize on less moderated platforms, like 4Chan or Telegram, and then may retail the fringe content on more mainstreams with just a few clicks.

Second, the whack-a-mole approach to catching content after it has gone viral cannot work. Facebook employees themselves admitted this problem. They said that the mechanics of the platform were behind the hateful speech and misinformation, but most of their ideas for changing these in order to limit algorithm radicalizations were rejected. The content moderation system cannot win against huge volumes of algorithmic recommendation, but that system is further undermined by exempting many users with

large footprints. No wonder that at Facebook the researchers said they catch only 0.6 percent of content that depicts violence or could incite serious violence.

Third, it is critical to change the platforms' incentives. While Congress works on more comprehensive legislation regarding privacy and antitrust, a digital code of conduct could help tackle algorithmic radicalization while protecting free expression. Congress or the Federal Trade Commission (FTC) could demand platforms commit to common-sense design changes and transparency, and the FCC would enforce the companies' commitments.

Platforms should implement the kinds of design changes that research has already shown would enable more consistent application of their own terms of service. For example, a circuit breaker, like those used by Wall Street to prevent market crises, could prevent the viral spread of sensitive content in topics areas with high harm potential while human reviewers have time to determine whether or not it violates platform policies.

Second, platforms should commit to transparent third-party audits, the equivalent of a black box flight data recorder, like the National Transportation Safety Board (NTSB) gets when a plane goes down or the data available to the Food and Drug Administration (FDA) or the Environmental Protection Agency (EPA), which should not need a whistleblower to access data.

Third, the FTC should enforce commitments to this code under its Section 5 consumer protection authority. Of course, Section 230 reform, as contemplated in a number of current bills, would also allow users to sue in cases of terrorism or serious harm. Or, they could require a code as a condition of limited liability, but this would require legislation.

Mr. Chairman, the whistleblowers documents are a look in the rearview mirror, but Web 3.0 is being built today. It is essential that we act now to set sensible rules of the road. I thank you for holding this hearing.

Chairman PETERS. Thank you, Ms. Kornbluh.

Our next witness is Dr. David Sifry. He is the Vice President of the Center for Technology and Society (CTS) at the Anti Defamation League (ADL). Mr. Sifry leads a team of innovative technologists, researchers, and policy experts developing proactive solutions and producing cutting-edge research to protect vulnerable populations. Additionally, Mr. Sifry is an advisor and a mentor for companies and was selected as a technology pioneer at the World Economic Forum (WEF). He joined the ADL after a career as a technology entrepreneur and as an executive at companies including Lyft and Reddit. Mr. Sifry is also an advisor and mentor for companies and was selected as a technology pioneer at the World Economic Forum.

Welcome, Mr. Sifry. You are recognized for your opening comments.

TESTIMONY OF DAVID L. SIFRY,[1] VICE PRESIDENT, CENTER FOR TECHNOLOGY AND SOCIETY, ANTI-DEFAMATION LEAGUE

Mr. SIFRY. Mr. Chairman, Ranking Member Portman, Members of the Committee, good morning. It is an honor to appear before you today to address the ways social media platforms amplify hate and foment domestic terrorism.

For over a century, ADL has been a leading voice in fighting hate in all forms, and we have been tracking online hate since the days of dialup. In 2017, ADL launched our Center for Technology and Society to respond to the threat of online hate. My team advocates for targets of online hate and harassment. We deeply engage with, and call out, tech platforms to hold them accountable for their actions and their deliberate inaction.

Before joining ADL, I spent my career as an entrepreneur and executive in the tech sector. A trained computer scientist, I founded six technology companies and served as an Executive at Lyft and Reddit. I have been on the inside, and I know firsthand how big tech companies work and how business incentives drive product, policy, and strategy.

These platforms maximize profits by providing hyper targeted ads to an audience that spends large parts of their life online. Core product mechanics like virality and recommendations are built around keeping you, your friends, and your family engaged. The problem is that misinformation, hate filled and polarizing content is highly engaging. So algorithms promote that content.

As ADL's own research has long suggested and Facebook leaks confirm, these platforms exploit people's proclivity to interact more with incendiary content, and tech companies do so with full knowledge of the harms that result. Ultimately, these companies neglect our safety and security because it is good for the bottom line.

With no accountability, no regulation, and no incentives beyond growth and increasing ad revenue, extremists find a haven to reach, recruit, and radicalize. Platform algorithms take advantage of these behaviors, especially to our attraction to controversial and extremist narratives. As a result, some users get trapped in a rabbit hole of toxic content, pushing them toward extremism. This has deadly consequences. ADL reports show that extremists on mainstream platforms push people into fringe communities that further normalize hate and violence. Extremist ecosystems inspire individuals to commit acts of domestic terrorism as we saw in Charlottesville, Poway, and El Paso.

Senators, three years ago yesterday, in what was the most lethal anti-Semitic attack in American history, 11 congregants were massacred at the Tree of Life Synagogue in Pittsburgh. Before he attacked, the terrorist posted his anti-Semitic manifesto, which then spread online and was expressly cited as inspiration by the Poway and El Paso shooters. How many lives will be lost before big tech puts people over profit?

The leaked Facebook documents revealed that company researchers flagged Facebook's key role in spreading conspiracy theories, inciting extremist violence, and contributing to the events of January 6th. Company executives were fully aware of the problem and

[1] The prepared statement of Mr. Sifry appears in the Appendix on page 46.

chose not to act. Self-regulation is clearly not working. These billion and trillion-dollar companies have the resources to improve systems, hire additional staff, and provide real transparency. Yet, they claim it is too burdensome. Without regulation and reform, they will continue to focus on generating record profits at the expense of our safety and the security of our republic.

The leaked Facebook documents, January 6th, and rising domestic terrorism all confirm what ADL has been stating for years; social media is a super spreader of the virus of online extremism. It is time for a whole-of-government, whole of industry, whole-of-society approach to fighting online hate. ADL built the PROTECT Plan to address the rise in domestic extremism and our REPAIR Plan to push back hate to the fringes of the digital world.

Congress must establish an independent resource center to track online extremists and make appropriate referrals, regulate platforms including through targeted Section 230 reform, ensure academic researchers access to data, and require regular and meaningful transparency records and independent third-party audits.

It is well past time to hold social media platforms accountable. Thank you for your leadership in working to bring an end to this cycle of hate.

Chairman PETERS. Thank you, Mr. Sifry.

Our next witness is Dr. Cathy O'Neil, the Chief Executive Officer (CEO) of O'Neil Risk Consulting and Algorithmic Auditing (ORCAA), an algorithmic auditing company that helps companies and organizations manage and audit algorithmic risk. As an independent data science consultant, Dr. O'Neil works for clients to audit the use of particular algorithms in context, identifying issues of fairness, bias, and discrimination, and recommending steps for remediation. Dr. O'Neil earned a Ph.D. in math from Harvard, was a post-doc at MIT Math Department and a professor at Barnard College. She has authored the books, Doing Data Science and Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.

Welcome, Dr. O'Neil. You may proceed with your opening comments.

TESTIMONY OF CATHY O'NEIL, PH.D.,[1] CHIEF EXECUTIVE OFFICER, O'NEIL RISK CONSULTING AND ALGORITHMIC AUDITING

Ms. O'NEIL. Thank you so much for having me today. I am in the lucky position of just being sort of a background expert here. I am going to try to explain three things: first of all, what is an algorithm; second of all, what is a recommendation algorithm; and third, what is a filtering algorithm because those are the two types of algorithms that you see the most on social media.

But I am just going to start with what is an algorithm. I am going to do the opposite of what most people who talk about AI and big data will do. I am not going to try to make it complicated; I am going to try to explain it in simple terms because it is quite simple. It is predicting the future, predicting what will be success-

[1] The prepared statement of Ms. O'Neil appears in the Appendix on page 75.

ful in the future based on patterns in the past, what was successful in the past.

It does not even have to be formal. It could be something you do in your head. Like for example, I cook dinner for my children every day. I look for patterns in the past. That is historical data. Well, historical data is just my memories. I know what my kid eats. He only eats carrots but not broccoli. That kid will eat raw carrots but not cooked carrots. I have a lot of information, and I can predict what will be a successful dinner.

But here is the thing that is really important. Besides the historical data I am talking about is the definition of success. I have to be very precise when I make an algorithm, and I have to say exactly what I mean by success. In this case of making dinner for my kids, I am going to define success as my kids eat vegetables.

The reason it is so important how you define success is because you actually optimize to success. So I am going to make meals that are likely to be successful. Time after time, I am going to learn from my past mistakes or successes, and I am going to make meals that will be successful in the future.

Now I just want to make the point that a different choice of success changes everything. If my son were in charge—he is a fan of Nutella and not so much of vegetables—then we would have very different meals. We would be optimizing to Nutella rather than optimizing to vegetables. So just imagine what kind of meal that looks like. It is completely different.

I want to make the point that algorithms depend a lot on patterns in the past, but they depend even more on what you define as success.

I will make a last point about algorithms just in general, which is like whoever is in power, whoever owns the code, typically is the one that defines success. I would say, to the points I have already heard, success for social media platforms is about money. They are always going to optimize to money which is, of course, ads, ad clicks.

Now we are going to talk about recommendation algorithms, which is how social media decides what content to show you or what groups to offer you, membership for.

I want you to think about this. I want you to think about your behavior on these platforms, it is a sort of scoring system, like you have scores in multiple dimensions.

Actually, let's start not with social media platforms, but let's start with like Netflix. Let's say you watch Diehard, the movie, twice a week, every week. Then you are going to be scored in multiple ways by the platform, and for example, you would be scored along the lines of: Do you like male characters in your movies or female characters in your movie? Do you like violent movies or nonviolent movies? Do you like suspenseful movies or nonsuspenseful movies?

If you watch Diehard twice a week, your sort of male character, violent, suspenseful movie scores will go up every time you do it. Every time you watch a movie that is like a chick flick or a romantic comedy, those scores will go down a little bit and romantic comedy scores will go up. So you should think about your profile from the perspective of a recommendation algorithm as just a series of

scores along these various dimensions that profile you, like what is your taste.

Now for the definition of success of those algorithms, the point is that they want you to stay on the platform as long as possible. For social media algorithms, the definition of success is, again, staying on the platform as long as possible. They will profile you and score you in all sorts of ways to figure out how to keep you on the platform.

I want to make the point that this is completely neutral so far. They would likely score me as very interested in crafting, and yarn in particular, and they would offer me yarn-type things, and every time I click on them my yarn profile score goes up. They would peg me more and more over time as somebody quite interested in yarn. In that sense, I would become an extremist with respect to yarn.

Every single person is sort of nudged and profiled with respect to their interests. I would even add that it is not just what they are interested in initially, but they can become more interested in certain things over time because of the content that is offered them. Similarly, if I watched Diehard enough on Netflix, I would be offered more and more movies like Diehard, and that would actually inform my profile and my tastes in the future.

I would spend time a little bit on filtering algorithms, but just suffice to say that whereas recommendation algorithms work quite well for the social media platforms and they make them very profitable because they do succeed in keeping people on the platforms, the opposite is true for filtering algorithms. The idea of getting rid of harmful content, they do not work well at all. They are very facile, sort of keyword search-based algorithms, and they are quite unsuccessful compared to recommendation algorithms.

I will stop there for now. Thank you.

Chairman PETERS. Thank you, Dr. O'Neil.

Our next witness is Dr. Nathaniel Persily, the Co-Director of Stanford Cyber Policy Center and the James B. McClatchy Professor of Law at Stanford Law School. Dr. Persily's scholarship and legal practice focuses on American election law. He is a co-author of the book, Law of Democracy. His current work has been honored as a Guggenheim Fellow, Andrew Carnegie Fellow, and a Fellow at the Center for Advanced Study in the Behavioral Sciences, examining the impact of changing technology on political communications, campaigns, and election administration.

Dr. Persily, you may proceed with your opening comments.

TESTIMONY OF NATHANIEL PERSILY, PH.D.,[1] CO-DIRECTOR, STANFORD CYBER POLICY CENTER, AND JAMES B. MCCLATCHY PROFESSOR OF LAW, STANFORD LAW SCHOOL

Mr. PERSILY. Thank you very much, Chairman Peters and Ranking Member Portman and Members of the Committee.

I am going to spend my time today talking about what we know, what we need to know, and then what to do about it. Before I do that, let me sort of admit to what we all know, which is why we are here, and that is that a courageous woman revealed thousands of pages of documents that previously no one had seen outside of

[1] The prepared statement of Mr. Persily appears in the Appendix on page 80.

the company. The revelations themselves, the content, were quite striking, but the fact that she did it was really momentous. It sort of brought front and center the fact that the internal researchers at these companies know an enormous amount about us and we know very little about then.

That is an equilibrium which is not sustainable. Right? We cannot live in a world where all of human experience is taking place on these platforms and the only people who are able to analyze the data are the people who are tied to the profit maximizing missions of the firms.

As Senator Portman mentioned earlier, I have been working both with his staff and others trying to figure out a way to open up these companies to outside research because they have lost their right to secrecy. All right? We are at a critical moment where we need to understand what is happening on these platforms.

Let me talk about three areas of critical importance when it comes to research on online harms. The first is how much harmful content, appears on these platforms. The second is the role of the algorithm. The third has to do with advertising.

First, how much harmful content is there on these platforms? The answer is a lot. Right? If you listen or read the transparency reports, there are millions of examples, whether it is hate speech or disinformation or inciting content and the like.

But those numbers are really quite meaningless, so is the claim that, for example, Facebook takes down four billion accounts a year. I mean, that is interesting because it seems like a lot, but none of us really know what the denominator is. We do not know how much of the offending content is being shared and viewed by users of the platform and what the experience is both of the average user and certain subsets of users.

For most people it is quite sort of important to understand their experience online is not filled with hate speech; it is not filled with disinformation and the like. But that is also the wrong question to ask.

The question—and where I think research has progressed in the last few years—is to suggest that there is a sizable minority on these platforms who are experiencing and producing an enormous amount of this terrible content, whether we are talking about incitement, hate speech, disinformation and the like. Particularly because these are folks that it is going to be very hard to survey, it is going to be very hard to get them in sort of outside strategies for research, we need the internal data to figure out exactly what is going on.

Second, on the role of the algorithm. This is an area where the platforms and outside observers seem to have diametrically opposed views. Whether it is the Haugen revelations or others that—from Facebook who have made this point or even conventional wisdom, sort of the argument is made that they are maximizing for engagement, and salacious disinforming and hateful content generates the most clicks, and therefore, that is what is favored in the algorithm.

If you ask the companies, they say, no, that there are all kinds of measures that the firms are taking in order to take down this

offensive content, that they take down, as I said, millions and millions of examples of hate speech and the like.

I will say one other point that they make is that on encrypted platforms such as WhatsApp and the like that you see the same types of offending content, and if you go outside the United States, where those kinds of platforms are much more ubiquitous and they do not have algorithms, that you see, as much, if not more, of the hateful content.

Finally, let me talk about advertising. These firms are advertising monopolies, and we need to treat them as such. That is what makes them distinctive. We do not really know a whole lot about how advertising is affecting sort of this information ecosystem problem that is the subject of this hearing. We know, of course, about the meddling in the elections, incendiary content, whether it is from Iran, whether it is China, whether it is Russia and the like. When outside researchers try to study advertising, as a group of New York University (NYU) researchers tried to do, they were kicked off the platform because they were trying to scrape and to try to find out exactly how people were targeted and the like.

So now what to do about it? There are many areas of reform that I think are on the table. Karen Kornbluh mentioned some important ones. I want to focus on this question of researcher transparency. We cannot live in a world, as I said before, where the only people who understand what is happening on the platforms are the internal researchers to the firms. Whether it is immunizing outside researchers who want to scrape the platform as these NYU researchers do or to develop a secure pathway, maybe administered by the FTC, in order to vet researchers so that they can analyze privacy-protected data, that has to be our future. Right? We cannot live in this world where the platforms hide behind their right to secrecy. Only if we can get access to this data can we then regulate intelligently.

Thank you.

Chairman PETERS. Thank you.

Our final witness this morning is Dr. Mary Anne Franks, who is a professor of law and the Michael R. Klein Distinguished Scholar Chair at the University of Miami. Dr. Franks is also the President and Legislative and Tech Policy Director of the Cyber Civil Rights Initiative, a nonprofit organization dedicated to combating online abuse and discrimination. Her work is at the intersection of civil rights and technology. Dr. Franks authored the award-winning book, The Cult of the Constitution: Our Deadly Devotion to Guns and Free Speech and has been awarded a grant from the Knight Foundation to support research for a book titled Fearless Speech.

Welcome, Dr. Franks. You may proceed with your opening comments.

TESTIMONY OF MARY ANNE FRANKS, D.PHIL.,[1] PROFESSOR OF LAW AND MICHAEL R. KLEIN DISTINGUISHED SCHOLAR CHAIR, UNIVERSITY OF MIAMI

Ms. FRANKS. Thank you. On October 14, 2021, Facebook announced a new artificial intelligence project called Ego4D. The name derives from the project's focus on ego-centric, or first-person, perception. Among Facebook's plans for this data include equipping augmented reality glasses with the capacity to transcribe and recall recordings of what people say and do around the user.

Asked whether Facebook had implemented measures to address potential privacy and other abuses of these capabilities, a spokesperson replied that the company "expected that privacy safeguards would be introduced further down the line." As underscored by multiple internal documents recently released by whistleblower Frances Haugen, this approach is characteristic of Facebook, aggressively pushing new untested and potentially dangerous products on the public and worrying about the consequences later, if at all.

Documents shared with the FCC note the asymmetrical burden on employees to demonstrate the legitimacy and user value of harm mitigation tactics before implementation, a burden not required of new features or changes aimed at increasing engagement or profits. While it may no longer be an official motto, "move fast and break things" still seems to be Facebook's animating philosophy.

It is notable that Facebook chose to announce such a highly controversial new project just as the company faces a storm of criticism and scrutiny over documented evidence that it knowingly allowed violent extremism, dangerous misinformation, and harassment to flourish on its platforms. One might have expected Facebook would be more circumspect about drastically increasing the capacity of individuals to record people around them without consent in light of the revelation, for example, that it allowed nude images of an alleged rape victim to be viewed 56 million times simply because the man she accused of raping her was a famous soccer star.

Is it arrogance? Is it callousness? Or, is it merely confidence? Confidence that no matter what is revealed about Facebook's role in the disintegration of our shared reality or the dissolution of our democracy—not its acceleration of a conspiracy theories from QAnon to Stop the Steal, its amplification of deadly disinformation about Coronavirus Disease 2019 (COVID–19), its endangerment of teenage mental health, its preferential treatment of powerful elites, or its promotion of violently racist and sexist propaganda—that it will face no real consequences?

After all, that seems to be the lesson that not only Facebook but other dominant tech companies have learned from previous scandals. Media attention will be intense for a while. They may be called before Congress to answer some uncomfortable questions. They may face some fines. But the companies will reassure the public that their purpose was never to cause harm. They will promise to do better in the future.

---

[1] The prepared statement of Ms. Franks appear in the Appendix on page 104.

It should be clear by now that debates over tech companies' intentions are a distraction and an obstacle to real reform. Moral and legal responsibility is not limited only to those who intend to cause harm. We hold entities accountable also when they know their actions will cause harm or when they are reckless about foreseeable harms and even sometimes when they are negligent.

Facebook and other tech companies have known for years that a business model focused on what is euphemistically called "engagement" is ripe for exploitation and abuse. These companies have, at a minimum, consciously disregarded substantial and unjustified risks to Americans' privacy, equality, and safety.

These risks are not politically neutral. Contrary to oft repeated claims that social media is biased against conservatives, the algorithms of major social media sites disproportionately amplify right-wing content. Facebook allows right-wing news sites to skirt the company's fact-checking rules and changed its algorithm in 2017 to reduce the visibility of left-leaning news sites. The day after the 2020 election, 10 percent of all views of political content on Facebook in the United States were of posts that falsely claimed that the vote was fraudulent.

As one Facebook employee wrote, if a company takes a hands off stance for these problems, whether for technical or philosophical reasons, then the net result is that Facebook will be actively, if not necessarily consciously, promoting these types of activities.

According to recently released internal research, Twitter's algorithms also disproportionately amplify right-wing content. Research on YouTube's algorithms show that they create a far more robust filter bubble for right-wing content than left-wing content and that Fox is, by far, its most recommended information channel, influence that illustrates how the ecosystem of extremism and disinformation is driven by forces beyond social media.

Lopsided political amplification is all the more troubling given the disproportionate rate of right-wing offline violence. Since 2015, right-wing extremists have been involved in 267 plots or attacks and 91 fatalities.

To be clear, the object of concern here is not conservative content as such but content that encourages dehumanization, targets individuals for violence and harassment, traffics in dangerous disinformation, and promotes baseless conspiracy theories that undermine our democratic institutions. Social media, along with, in some cases, mainstream media, elected officials, and others with influential platforms amplify these anti-democratic forces.

Structural reform, including reform to Section 230, that limits its protections to speech protected by the First Amendment and denies immunities to intermediaries who exhibit deliberate indifference to unlawful content, is necessary to ensure that no industry and no individual is above the law when it comes to the reckless endangerment of democracy.

Thank you.

Chairman PETERS. Thank you, Dr. Franks.

Recent reports based on the leaked Facebook Papers indicate that Facebook was certainly aware that changes that they made to their news feed configurations spread dangerous content more rapidly. Each of our witnesses have mentioned this in one form or an-

other already. Yet, company leaders repeatedly argue that they are investing in trust, safety, and civic integrity efforts. I am certainly struck, as I think you are, by this fundamental conflict between efforts to take down hateful and violent content and company algorithms that seem to amplify extremist views simultaneously.

Mr. Sifry, can you speak to this apparent conflict you have mentioned in your opening comments but a little more in depth if you would, please? How successful can these social media companies be at content moderation if they continue to design their own platforms to amplify extreme content?

Mr. SIFRY. Mr. Chairman, thank you so much for that excellent question. I think it really cuts to the heart of the conversation here, that at its core, what we are talking about is the incentive systems that drive this business model, a business model that is based around getting you engaged with, unfortunately, the natural human biases that we have toward engaging with controversial, with polarizing, with content that makes us afraid, with incendiary content.

What happens? All of those indicators that were talked about earlier, right, likes, shares, retweets, you name it, that those indicators go up. And so these platforms, Senator, are working as designed. What ends up happening is that we end up spending more time on those platforms and they end up tracking us more and then they get to send us more hyper targeted ads.

This will not change until there is a clear shift in the incentive systems that they use to be able to do this business, and that is where Congress must act. In creating systems that actually bring about a change in their incentive systems, that is how we end up getting them to behave rationally in this sense, right, and move toward those different incentive systems.

Chairman PETERS. Professor Kornbluh, just a follow-up basically on that question, in your opinion, can investments in trust and safety ever overcome a business model that prioritizes advertising revenues based on extremist content and the desire to keep people on a platform as long as possible?

Ms. KORNBLUH. I think that goes right to the heart of the question. The algorithm is this machine that is pushing misinformation or extremist content into a user's feed. It is recommending these extremist groups. It is targeting small groups of people with ads designed to agitate them. It is the mechanics of the system that are working.

Then these poor human content moderators, or even outmatched AI systems as Dr. O'Neil was talking about, do not stand a chance. So their own data shows that less than 1 percent of content that depicts violence or could incite violence is caught, and this is content that violates their own terms of service.

This is true; there are new revelations today about even something that was a high priority from Mark Zuckerberg, which is catching COVID misinformation. Apparently, they are unmatched at the content moderation even there. An international vaccine expert that I talked to told me he was so frustrated because he felt that the disinformation purveyors were working with the motor of social media and that the public health officials were fighting against the engine of social media.

I always think of that "I Love Lucy" skit when she and Ethel are on the candy conveyor belt and they are desperately trying to go as fast as the conveyor belt. These content moderators just do not have a chance.

Chairman PETERS. I want you to continue on this line of questioning, Ambassador. Earlier this year, a report found that Facebook was posting ads for body armor, gun holsters, and other military equipment right next to content promoting the 2020 election misinformation and news about the attack on the Capitol on January 6, so getting this content and ads for military equipment. What do we know about how Facebook and the platform targets their advertising and this kind of link? Can you talk to me a little bit more about what you are finding?

Ms. KORNBLUH. Yes. There is so much here, and it is overwhelming. It is hard to get our heads around this advertisement of these kinds of weapons, but there are three ways in which ads in general drive extremism.

First, it is very different than what we think of in terms of cable and broadcast ads, where they are micro-targeted specifically to the people who would be most moved by them and other users who might object, who might say, "Oh, that is violent" or "Oh, that is wrong," they do not get to see them. The ad content is not reviewed by humans before it is placed as they are on broadcast and cable. Third, users cannot really find out who is paying for these ads. Even political ads can just list a dark money group instead of their true sponsor, which could be a foreign government. So in 2016, remember, there were ads paid for in rubles that got through because there is no human monitoring this and there is very little transparency.

Then as we have talked about, the algorithms are designed to maximize this ad revenue. The ads are driving the whole thing. They are keeping you online to be fed the ads, and the incendiary content keeps users on the platforms longer.

Then there is this third element that I think is really important to think about. The algorithms that are trying to keep you online to feed you these ads, they are creating a marketplace that values extremism. The more extremist the content the better the distribution will be, the cheaper the ads will be actually. It is a cheaper ad if it is more incendiary per unit because it gets wider spread distribution. So it is creating this marketplace.

Political parties apparently came to Facebook and said, your algorithm is driving us to put out more incendiary content because it is the only way we can get distributed. If we just put out our 10-point plan, it does not get distributed online. So the ads are really the heart of this matter.

Chairman PETERS. Thank you. The buzzer you heard is a vote. So we are in a series of votes. You will see Members running in and out, and that is what I will do. I will go and vote. I will recognize Ranking Member Portman for his questions, and Senator Hassan will chair the hearing in my absence.

Ranking Member Portman.

Senator PORTMAN. Thank you, Mr. Chairman, and thank you to the witnesses. This is a really complicated area, and I am glad that Dr. O'Neil gave us a little one-on-one on algorithms and how they

work. Ultimately, she came down to the conclusion that this is about money. This is about what works, kind of along the lines of what Ambassador Kornbluh just talked about, that you know, what works is what sells more advertising and these algorithms are at the core of that. In other words, they are determining what we want to hear as online participants and amplifying that.

I will say two things that I just want to try to stipulate here at the start. Some may not agree with me, but it seemed to me, Ambassador Kornbluh, in your comments in particular, you were focused on right-wing extremism as being the problem. It is not. It is everything.

I hope we can agree to that because, again, the work that we did early in this Committee with regard to what ISIS was doing online, and in terms of recruiting and spreading violence, and in terms of what happens today even—I mean, there is, as you know, a lot of concern about these platforms not allowing what happened at the Wuhan lab to leak, to come out, or you know, concerns about what Hunter Biden is doing or not doing being blocked, or other things that lead me to believe that, whether it is Antifa on one side or whether it is White Supremacy on the other side, we need to look at this as a problem that is impacting ideologies across the spectrum.

I am just going to stipulate that because I want to get to some questions. Some of you may not agree with me on that, but I think that is really important, for us not to make this a partisan exercise.

Second is the First Amendment. This is impossible. How do you figure out what is speech that is peaceful expressions of points of view that we should be encouraging and what is content that ought to be filtered in some way? And there is lots of examples of this.

Recently, parents at a school board meeting by, I guess it was, a National School Boards Association (NSBA) said these were domestic terrorists. They are not domestic terrorists. These are parents trying to—and I think they later apologized for saying that. But, parents expressing their strongly held views about their kids' education.

You have to be sure that we are not taking content which is people expressing political views that are peaceful and somehow filtering those out. Any thoughts on that as we start, for any of the panelists, either with us or virtually?

Ms. KORNBLUH. I just want to agree with you. The algorithms are not partisan. The algorithms are economically driven, as Dr. O'Neil said, and they are trying to keep us all online. I think it is really important to keep that in mind.

I think for the First Amendment, the freedom of speech, freedom of association, it is extremely important that the government not be in the business of deciding what is true and what is not true and that instead—that is why I think some of these revelations from the whistleblower are so important, because she focuses upstream of the content, at the mechanics of the platform, and how it is driving this content just to service itself and to service its ad revenues.

If we focus on those design elements, and we focus on transparency especially, that furthers First Amendment concerns. It furthers freedom of speech and freedom of association. Transparency

is such an important principle, if we let consumers know who is behind what they are seeing, what interest is it serving, that——

Senator PORTMAN. That leads me to my core question today really, again, just stipulating this is a hard area. We have talked about a couple of those issues.

But it seems to me that, as I said earlier about getting under the hood and looking into what these design elements are, that you talk about the transparency issue. What you talk about is really important because—again, it is proprietary information. I understand that. These are private companies.

Again, this is not an easy issue for government to be involved with, but everybody is talking about regulation right now. I mean, everybody. Facebook is talking about it. Google is talking about it. Twitter is talking about it. We are talking about it. Everybody has a different view what that regulation might be, but shouldn't it be based on better data because we really do not know what we are trying to regulate, if there is a lack of transparency, as to what that design is or how these algorithms are derived.

So you know, we talked a little again in your testimony about this. Dr. Persily, in particular, you talked about your thoughts on how to give access to impartial researchers to be able, I assume, to publish about what is actually behind all this. What is the content-directing mechanism and how does it work. I am intrigued by that. I do not know if that is the answer. You mentioned that the FTC could play a role in this.

But can you amplify that a little bit and talk about what you think could lead us to more transparency and better understanding?

Mr. PERSILY. Thank you for that question. The model that I put out there is that the FTC, working with the National Science Foundation (NSF), would vet researchers who would not have the permission of the company, but the company would have to basically develop research portals for outside, independent research to study all of these societal problems that we are saying are caused by social media. The key features of this are simply that the firms have no choice in who has access to the data and we have some way of vetting the researchers to prevent, say, another Cambridge Analytica and the like. We need to have some process in place so that someone other than those who tied to the profit-maximizing mission of the firm get access to this data.

As you mentioned, it is proprietary data, but it is data about us. Right? It is essentially data about everything that is happening in American politics and, frankly, around the world. We need to figure out some way for the firms to be opened up so that outsiders can see it.

My view is that they should not turn the data over to a government agency, that that would pose real privacy and surveillance problems. We want to make sure that there is a vetted, independent third party that is able to do this kind of research, on all of these questions that have come up today, so that we can get to the answers to some of the questions that you asked earlier about the propensity on the left and the right to engage in hate speech or engage in violent content and the like, as well as potential bias

in the content moderation platforms and the like. Only if we have access to the data can we really answer those questions.

Senator PORTMAN. Otherwise, very hard to come up with regulation, which is what, again, everybody is talking about. I know there are different views on what that means, but it seems to me that there should be a consensus that if we are going to try to regulate this we need to have better information as to what the actual design and what the intentions are and the impact is. So this hearing is helpful, I think, in that regard.

I am at the end of my time. Hopefully, we will come back for a second round. I have so many other questions for this team. But again, I thank you for your expertise today.

Senator HASSAN [presiding]. Thank you, Senator Portman.

OPENING STATEMENT OF SENATOR HASSAN

it is now my turn for a round of questions. I want to thank Senators Peters and Portman for holding this hearing and to all of our witnesses, both in the room and virtually. I really appreciate your testimony. This is an excellent panel, and I appreciate your work very much.

I want to start with a question to Dr. O'Neil and Dr. Persily. I am concerned that extremist groups, including ISIS, continue to develop and refine online radicalization techniques that make it easier and quicker to radicalize people. At this Committee's annual threats hearing in September, leaders from the Department of Homeland Security (DHS), the Federal Bureau of Investigation (FBI), and the intelligence community (IC) expressed similar concerns. Extremists take advantage of the speed of social media platforms but also their algorithms, which often facilitate the quick spread of extreme content.

Dr. O'Neil and Dr. Persily, in your view, what are the weaknesses in the algorithms used by large social media platforms that can make it possible for anyone, including violent extremist groups, to expand and capture their audience? What steps can social media platforms take to curtail extremist efforts? Why aren't these companies taking these actions already? I will start with you, Dr. O'Neil.

Ms. O'NEIL. Thank you for the question, yes, and it is an important one. I want you to think of the filters, these things that are trying to catch content like that is hateful or otherwise not allowed, as sort of nets that fishermen use in the ocean. They sort of pull the net, and they see what they have got. They have some fish there, and they count the fish. They say, oh, we got a lot of fish. What they are not counting, of course, is the fish that got through the holes of the net.

What I am talking about are the people who are paid, actually, to put hateful content on Facebook and elude the nets. The thing about is that they can tell when they have been caught, and then they will double, redouble their efforts to change their content somewhat so that it gets through the net.

It is kind of like if you think about, the early spam on Viagra that would filter into your e-mail, and then you know, the spam filters got rid of the things that said Viagra. Then they started saying Viagra, but it was spelled—instead of with an "I," it was spelled

with a "1." Those got through for a while until they did not get through.

Spam filters work pretty well to remove Viagra ads in part because Viagra is the same word over many years. But in the case of social media, the stuff that they are putting on social media changes very quickly, and like the spam filters essentially cannot keep up.

I think you should think of it as an arms race. It is the filters on the one hand owned by the social media, and of course, the people that get sent the high scored, like high risk content——

Senator HASSAN. Right.

Ms. O'NEIL. Then they have to decide whether it is in fact against policy, and then on the other hand, all the propagandists who are actively trying to evade the filters. The simple truth is that the propagandists are winning that war.

To the extent that social media companies can combat it more, it would require much more expensive work, and they simply do not want to do it. So their policy has been, we are going to count how many fish we got, we are not going to count how many fish we did not get, and we are going to hope that it sounds good enough for you guys to stop asking questions.

Senator HASSAN. Thank you.

Dr. Persily.

Mr. PERSILY. So this is one of those areas where I wish I had the answer, and in order to get the answer we do need to have access to the data. I will say that, having talked to many of the integrity teams at these companies, I mean, these are pretty sophisticated operations——

Senator HASSAN. Right.

Mr. PERSILY [continuing]. That they have stood up in the last five years, and there is more that they could do.

As bad as things may be in the United States, by the way, they are a lot worse around the world, and that is something that I think naturally we are focusing on problems unique to the United States. But if they do not have the competencies in the local languages around the world, especially if you are dealing with terrorist content and the like, then they are going to be hindered in their ability to really attack these problems.

Again, if you look at the work that we have done at Stanford, if you go to the Stanford Internet Observatory, we have been trying on the outside to flag this kind of content, the violent content, terrorist content, and foreign election meddling and the like. We have been trying to do on the outside what they do on the inside, but it is extremely difficult.

Senator HASSAN. OK. Thank you.

Dr. Franks, the proliferation of nonconsensual intimate imagery, sometimes called "revenge porn," is a pervasive problem on the Internet. There are a number of truly despicable sites dedicated to hosting that material, but often users also share these private images and videos on large social media platforms with absolutely devastating consequences for those whose images are shared without their consent.

Congress and the States have taken notice of the tremendous harms from these situations, and there is some work going on to

address the problem. But, what more should social media companies be doing to prevent this content from being hosted on their platforms and to more immediately remove it when found? What additional tools could these companies give to people to help ensure that their images are not shared on social media platforms?

Ms. FRANKS. Thank you. Social media companies—I will say this about some of the more dominant companies—have been trying to tackle this issue, Facebook among them. There are teams at all of these various companies that are quite concerned about these issues and have worked with organizations like mine, the Cyber Civil Rights Initiative, to think about ways to impose their content policy restrictions and to encourage people not to participate in this kind of abuse and to try to empower victims to be able to remove the content.

That being said, those kinds of measures are essentially putting victims at the mercy of these companies. They may choose to make this priority. They may choose to impose some sort of policies. But there is not any necessary reason for them to continue doing so, and tomorrow they could simply stop.

I think it really is important for the work that is being done by the State legislatures and by Congress. The SHIELD Act is included in the Violence Against Women Reauthorization Act of 2021. That would federally criminalize nonconsensual pornography. That would be a real incentive, I think, for these companies to take it seriously.

This is connected to what I have suggested about Section 230, and it also goes to the broader question of incentives, about transparency or about regulation, about any of these questions. So long as these companies enjoy essentially blanket immunity for these harms, there is no real incentive for them to do anything, and therefore, it would be very important for there to be changes in Section 230 to take away some of that preemptive immunity.

Senator HASSAN. Thank you very much. Again, I appreciate the testimony of all of you.

I am going to recognize Senator Johnson now for his round of questions.

OPENING STATEMENT OF SENATOR JOHNSON

Senator JOHNSON. Thank you, Madam Chair. First of all, let me say anybody who has watched the documentaries, The Social Dilemma and The Creepy Line, I think has to understand that this can do great harm, particularly to children. I think as a result I think our first line of defense is with parents. I think that they need to really do everything they can to make sure that their children are not affected by this in a negative fashion.

Professor Franks, you made the comment that it seems like from your standpoint most of the bad content leans toward conservative and that type of misinformation. I would just ask you, what is your assessment in terms of the leadership, the people who work in these companies? How do they lean politically in your assessment?

Ms. FRANKS. I am not sure that it is relevant. I think the——

Senator JOHNSON. Pardon?

Ms. FRANKS. I am not sure that that is relevant. The particular individual leaning——

Senator JOHNSON. OK. I will just determine what is a relevant question or not, but let me just ask you again, what political affiliation do you think most people in leadership positions and that work in whether it is Facebook or Google or Twitter? Are they right-wing extremists? Are they conservative? Are they liberal? Are they radical leftists? I mean, what part of the political spectrum do you think they fall on?

Ms. FRANKS. Perhaps you could explain to me the relevance, and then I could answer the question.

Senator JOHNSON. Can you speak a little bit louder?

Ms. FRANKS. Sure. Perhaps you could explain to me the relevance of political affiliations of individual employees.

Senator JOHNSON. I was saying, just listening to the testimony here, it seems like the big concern here is about right-wing extremism, which I completely condemn, or right-wing misinformation. So again, I would just argue when you take a look at what Mark Zuckerberg has done through his Center for Tech and Civic Life, a couple hundred million dollars spent, pretty well took over the election system—I think in violation of Wisconsin law—of Green Bay and three or four other cities, it does not seem to me that the impact or the intent of their manipulation of data would tend to favor conservative groups or conservative thought. It seems to make more sense that they would tend more to push a liberal ideology.

Ms. FRANKS. As other experts have testified, the main issue for most of these companies tends to be profit, and profit is usually going to be built around engagement. Engagement is usually going to be built around outreach, misinformation, half-truths, things that provoke people into thinking that they are under attack, that they are being victimized.

Senator JOHNSON. OK. We heard that in testimony. Let me ask you, have you heard of the work of Dr. Robert Epstein?

Ms. FRANKS. I am not sure.

Senator JOHNSON. So I saw one of you shake your head. Is that Mr. Persily?

Mr. PERSILY. Yes. I am familiar with it.

Senator JOHNSON. You are familiar with it?

Mr. PERSILY. Yes.

Senator JOHNSON. Can you summarize it, or should I summarize it for you?

Mr. PERSILY. Whichever you prefer, but, yes, that he makes the argument that the platforms favor the Democrats in his——

Senator JOHNSON. He makes the argument that through manipulation of search, Google, as it ramps up toward elections, starts manipulating the search to push users of Google into the type of information that is going to tend to have you vote or decide to vote for a Democrat, delivering, according to him, millions of votes to Hillary Clinton, millions of votes to congressional Democrats in elections in 2018.

Mr. PERSILY. Yes, that is what he says.

Senator JOHNSON. OK. I think the point I am trying to make here is I think manipulation is potentially going to swing both ways. Who is going to be the arbiters of truth? You know?

Professor Franks, you said that the reckless endangerment of democracy. Now I would happen to think that if one of these platforms is utilizing their awesome power of manipulating a search to turn votes toward one political party versus the other, if one of these tech giants uses—spends hundreds of millions of dollars to turn out voters in Democrat-leaning areas and regions, that certainly impacts our democracy.

I will give you another example in terms of Facebook. I provided a forum for people who were vaccine injured to just tell their story. Following that, there were about 2,000 people involved in groups supporting each other. Some of these women have such severe vibrations from vaccine injuries that they committed suicide. So this Facebook group was a support group. It was literally helping people prevent suicides. Within a week, their group grew from 2,000 to 5,000 people, at which point Facebook dismantled the groups. These individuals who were counseling and helping prevent suicides lost contact with the people who were suicidal.

So who is going to regulate this? How is free speech different when it is on a platform versus when it is just spoken in the town hall?

Ms. FRANKS. I would suggest——

Senator JOHNSON. Who is going to be the unbiased arbiter of truth? I do not think it exists. I am certainly questioning the Section 230 liability protections when you have these platforms acting as publishers, which is what they are doing when they censor primarily conservative thought. I have been censored myself, repeatedly.

So again, I am just pushing back and challenging the fact that this is something that is fomenting right-wing conspiracies and highly advantageous to the conservative movement. I would say, if anything, it is more likely it is, from a political realm, advantaging left-wing ideology.

But again, I will come right back to we have a constitution that protects free speech. Who is going to regulate that fairly, in an unbiased fashion? It is just not possible. Along the way we are violating people's constitutional rights.

Anybody want to just take a stab at that one?

Ms. FRANKS. I would be happy to respond to that. Yes, we do have a First Amendment. We do have right to free speech.

But we also know, of course, that private companies are not obligated under the First Amendment to take all comers. They are allowed to make their own decisions about what is considered to be high quality or low quality content. They can make any number of decisions, and I think we would applaud them in many cases to make those decisions. As we were talking about just before, in terms of nonconsensual pornography, I, for one, am very happy that Facebook has made the decision to say that that is not welcome on their platform.

When it comes to the questions of conservative versus liberal bias, this is not a preconceived notion that I am suggesting here. This is not about intuitions or impressions although I know that those can go in many different directions. This is about what the data actually suggest. The data actually do indicate that right-wing content is more amplified on these social media platforms than left-

wing content and that right wing content is more disproportionately associated with real-world violence, not hurt feelings, not people being upset, but in fact actual violence, actual armed insurrections, actual notions of terrorism and anarchy.

Senator JOHNSON. Thank you, Mr. Chairman.

Chairman PETERS. [presiding]. Thank you, Senator Johnson.

Senator Ossoff, you are recognized for your questions.

OPENING STATEMENT OF SENATOR OSSOFF

Senator OSSOFF. Thank you, Mr. Chairman.

Ms. O'Neil, based upon your experience reviewing the algorithms underpinning many of these platforms and similar products, can you please connect the dots for the Committee, the link between the scale that these companies manage to achieve and the algorithms that they use to feed content to targeted users?

Ms. O'NEIL. Thank you for your question. I am going to make a confession. I have not audited these algorithms because these companies have not welcomed me or invited me to—inside their systems. They would not want me there.

For that matter, I do not think I would take that job for the very reason you are asking the question. As an algorithmic auditor, I want to consider who would be harmed, who are the stakeholders, and what kind of harm would come to those stakeholders or might come to those stakeholders. I do not go in assuming there is harm, but I do go in thinking about who are the stakeholders and what could harm them, and then making and developing tests and experiments to see the extent to which these harms are actually occurring.

For example, the kind of research we learned about from the whistleblower around teenage girls and suicidal ideation would be the stakeholder, teenage girls, and the harm would be suicidal ideation caused by their experiences on Instagram. That would be an example of the kind of stakeholder and harm that I would examine.

But if I were to actually be given the job of auditing the Facebook algorithm or the Instagram algorithm or any of the other algorithms, it would just be too large. There would be too many stakeholders. For example, it is very clear to me that if I had been given that job four years ago I never would have imagined the stakeholder that would be the Rohingya Muslims who were going to be there was going to be a genocide against them in Myanmar. That simply would not have occurred to me.

So it is just too big a job to do that, and it is because of scale, because it is international and because even within the United States it is too large to imagine who are the stakeholders and what kind of harm could befall them.

Having said all that, there are specific stakeholders that you can imagine right now that are interested—that are the focus of this particular Committee, that you could be saying, well, wait a second. Are these stakeholders having these harms that are actually illegal or a national security threat? That is a kind of algorithmic audit I would be happy to do. It would not be something that Facebook would invite me to do. You would have to somehow subpoena the data for me on my behalf, but that is something I could do.

To summarize, auditing for me is a stakeholder and a harm and an experiment to see the extent to which that harm is falling on that stakeholder. I would be happy to do that, but you would have to choose a few of them because there is just almost an infinite number to consider a priori.

Senator OSSOFF. OK, Ms. O'Neil. Thank you.

Dr. Franks, having observed this hearing, paying close attention to the broader discourse in politics, culture, society on these issues, what do you think is being missed by policymakers, and how is the nature of our debate perhaps overlooking key considerations or facts relevant to the policy discussion?

Ms. FRANKS. Thank you. I think a lot of what is being missed is—or I should say the focus is sometimes on purposes, intentions, listening to the companies say, we are working on this; we wish we had caught that. I really think it is long past time that we look beyond what the companies are saying they care about and what they are intending to do and simply look at what is happening, that the question of intentions just be something we leave in the past. This is why I suggest, for instance, that Section 230 is ripe for reform because it gives too much deference to the idea that this industry will be able to regulate itself.

I think the other important issue is to recognize that certain types of changes to Section 230 need not raise or settle all of the First Amendment concerns that have been brought up, that are, of course, legitimate to bring up and be concerned about, that modest reforms to Section 230, for instance, denying immunity when it comes to harm that is caused, that is foreseeable, and to which intermediaries have exhibited deliberate indifference.

All that would do would be to allow people to be able to sue these companies if they had a theory. It does not mean that they would be vindicated. It does not mean that some of that speech would not ultimately found to be First Amendment protected.

It means that the industry would not continue to have this kind of preferential treatment that rarely any other industry really has. They would be called to account. They would have to reveal documentation. They would give us some transparency about the inner workings of what these companies actually do. They would allow, at least in some cases, for people who have been harmed to prevail and to actually get some kind of compensation for their injuries.

Senator OSSOFF. Thank you, Dr. Franks. What information about the business practices of these firms that may not currently be public do you think would be of the greatest value to Congress as we weigh potential statutory revisions?

Ms. FRANKS. There are any number of things that I would particularly be interested in hearing what these companies are doing, but just to take a few examples, when companies implement policies against certain types of harms and say that they now have removal policies, let's say, with nonconsensual intimate imagery: What is the data in terms of what kinds of reports they are getting? What is the data in terms of whether they are taking those requests seriously? How quickly are they responding to those requests? How often are they aware that those types of material are flourishing on the platform and increasing engagement? How many

times are they willing to, nonetheless, hold to their principles and take it down as they have said that they would?

In essence, are they actually fulfilling the promises that they are making to the public?

Senator OSSOFF. Thank you, Dr. Franks. Thanks to the panel.

Mr. Chairman, I yield back.

Chairman PETERS. Thank you, Senator Ossoff.

Senator Rosen, you are now recognized for your questions.

OPENING STATEMENT OF SENATOR ROSEN

Senator ROSEN. Thank you, Chair Peters, of course, Ranking Member Portman, for holding this very timely hearing.

To our witnesses, I appreciate you being here and sharing your expertise with us.

I want to talk a little bit about algorithms because the tools of violent extremism, like conspiracy theories and disinformation, they frequently begin online as building blocks of hate. As we have seen time and time again, hateful online words can morph into deadly online—offline actions, excuse me, and then amplified once again online. This is a vicious cycle. This was the case three years ago yesterday at the Tree of Life Synagogue shooting, and it is true in far too many cases of hate.

As we have been discussing today, what often enables extremist groups and individuals to disseminate hate messaging are social media algorithms. Platforms generate algorithms that promote content, even if it is harmful, to keep people engaged. And engagement is what drives advertising revenue.

We have learned recently that social media platforms have known for some time—and I am going to quote here—"hate speech, divisive political speech, and misinformation," on their apps are having a negative impact on society and that—again I am going to quote—the core product mechanics, such as virality, which means how viral someone goes or some post goes. Those recommendation algorithms optimizing for engagement, they are a significant part of why these types of speech flourish.

So as a former computer programmer, I know that platforms have the capability to remove bigoted, hateful, and incendiary content that leads to violence, and I also know that they have an obligation to do so.

So to Mr. Sifry and then Ms. O'Neil, when a platform announces a new policy banning hate content, do you know how often do they or should they adjust their algorithms to reduce this content and, more importantly, to reduce and remove the recommendation algorithm for hate content so they do not continue to spread like we see that?

Mr. SIFRY. Senator, that is an excellent question. Thank you so much. The core issue, right, that these policies that they are creating—and many of the large tech companies have, on the face of it, admirable policies against hate, against incitement to violence, and against harassment. However, the issue is enforcement at scale, and what we have seen over and over again is that the platforms are falling down at being able to enforce at scale.

For example, for nine years, Facebook did not have an official policy on Holocaust denial. They allowed this content to stay up on

their site with no policy. Last year, after years of advocacy by civil society groups, including ADL, they finally changed their policy and said, OK, this terrible, nefarious content needs to be off the platform.

Yet, in January of this year, just 3 months later, when you would expect they had time to actually enforce said policy, we went and tested that policy. We were able to find groups with tens of thousands of members that were still advocating Holocaust denial, and we were able to find these kinds of nefarious content still being pushed to the tops of people's feeds.

So you are so right in the sense that not only the policies are, of course, important, but it is how do you enforce these policies at scale and how these algorithms, because of these engagement mechanisms, will then push them to the top of our news feeds right next to content from our friends and family.

Senator ROSEN. Thank you. Mr. Sifry, what do you think we can do to reduce or remove these algorithms? I am sorry, I meant Ms. O'Neil.

Ms. O'NEIL. Yes. Thank you.

Senator ROSEN. I was looking at Mr. Sifry but seeing you on the screen. I knew what I meant. Sorry about that.

Ms. O'NEIL. Thank you, Senator Rosen. It is a really good question and an important question. I personally, and probably you as well—when I heard Mark Zuckerberg say a few years ago that AI is going to solve this problem, I knew that was a lie, and he knew that was a lie. AI does not have a notion of truth. It does not understand the English language or never mind other languages. It simply looks for keywords. It is like a little bit of a gussied-up version of a keyword search.

So the point is that if we cannot decide what is true, right, AI can definitely not decide what is true. Its track record is so very weak. We have been hearing about less than 1 percent of certain types of violent content being caught by this particular type of AI. I hesitate to call it AI because that is giving it too much credit. It is just an algorithm that is a filter.

I also want to say that I do not have any confidence that this will work better in the future. The reason is that, whereas the propagandists are working with the recommendation engine, should get more attention because they want more attention—so you can say that it is like they are working with that algorithm—they are working directly against the filtering algorithm. They are trying to bypass it, and they are very good at that. The filter is just a very weak thing, and it will never work as far as I am concerned.

Senator ROSEN. I want you to go on. We think about enforcement at scale, some of the things you are talking about. What can we do here in Congress to really be sure that we can propose guardrails there for this misinformation, hate, things that promote violence in the real world and then amplify that violence and celebrate it after it happens? What can Congress do to help stop this?

Ms. O'NEIL. I think I am going to go back to the first person who spoke—I believe her name is Karen—that it is about incentives. Right? So the thing that I suggest should happen would be for you guys to have a specific definition of the kind of harm that you want them to keep an eye out for and that you have a way of making

sure it is happening, a monitor, and every single time their monitor fails they get charged money because it is always going to be about money.

So in other words, you have to make it more expensive for them to let it happen than the profit that they gain by letting it happen. You have to really counter the profit motive, and that is the only way you can do it.

Of course, it will be very expensive for them to stop these kinds of things. It will not work with AI. They will have to put humans on it, and they do not want to do that.

Senator ROSEN. Right. Thank you.

[Simultaneous discussion.]

I was going to say, Mr. Sifry, do you have something to add? Because I really want to work to stop this escalation and celebration and this vicious cycle. So, please.

Mr. SIFRY. Absolutely. What is so critical here is it starts with transparency. So No. 1, what are the actual policies that these platforms actually have? Facebook had a policy that they never reported on, called Cross Check, where over five million people were essentially absolved from any kind of machine review, and this included a virtual whitelist where celebrities, politicians, and others could say whatever they wanted without review.

Then for each of the policies they have, what are the enforcements that they have done? We expect these kinds of reports when companies go public and go to the FCC, and we expect it, and we penalize them when they lie. We should be doing the exact same things with these companies that are so vital to the future of our democracy.

Senator ROSEN. Thank you. I appreciate both.

I have some other questions for the record. Mr. Chairman, I will be submitting those. Thank you for your time today.

Chairman PETERS. Thank you, Senator Rosen.

Senator Lankford, you are recognized for your questions.

OPENING STATEMENT OF SENATOR LANKFORD

Senator LANKFORD. Mr. Chairman, thank you very much.

Thanks for all the testimony. This is an ongoing national conversation, obviously. We do not have an anticipation we are going to solve it all today in this hearing, but we do have an anticipation we are going to narrow down what are the key things that we are going to engage in and that we have to be able to find a way to be able to solve. So it is helpful to be able to have your dialog in the conversation.

Five years ago, I was with an executive from a Silicon Valley company, which I will leave out the name of the company, and as we were chit-chatting I made just some random comment about his social media page. he looked at me point-blank and said: Oh, I do not do social media at all, and my children do not do social media at all. I know better because I know what it is and what it is designed to do. I would never do that.

It was stunning to me to hear someone in the middle of Silicon Valley to say: Oh, I know how toxic this is. There is no way I am going to allow my children or my own family to be affected by this.

For some reason, that message has not gotten out to a lot of other folks, and for some reason there are folks that are in social media companies that have no personal challenge with being a part of that. The original sale of this seems to be we are going to swap pictures of our children and what we are eating for dinner, and it has moved to something very different.

So saying all that, I want to talk a little bit about ways that it could get better, practical aspects. Some of them were just mentioned online, in some of the conversation just now. What are some practical things that could be done to make this platform, this type of platform, better to be able to engage? Anyone who wants to jump in with a practical idea, you are welcome to.

Mr. SIFRY. Senator, if I may, so we have talked a lot about changing the incentive systems that then get these companies to change their algorithms. I think, No. 1, we have been looking at Section 230 reform in particular targeted ways. But this is a complex issue. We need to be careful about how we do so and how to, in fact, incentivize companies to maintain—they will keep the 230 protections if they actually reform and behave more responsibly.

Second, we have talked about transparency reporting, a number of us have, and I think that this is something that again has real First Amendment considerations here. This is something that we should be asking these companies to do tout de suite.

Third, ADL has been bringing forward the idea of a resource center, an independent resource center much like the National Center for Missing and Exploited Children (NCMEC), where we have for the exploitation of a child and sexual imagery. It is a public-private partnership to help to track what is going on in the world of extremism as well.

Most importantly as well, academic and other good faith, third-party researcher access to this data so that they can actually be looking at what is going on and report back on what is happening beyond what is just mandated.

Senator LANKFORD. OK. Other specific issues?

Ms. FRANKS. I do think that Section 230 reform is going to be one of the most important planks of this. Changing incentives for the industry is of the utmost priority. If companies are essentially given a blank check, told that they can engage as much as they want, as often as they want, and take profits from that, that they will face no repercussions for this, that they are immune sort of preemptively from suit, that is something that gives them a terrible incentive to do anything responsibly. So that would be one way.

Senator LANKFORD. Is this 230 reform, or is it 230 enforcement because the issue about not being an editor or publisher? Clearly, multiple of these entities are already trying to be able to alter content and to be able to engage in something that looks like a violation of 230 already. If there is a 230 issue, is it enforcement, or is it changing the way that it is written?

Ms. FRANKS. The 230 issue I think can be thought of in either way. There are ways to read 230 more in a limited fashion, but I do think that the primary problem with Section 230 currently mostly has to do with the (c)(1) provisions which provide immunity for things that the companies are leaving up.

Senator LANKFORD. Right.

Ms. FRANKS. I think if we were to open the door to that they would be subject to the same kinds of litigation that other industries are a part of, we would see more safety and better standards in addition to, I think, far more FTC interventions against what some of these companies are doing, recognizing that there is a fundamental misalignment of incentives here because these companies, their customers are not the users; their customers are advertisers.

Therefore, we have a big problem when it comes to how much those incentives are being thought of as advertiser issues as opposed to user issues——

Senator LANKFORD. Right.

Ms. FRANKS [continuing]. How often users believe somehow that they are getting a free service. There needs to be more oversight from the FTC about what that means for consumers and what it means for their consent.

Senator LANKFORD. With that, how do you balance out the issue of censorship? Because Focus on the Family and the Heritage Foundation and countless other conservative groups or faith based groups would tell you they put up content and immediately it gets pulled down and blocked or they get their account suspended. They are saying, "we are trying to keep, hateful content away, and typically, hateful content is conservative content". And so they are saying, "we are trying to abide by that". How do you balance that out?

Ms. FRANKS. I think there again we have to underscore the fact that these companies do have a First Amendment right of their own to exclude content and not associate with content that they find objectionable for any reason.

Senator LANKFORD. Right.

Ms. FRANKS. Therefore, that is the part of Section 230, the extent to which it gives them procedural rights as far as that enforcement goes, that I think we probably should leave alone.

Senator LANKFORD. We are, on the other side of this then, dealing with antitrust violations, where you have an issue where a company intentionally goes in and blocks out competition to be able to make sure there are no other voices than them in the marketplace and only their point of view gets out. That leaves the First Amendment issue, which I agree with, and moves into an antitrust issue that we still have to be able to resolve for several things.

Mr. Persily, do you have something to add to that?

Mr. PERSILY. Certainly. Thank you. Let me begin where you left off, which is so we were thinking about the antitrust problem in a very classic way, and we realize that these firms do not fit the normal model of antitrust. Let us talk about it in terms of competition law or the like. There are a lot of measures on interoperability and trying to break up their monopoly on data that I think Congress should explore.

Also, there are sort of non-speech-related reforms that I think will have an impact on these speech-related issues, right, in a beneficent way. So privacy legislation, right, is tied to advertising regulation because the amount of data that they hoover up from all of us that is what enables certain types of targeting, certain types of messaging which a lot of us here today were complaining about.

Then on transparency, everybody is in favor of transparency. What I want to emphasize—and Mr. Sifry mentioned it before—is researchers or just getting some third party in there to figure out what is happening, whether it is on the political bias, on content moderation or on the actual nature of the platforms, that this will change their behavior if they know that they are being watched. Right?

It is not just about providing a subsidy to outside researchers to figure out something for their publications. It is about making sure someone is in the room to figure out sort of what is going on and that the researchers are not the only ones in the firms who are tied to the profit-maximizing mission of the firms, who then have access to that information.

Senator LANKFORD. Yes. Mr. Chairman, it is interesting. I brought to Facebook several years ago now just an idea, to say, if you want to turn down the volume in some of these pages where there is people attacking other people on the page, allow the opportunity for the page owner, quote-unquote, to be able to just say, the comments just come to me, but they are not public. If you want to comment to me, you can comment to me. But then that discourages people from attacking each other back and forth.

Their response was, well, we like the interaction.

It is interesting. The preference for them was very clearly we like people hating on each other on this, even on pages where they know that is the dominant theme, because it helps with advertising dollars. It helps with the income side. These platforms are very aware of methods to be able to turn down volume, turn down hateful rhetoric, but they choose to be able to leave it up for advertising dollars.

I do think that is something we need to continue to be able to engage in, to be able to show how this could be better, that you can get back to a point of sharing pictures with loved ones and what you had for dinner, though I have no idea why people care what picture you had for dinner. But there is a way to be able to turn down volume, and we expect platforms to be able to step up and do that.

Mr. Chairman, thank you for holding this hearing.

Chairman PETERS. Thank you, Senator Lankford. Clearly, a lot of work ahead for us. I agree with you.

Senator Romney, you are recognized for your questions.

OPENING STATEMENT OF SENATOR ROMNEY

Senator ROMNEY. Thank you, Mr. Chairman. I, like all Americans, are very concerned about what we are seeing in social media, the impact it is having on our democracy, the disinformation that we are seeing. I have grandkids whose parents are wisely telling their kids not to get on social media. They are not giving them smartphones until they get a good deal older.

But frankly, as I have listened to the conversation this morning, it strikes me that this is a story we have heard before in a different context. The idea that there are companies that are trying to give people what they want to read, well, that is what newspapers do. There are companies that are trying to let people hear what they want to hear; that is what radio stations do. There are TV stations

that have something known as Nielsen, which tells them what people watch, and they found more salacious things, I presume, get more eyeballs.

I have heard on some of the nightly shows that they can look at which guest comes in, what that guest is wearing, and they can see their ratings going up based upon the gender of the guest and what they are wearing.

I mean, this is not something that is just particular to social media. It is something that is part of our entire media system. The idea that media companies, like social media companies, are trying to maximize their profitability, so is TV, radio, newspapers, magazines. They are trying to stay in business and maximize their revenue.

Disinformation. Have you ever read the National Enquirer or the Star? I mean, those are out there. They are at the newsstand. People can pick them up. There is all sorts of disinformation.

How about political bias? How about Fox, MSNBC, New York Times, and Washington Examiner? They have a point of view. They have a bias.

And we say, people only have access to those sites. Actually, there are a lot of social media sites out there, and more are being introduced, and there will be more introduced over time.

The idea that saying, hey, we are going to subject them—we are going to break them up for antitrust reasons. Do not forget, data and social media is international. It is not like it is just a U.S. enterprise. TikTok is owned by the Chinese. We are not going to break them up, even though some of us might prefer that we could. We have to think about the competition coming from around the world.

I must admit I have a hard time seeing how we are going to change the algorithms. If we are going to change the algorithms for social media, how about changing the algorithms for Fox, MSNBC, the New York Times, NBC, and CBS? I mean, each of them have their own bias. They have their own way of assessing how they get the most eyeballs, how they get the most clicks, if you will, even if it is just the channel changer. I do not see how that works.

Likewise, to say we are going to set the standards for what is truth and not truth, I do not know how government does that consistent with the First Amendment.

I understand why a CBS decides what they are going to show and why a New York Times decides what they are going to print. That is their standards. By the way, their people can respond to it, and people can boycott or stop subscribing as a result of what they are saying.

This is an area, in my opinion, that is extraordinarily fraught for government action. What we can do that does not violate the freedom of expression, of which we are intent on having in our country, is something which I think we have to be careful in considering.

I would note that what is unusual about social media, among many things, is the precision with which data is able to be gathered and people are able to be targeted. The New York Times may say, I want to get a left-of-center subscriber base, but they cannot get it literally down to the home to print the articles that you might want to watch although as they go online, and increasingly

they are an online forum, they will be able to provide the articles that you want to watch and will compete on that basis.

Anyway, I guess I raise a couple of things as possibilities in a topic I know very little about but care very deeply about because of my kids and grandkids. That is, one, to require social media companies to certify that a person who is posting or commenting is an actual person, so very simply saying you have a responsibility as a social media company to determine if, "JAB123" is actually a human being and also what country they are from perhaps but whether it is a government or a corporation but whether it is human being. That would be No. 1, and that is probably the easiest constitutionally.

The other is something I would like to do. I would like to require actually people to be identified as to who they are, what their name is, if not their address, at least their name so people are responsible for what they post and what they comment on. I think that would fly in the face of the First Amendment and the right to be anonymous.

Perhaps having more emphasis on "blue checks," if you will, verifying people and really insisting that social media companies encourage more people to be identified and perhaps even having an alternative site, which is, if you will, there is Facebook where there is no blue check, but then there is another portion of Facebook which is all blue check. To comment or to post, you have to have a blue check on that particular wave. That would give people some confidence that there is a real human being that is willing to stand behind that comment.

Those are the only two ideas I could come up with that I thought would be helpful here and not fly in the face of the same challenge we have with all of our media, that there is disinformation that we do not make illegal, that there is bias that we do not make illegal, that there is profit motive that we do not make illegal, that there are algorithms that draw more articles that people want to read that we do not make illegal.

I do not see the course that you all are talking about as being consistent with what the rest of our media system is providing.

Do any of you have any comments about those ideas that I suggest? I am concerned about the topic, but I just cannot seem to find a way to resolve some of the problems we describe.

Yes, Dr. Persily?

Mr. PERSILY. Yes.

Ms. KORNBLUH. Can I just try real quick? Just because I worked at Nielsen, so you are absolutely right, and I think it is great that you are pulling us back to this history of newspapers and broadcast. What is different, I think, is that in the newspaper, first of all, speakers did not have this much power. That is true in broadcast, too. We had ownership restrictions. You could not own a newspaper and a broadcast in the same market. You could not own——

Senator ROMNEY. But you could have the whole country, like Sinclair does.

Ms. KORNBLUH. But you could not in the past. We tried to deal with some of these issues but mostly through transparency. Right? So there is a masthead.

Senator ROMNEY. Yes.

Ms. KORNBLUH. There are codes and standards that you hold yourself to if you are a newspaper. If you are a broadcaster, we had the payola rules so that the DJ would have to reveal if he was being paid to play that record. The political ads that you run, you had to be transparent and say, I paid for these ads.

What we are seeing on social media, I think, is we have not worked through those systems. We do not have that transparency. That buyer cannot beware because they just do not know who is pushing this at me, why are they pushing it at me. The codes and standards that they proclaim they are not really honoring. I have seen——

Senator ROMNEY. Yes, the transparency, that is the direction I am going is transparency. Yes.

Ms. KORNBLUH. I think what you are talking about is really interesting. I might tweak it a little bit and say, if you are Facebook and you say you have a real name policy, then you better have a real name policy. If you have a blue checkmark and it says it is a human, it better be a human. If it is a bot, it should be labeled as a bot.

Senator ROMNEY. Yes.

Ms. KORNBLUH. If it is a deepfake, you need to let people know it is a deepfake.

Senator ROMNEY. Totally agree.

Ms. KORNBLUH. That is really consistent, I think, with the way we approach—with the way the industry had norms and standards in the journalist industry and in broadcasting, what the public interest standard required. In fact, those kinds of user empowering——

Senator ROMNEY. That course makes sense.

Ms. KORNBLUH. Exactly.

Senator ROMNEY. Yes, sir?

Mr. PERSILY. I agree with that. Just in terms of the differences with TV, which is that we know what Tucker Carlson or Rachel Maddow is saying every night. There is a record of it. We know who was able to see it. We do not know that with speakers on the internet. Right?

That is where the transparency legislation would come in—and I attached some to my testimony—because we need to figure out sort of who saw what, when, and how. Right? With television and other kinds of broadcast media and even some print media, we can figure that out, but not with social media platforms.

Senator ROMNEY. Yes. Thank you.

Mr. Chairman, thank you.

Chairman PETERS. Thank you, Senator Romney.

Senator Padilla, you are recognized for your questions.

OPENING STATEMENT OF SENATOR PADILLA

Senator PADILLA. Thank you, Mr. Chair.

Let us just jump right into it. When I was Secretary of State of California, prior to joining the U.S. Senate, I saw more than our fair share of bad actors seeking to discourage communities, particularly communities of color, from exercising their right to vote by

gaming social media and exploiting gaps in trusted sources and data voids.

Like many of you, I am alarmed by the recent Facebook Papers and, in particular, their absolute failure at Facebook to invest in integrity systems responsive to non-English languages and cultures around the world and right here in the United States as well.

While the focus of today's hearing is on social media, I hope we can keep in mind the broader information ecosystem. We need to equip our kids and neighborhoods with media and information literacy skills. We need to address the collapse of local journalism—and I will be asking a question about that here in a minute—that are expanding news deserts across the country.

Now this hearing is about platform design choices, and it would be an oversight not to reflect on user data. User personal data is indeed what fuels targeting on social media, informing the content users see and how ads are targeted to them. It is all driven by their data.

Now strong data privacy protections may help address some of the unhealthy dynamics that we see online. We have been talking about a national privacy law for a long time, and I think it is time that Congress finally gets it done.

Question for Dr. Persily: How can a strong privacy law reduce the risk of echo chambers, micro-targeting of disinformation or exploitative advertising which targets specific individuals or groups based on profiling?

Mr. PERSILY. Thank you. Thank you, my Senator, for that question. It is good to see you. I hope to see you back in California.

As I mentioned before, we tend to think of internet regulation in two domains: One is on speech, sort of explicitly, and Communications Decency Act (CDA) 230 reform is one of them. Then kind of structural or infrastructure design questions, and privacy would be one of those. But I actually think that we need to start recognizing that they sort of bleed over into each other.

I think you are right to point out that through national privacy legislation and regulating the kind of data that the firms can collect, that we will be able to get at some of these problems because if you think that part of the problem here is the micro-targeting of messages that necessarily select out audiences for manipulation and persuasion and the like, that it is only enabled because of the amount of data that the firms have. If we had rules on what particularly the big platforms could do in terms of collecting data, I think it would go a long way in addressing some of these speech problems as well.

Senator PADILLA. Thank you. Now on an extended topic because I mentioned reporting of the Facebook Papers reveals an abdication of responsibility to meet the needs of non English speaking Facebook users around the world, and it is happening here in the United States as well. We are blessed with a very diverse population. Spanish-language disinformation about how to vote, where to vote, when to vote, et cetera, ran rampant on platforms in 2020 as compared to similar content in English.

It is not limited to just election information, by the way. I saw it significantly when we were doing census outreach and assistance at that critical time in 2020. We continue to see it in regards to

the COVID–19 pandemic, the safety of vaccines, et cetera. It is absolutely unacceptable.

Question for Dr. O'Neil: Why do you think platforms are failing even more for non-English language speakers, and in what ways can Congress be helpful in this space?

Ms. O'NEIL. Thank you for the question. You are absolutely right that they are failing in non-English language spaces. In India, which is a huge problem as we have read about in the last few days, but for many months and years we have known this, there are just too many language dialects. Facebook just does not want to hire people to know those languages. It would be very expensive.

It is a cost issue on top of the fact that I already mentioned, that the filters for hateful or extreme content are essentially keyword searches, so you need to understand what keywords to search for. You need a lot of experts working full-time on this, and they just simply do not want to pay for that.

It would be very expensive. So it is clear.

But I want to make it also clear that there is no simple solution. I am not suggesting that they are avoiding doing something simple to solve these problems. This is actually really hard. Their mistake is not that they are not doing it well. It is that they are pretending they can do it. They simply cannot do it because of what I have said before, that AI does not understand truth so it is just simply looking for keywords.

To the extent that Facebook cares about looking good, they care much more about looking good to English-speaking Americans and to people like you.

I would say one quick story. I gave a talk in the Ukraine recently, and one of the audience members was a parliament member of the Ukraine. She said, what can we do here in the Ukraine about Russian propaganda that undermines people's trust in our elections?

I was like, wow, I really do not know. I mean, you have even less power over Facebook than the Senators in the U.S. Senate. It is a really important question.

Senator PADILLA. I will just add to your commentary about looking good in front of Americans or looking good in front of people like us, members of the U.S. Senate. Sadly, looking good in front of investors and Wall Street seems to trump it all.

My final question in the time remaining: We know today's information ecosystem is complex. In addition to facing organized propaganda campaigns, social media users encounter more content at higher speeds. Right? The innovation technology has a role to play here.

I worry that efforts to help communities critically engage with information is not keeping pace. It is also not lost on me that we have seen an explosion of propaganda campaigns aimed at manipulating and intimidating communities online while we are in the midst of a collapse of local journalism and independent media.

My final question is for Ambassador Kornbluh. I welcome any thoughts you may have on how the shuttering of local news outlets has impacted how users engage with content that they consume online.

Ms. KORNBLUH. That is such an important question. As you say, it creates this vacuum, and people are served things online. Again, I think we have to underscore that so much of what happens online is manipulation. People do not know what they are being served or who is behind it.

There are these pretend local outlets that they see online that seem to have a name that suggests that they are local, but they in fact are often controlled centrally. The news stories can even be constructed by AI. They think they are getting local news, but they are actually being fed information that serves a political interest or a financial interest, and they are not aware of it. There is no alternative, so they do not have access to the civic information that they need to be a citizen.

The Secretary of State of Colorado just made a really interesting point a couple of weeks ago. She said, if I am standing up at a podium and having a press conference, and the voters in my State are reading about, something completely different, SharpieGate or whatever it is, online, I am not in conversation with them because she is communicating to them over these social media platforms and that is sort of a funhouse mirror of what is going on.

I think we really have to think about how is the civic information, public health information, election administration information, how is it going to get to citizens at a time when local news is so undermined.

I should say part of the reason local news is undermined is because it was supported by advertising and all those advertising dollars have now gone to the platforms. There is no revenue base for local news. So this is a problem.

It is really a fundamental democracy problem. The press is mentioned in the Constitution. It is something we really have to address.

Senator PADILLA. Thank you very much.

Thank you, Mr. Chair.

Chairman PETERS. Thank you, Senator Padilla.

I would like to take this opportunity to thank once again our witnesses for joining us here today. I think I speak for the entire Committee when we say we appreciate your very unique insight and expertise and helping us examine this critical issue and navigate the tough challenge today.

Today's hearing provided an opportunity for us to learn about the role social media platforms play in the amplification of extremist content, including White Nationalist and anti government ideologists or ideologies. We heard expert testimony about how their algorithms and recommendation tools are driving extreme content users, how that exposure to harmful content can translate to real-world violence, and how their business models built on user engagement and targeted advertising appear to prioritize profits over safety.

The connection between extremist content on social media platforms and domestic terrorist attacks in our communities is, without question, a national security threat and one this Committee will continue to examine. Our next steps will be to include a hearing from the social media companies themselves, and we will work to bring more transparency to this pressing issue. I am also seeking

information from both the Department of Homeland Security and the FBI on this threat and will be looking for needed reforms.

The record for this hearing will remain open for 15 days, until 5 p.m. on November 12th, 2021, for the submission of statements and questions for the record.

This hearing is now adjourned.

[Whereupon, at 12:12 p.m., the Committee was adjourned.]

APPENDIX

Chairman Peters Opening Statement As Prepared for Delivery
Full Committee Hearing: Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content
October 28, 2021

Thank you to our witnesses for joining us today. Our Committee has held several hearings this year examining the rise in domestic terrorism.

Today's hearing will focus on the role social media platforms play in the amplification of domestic extremist content, and how that content can translate into real world violence.

Yesterday marked three years since a white supremacist gunman opened fire in a Pittsburgh synagogue, killing 11 innocent people, in the deadliest attack on the Jewish community in the United States.

The attacker used the fringe social media platform Gab prior to the attack to connect with like-minded extremists, and spread his own hateful, anti-Semitic views online.

While that violent and hateful ideology has long terrorized communities, for many Americans, this shocking attack was a first glimpse at how quickly increased exposure to extremist content, and users with similar beliefs, can radicalize domestic terrorists, and drive them to act on their violent intentions.

Less than a year after the Tree of Life attack, we saw a white nationalist open fire in an El Paso shopping center. This attacker was one of many who viewed video of the Christchurch mosque massacres that widely circulated on social media just months earlier, a video he reportedly cited as inspiration for his deadly attack in a 2,300 word racist manifesto he also posted online.

And on January 6, 2021, we saw a stark example of how individuals went beyond seeing and sharing extreme content across numerous social media platforms. They were spurred to action by what they repeatedly saw online, and ultimately a mob violently attacked Capitol Police and breached the Capitol Building.

In attack after attack, there are signs that social media platforms played a role in exposing people to increasingly extreme content, and even amplifying dangerous content to more users.

Yet there are still many unanswered questions about what role social media platforms play in amplifying extremist content. We need a better understanding of the algorithms that drive what users see on social media platforms, how companies target ads, and how these companies balance content moderation with generating revenue.

For the majority of social media users who want to connect with distant family and friends, or stay up to date on their favorite topics, there is very little transparency about why they see the content, recommendations, or ads that populate their feeds.

While social companies have promoted how they are providing more curated content for their users, we have seen how users can be shown increasingly polarizing content. In worst case scenarios, users are reportedly recommended more and more extreme content, nudging them down a rabbit hole.

Recent reporting, Congressional testimony, and revelations in the Facebook Papers, have shed some light on business models that appear to prioritize profits over safety, and decisions that appear to disregard the platforms' effects on our homeland security.

It's not enough for companies to simply pledge that they will get tougher on harmful content, those pledges have gone largely unfulfilled for several years now. Americans deserve answers on how the platforms themselves are designed to funnel specific content to certain users, and how that might distort users' views and shape their behavior, online and offline.

As part of my efforts to investigate rising domestic terrorism, including the January 6th attack, I've requested information from major social media companies about their practices and policies to address extremist content, so that we can better understand how they are working to tackle this serious threat.

While we are continuing to work with companies to get answers and examine relevant data, I'm looking forward to hearing from our experts today about how these platforms balance safety and business decisions, and specifically how these decisions relate to rising domestic extremism.

U.S. Senate Committee on Homeland Security and Governmental Affairs
Hearing on "Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
Thursday, October 28, 2021

Karen Kornbluh Written Testimony

Chairman Peters, Ranking Member Portman, and distinguished Committee members, thank you for this opportunity to testify today on the relationship between social media platforms and the amplification of domestic extremism.

My name is Karen Kornbluh and I direct the Digital Innovation and Democracy Initiative at the German Marshall Fund of the U.S., where I am also a Senior Fellow. Previously, I served as United States Ambassador to the Organization for Economic Cooperation and Development.

The topic of this hearing is critical for the future of the country and our national security. In the words of Timothy Langan, Assistant Director of the FBI Counterterrorism Division, "The greatest terrorism threat…today is posed by lone actors or small cells who typically radicalize online and look to attack soft targets with easily accessible weapons."[1]

An internal Facebook test showed how these actors can be radicalized online. A fake Facebook account created for a fictional "Carol Smith," a 41-year old conservative mother from North Carolina, was recommended pages and groups related to QAnon within days of its creation and was recommended an account associated with the militia group Three Percenters within three weeks.[2]

This research and other documents released by the Facebook whistleblower underscore that design features of large social media platforms are creating a feedback loop that pushes some Americans toward violent extremist ideologies and are facilitating large-scale extremist organizing.

They also make clear that the companies' current strategy of after-the-fact, "whack-a-mole" take-downs is grossly insufficient to address this systemic vulnerability.

Until social media companies' incentives are changed, the problem of online radicalization and violent extremism will continue to grow.

Down the "rabbit hole" -- social media goes well beyond providing users tools to connect organically with others; it pulls users into rabbit holes and empowers small numbers of extremist recruiters to engineer algorithmic radicalization.

---

[1] Timothy Langan, "Confronting White Supremacy: Examining the Biden Administration's Counterterrorism Strategy," September 29, 2021.
[2] Ryan Mac and Sheera Frenkel, "Internal Alarm, Public Shrugs: Facebook's Employees Dissect Its Election Role," New York Times, October 22, 2021.

The documents released by Facebook whistleblower Frances Haugen reveal the risk that algorithms that rank and recommend content based on user engagement lead some users down information "rabbit holes" into increasingly narrow echo chambers where violent conspiracy theories thrive. People radicalized through these rabbit holes make up a small slice of total users, but at scale that means a great many users.

An August 2019 Facebook internal memo admitted very clearly that, "the mechanics of our platform are not neutral" and, in fact, that core product mechanics, including virality, recommendations, and optimizing for engagement are key to why hate and misinformation flourish on the platform.[3] Research shows this is true for other platforms as well.

Facebook groups are a key vector of recruitment. Internal research found 70% of Facebook political groups in the U.S. were rife with hate, bullying, harassment, misinformation and other rule violations, and that many of the most toxic civic groups were "growing really large, really fast."[4] Organizers benefited from a variety of tools to build groups and pages. For example:

- Facebook's own algorithms recommend extremist groups to users – to such an extent that they were responsible for a striking 64% of these groups' new members in 2016.[5] For example, Facebook directs users who "like" certain militia pages toward other militia groups.
- "Super inviters" or "invite whales" created invitation links that could be shared on or off Facebook and can easily coordinate their invitations.
- The platform provides group members recommendations of others to invite based on data about users' activities, allowing groups to draw from other conspiracy and militia groups.
- A report revealed the problem of bait-and-switch groups, in which pages that post about cute animals and other innocuous topics build tens of thousands of followers then sell the page to the highest bidder, at which point it becomes a vector for extremist content pushed at unwitting users.
- And groups that were punished for breaking the rules easily could evade take downs by re-establishing themselves with new names, from which they could continue viral recruitment strategies.

As an example of the way these tools can be used to grow groups, a Facebook internal report on the Stop the Steal movement revealed that 0.3% of group members were responsible for 30% of invitations to join.[6] Organizers sent hundreds of invitations to members of other groups, resulting in high membership overlap with Proud Boy and militia groups and fueling Stop the Steal Groups' meteoric growth rates.

---

[3] Mike Isaac, "Facebook Wrestles With the Features It Used to Define Social Networking," New York Times, October 25, 2021.
[4] Shannon Bond and Bobby Allyn, "How the 'Stop the Steal' movement outwitted Facebook ahead of the Jan. 6 insurrection," NPR, October 22, 2021.
[5] Jeff Horowitz and Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive," May 26, 2020.
[6] Ryan Mac, Craig Silverman, and Jane Lytvynenko, "Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here," BuzzFeed, April 26, 2021.

Ads are another vector for radicalization. These can be targeted to small audiences based on detailed data gathered on users. And, as the Facebook whistleblower emphasized to a parliamentary select committee, ads on Facebook are priced "partially based on the likelihood that people like them, reshare them, do other things to interact with them — click through on a link" and therefore, "An ad that gets more engagement is a cheaper ad."[7]

Similar algorithmic radicalization is evident on other platforms: TikTok's recommendation algorithm also promotes content from QAnon, the Patriot Party, Oath Keepers, and Three Percenters.[8] After users interacted with trans-phobic videos on TikTok, the recommendation algorithm fed users videos with hate symbols, white supremacist and anti-Semitic content, as well as coded calls to violence.[9]

YouTube has 290 extremist channels, according to new research. When researchers showed an interest in militant movements, YouTube suggested videos to them with titles like "5 Steps to Organizing a Successful Militia" and "So You Want to Start a Militia?" The platform also recommended videos about weapons, ammunition, and tactical gear to what the researchers at the Tech Transparency Project call "the militia-curious viewer."[10]

This algorithmic radicalization has already resulted in extremist violence. Air Force Staff Sergeant Steven Carrillo in 2020 shot and killed a protective security officer and wounded his partner, before killing a sheriff's deputy. He had begun engaging with the extremist group Boogaloo Bois on Facebook and eventually was in direct contact with prominent Boogaloo members. He purchased a device that converts AR-15 rifles into fully automatic machine guns from a website that advertised to Boogaloo Facebook groups. As Carrillo was being pursued by police, he sent a WhatsApp message to members of the heavily armed Boogaloo militia faction he had recently joined, telling them to join him.[11]

Radicalization is a cross-platform phenomenon.

Extremists can organize on specialized, less moderated sites and use other platforms to radicalize others, silence critics, or swamp the news cycle. The magazine *Nature* found individuals move from mainstream platforms to less moderated ones like 4Chan or Telegram in a few clicks and then reintroduce fringe content to the original mainstream platform.[12]

---

[7] Isobel Asher Hamilton, "Facebook whistleblower Frances Haugen says it's cheaper to run 'hateful' ads on the platform than other kind of adverts," Business Insider, October 26, 2021.
[8] Olivia Little, "TikTok is prompting users to follow far-right extremist accounts," Media Matters for America, March 26, 2021.
[9] Olivia Little and Abbie Richards, "TikTok's algorithm leads users from transphobic videos to far-right rabbit holes," Media Matters for America, October 5, 2021.
[10] Tech Transparency Project, "YouTube's Filter Bubble Problem is Worse for Fox News Viewers," October 24, 2021.
[11] Gisela Pérez de Acha, Kathryn Hurd, and Ellie Lightfoot, "'I Felt Hate More Than Anything': How an Active Duty Airman Tried to Start a Civil War," PBS, April 13, 2021.
[12] N. Velásquez, R. Leahy, N.J Restrepo, et al., "Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms," Scientific Reports 11, 2021.

The social media companies' whack-a-mole approach of taking down individual pieces of violative content or accounts after damage is done fails to prevent algorithmic radicalization.

When people inside Facebook discussed a more systematic approach – one that would be content and viewpoint agnostic -- to restrict design features that amplify incendiary and divisive posts, the company rejected most of these ideas. Head of Facebook Health Kang Xing Jin proposed in 2019 that the company dial back on automated recommendations. Other proposals included dialing back algorithmic virality.[13] Another option was to enforce its rules prohibiting individuals from operating multiple accounts, since many of these accounts are purveyors of violent political activity, according to Facebook employees.[14]

The company rejected most of these ideas and largely left it to content moderation to play after the fact whack-a-mole. But even then, Facebook tied moderators' hands behind their backs. It maintains a whitelist that exempts VIP users with the largest footprints from the stated rules -- even though that meant the company was "not actually doing what we say we do publicly," according to an internal report.[15]

The resulting moderation process is catching only small percentages of violative content: only 3 to 5% of hate speech and 0.6% of content that depicts or incites violent content.[16] Even though Facebook says militia groups are banned, in reality, roughly 70% of the Facebook militia pages identified in a Tech Transparency Project report had the word "militia" in their name.[17]

The platforms are considering doing more to address this serious problem. Twitter recently released an internal report on its algorithms.[18] YouTube instituted strict enforcement of their rules against false election claims.[19] Facebook itself has launched a new project to examine the pathways to radicalization.

While Congress works toward comprehensive privacy legislation and various antitrust investigations proceed, targeted steps are needed now to limit algorithmic radicalization.

---

[13] Jeff Horwitz and Justin Scheck, "Facebook Increasingly Suppresses Political Movements It Deems Dangerous," Wall Street Journal, October 22, 2021.
[14] Julia Arciga and Susannah Luthi, "How Facebook users wield multiple accounts to spread toxic politics," Politico, October 25, 2021.
[15] Jeff Horwitz, "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt," Wall Street Journal, September 13, 2021.
[16] Deepa Seetharaman, Jeff Horwitz and Justin Scheck, "Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts," Wall Street Journal, October 17, 2021.
[17] Tech Transparency Project, "Facebook's Militia Mess," March 24, 2021. For a detailed breakdown of Facebook's tiered process of addressing Dangerous Individuals and Organizations, see Facebook Transparency Center, "Dangerous Individuals and Organizations," last accessed October 26, 2021.
[18] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjera, and Moritz Hardt, "Algorithmic Amplification of Politics on Twitter," Twitter, 2020.
[19] YouTube, "Supporting the 2020 U.S. election," December 9, 2020; YouTube, "Updated Policy," Twitter, January 7, 2021.

First, a "black box flight data" recorder, or actionable transparency is needed. We shouldn't need a whistleblower to access data. The Federal Trade Commission should require more transparency, just as the National Transportation Safety Board gets access to data on airplane crashes or the Environmental Protection Agency releases data on pollution.

- It can require third-party audits of terms of service enforcement that are routine and publicly available.
- Researchers also need access to privacy-protected retrospective data.
- The bipartisan Honest Ads Act would provide the same transparency about ads as is required on broadcast but should be supplemented by Know Your Customer rules that prevent dark money or foreign actor ad funding. Platforms must have robust systems for archiving political advertisements that are searchable and sortable through an API.

Second, the industry must implement and regulators must enforce a digital code of conduct. Where regulators lack explicit authority or the First Amendment prohibits them from telling companies what to do, Congress or the FTC can drive platforms to clean up their acts with watchful oversight and enforcement. Such a code should include commitments such as:

- Eliminating design features, such as automatic group recommendations, that provide turn-key solutions for radicalizers.
- Using "circuit breakers" to prevent quick viral spread of radicalizing content, while human reviewers determine whether the content violates platform policies or poses a risk to public safety. [20]
- Best practices to avoid linking to platforms that consistently permit illegal and terrorist activity. This would prevent violent extremists from organizing on smaller platforms and then using the major tech platforms to spread their content—including so-called "manifestos" purporting to explain and justify acts of violence.

Any violation of the code could be enforced as an FTC Act consumer protection violation. Conditioning Section 230 immunity on companies following a robust code of conduct could provide further incentives to adopt a code.

It is essential that we act to protect our country from violent extremism and divisive hate. The Facebook papers provide a telling look in the rearview mirror. But the line between our offline and online interactions is disappearing. As we move to Internet 3.0, interacting through our avatars, it's essential that we build in protections against further radicalization and violent extremism.

[20] Ellen Goodman and Karen Kornbluh, "Social Media Platforms Need to Flatten the Curve of Dangerous Misinformation," Slate, August 21, 2020.

Congressional Testimony

Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content

David L. Sifry

Vice President, Center for Technology and Society

ADL (Anti-Defamation League)

SENATE HOMELAND SECURITY AND GOVERNMENTAL AFFAIRS COMMITTEE

Washington, DC

October 28, 2021

10:15 a.m.

Working to stop the defamation of the
Jewish people and to secure
Justice and fair treatment to all since 1913

I. INTRODUCTION

During the past several years, there has been a tectonic shift in the way communities across the world integrate digital and social networks into their daily lives. Anyone who has been paying attention to extremist activity across the country knows that online hate and extremism are amplified, reinforced, and spread by the chorus of disinformation and hatred that runs rampant across social media. New adherents to extremism are recruited, fed a radicalizing diet of conspiracy theories, and connected to others who share those views. Most recently—as underscored by the recent explosive proof provided by leaked, internal Facebook documents—we've seen horrifying evidence of how the influence of social media spurs extremist activity, terror, and anti-democratic offline violence in this country and around the globe.

For ADL, the spread of conspiracy theories and on-the-ground domestic terrorism has been shocking but not surprising. For example, what the Facebook leaks have provided is proof of what so many of us knew to be the case—that the platform and its top executives made intentional choices to allow and even spur online harm (with offline consequences) in service of growth and revenue. The difference between what these executives say and what they do is revealed in all its chilling hypocrisy. What is clear now is that companies like Facebook are not only prioritizing profit—they are doing so at the expense of our safety, security, and democracy because it is good for their bottom line. In the end, that seems to be the most important principle for the few individuals who run the largest, most powerful, and most unaccountable communications, news, entertainment, and surveillance system the world has ever known.

Social media's amplification of extremism, disinformation and conspiracy theories is one of the greatest threats to democracy in this country and to the safety of vulnerable individuals and communities worldwide. Hatred spread online has resulted in deadly terrorism in this country: from Charleston to Charlottesville to Pittsburgh, to Poway and El Paso, we have seen the fatal consequences of white supremacist extremism that often has a clear nexus to social media. We cannot afford to minimize the threat of social media's algorithmic amplification of extremism and hate. We need a bipartisan "whole of government approach"—indeed, a "whole of society" approach—to interrupt domestic extremism and harmful content amplified by social media companies in their pursuit of profit.

ADL brings unique expertise to the table in the fight against online hate. Our Center on Extremism (COE) examines the ways extremists and white supremacists exploit digital affordances to spread their messages, recruit adherents and commit acts of terrorism. Our Center for Technology and Society, which has deep policy and technical product expertise, generates advocacy-focused solutions to make digital spaces safer and more equitable. Our proficiency in these spaces, in addition to more than a century of work to fight against hate and for civil rights, informs ADL's analysis of the online hate and extremism ecosystem and what we can do to combat it. This testimony will explore how platforms spread hate and extremism, why hate-filled and extreme content is favored by platforms such that the entire enterprise is engineered and operated for its expansion—because it drives profit—and the ways in which online extremism can lead to offline violence. Finally, ADL lays out several policy recommendations for lawmakers and the technology sector to fight online hate and extremism meaningfully and significantly.

II. ADL'S FIGHT AGAINST ONLINE HATE

Since 1913, the mission of ADL (the Anti-Defamation League) has been to "stop the defamation of the Jewish people and to secure justice and fair treatment to all." For decades, one of the most important ways in which ADL has fought against bigotry and antisemitism has been by investigating extremist threats across the ideological spectrum, including white supremacists and other far-right violent extremists, producing research to inform the public on the scope of the threat, and working with law enforcement, educators, the tech industry, and elected leaders to promote best practices that can effectively address and counter these threats. As ADL has said time and time again, where people go, hate follows—including online.

ADL has invested resources and become a leader in fighting online hate since we launched the Center for Technology and Society (CTS) in 2017. CTS is a leader in the global fight against online hate and harassment. In a world riddled with antisemitism, bigotry, extremism, and disinformation, CTS acts as a fierce advocate for making digital spaces safe, respectful, and equitable for all people. CTS also plays a unique role among civil society organizations working on fighting online hate. It brings to bear decades of lived experience rooted in a community that has been targeted, often lethally, by bigots and extremists and leverages ADL's decades of expertise in tracking and fighting extremism and antisemitism.

One of the signature differentiators of CTS is the fact that it works in five key areas: policy, research, advocacy, incident response, and product development. It recommends policy and product interventions to elected officials and technology companies to mitigate online hate and harassment; drives advocacy efforts to hold platforms accountable and push hate back to the fringes of society; produces data-driven applied research by analysts and a network of fellows; sheds new light on the nature and impact of hate and harassment on vulnerable and marginalized communities; brings to market technical tools and products that meet the crucial need for independent data measurement and analysis to track identity-based online hate and harassment; and empowers targets of harassment by responding to online incidents and pushing platforms to create safer online spaces for all.

In our direct engagement with platforms, CTS has emphasized the need for them to adopt anti-hate-by-design principles. This concept was first popularized in the area of privacy (known as privacy-by-design) but can and should be applied to building less hate-filled platforms. Our recommendations include several steps that would help inculcate a culture of anti-hate-by-design to can be implemented across social media company systems, policies, and product developments.

III. PLATFORMS SPREAD HATE AND EXTREMISM

There is no question that the prevalence and impact of online extremism is growing. The spread of QAnon and its consistent elevation of antisemitism, the mainstreaming of the foundational white supremacist and neo-Nazi "Replacement Theory," the Big Lie about the 2020 presidential election, and COVID conspiracies, all are examples of extremism and hate that has become increasingly normalized and mainstreamed—in large part because of its viral spread online. Leaked documents from the Facebook whistleblower show this trend, our own research and that

of others confirm it, criminal prosecutions demonstrate it, and government and news investigations continue to provide a firehose of evidence.

Discovery in civil cases, like the lawsuit against the neo-Nazi and white supremacist organizers of the 2017 Unite the Right rally in Charlottesville, which began this week, provide still more chilling examples. Extremists' online presence has reverberated across a range of social media platforms. This content is intertwined with hate, white supremacy, racism, antisemitism, and misogyny—all through the lens of extreme ideologies. Such content is enmeshed in conspiracy theories and explodes on platforms that are themselves tuned to spread disinformation.

We need to look no further than the deadly insurrection at our Capitol, which ADL has repeatedly called the most predictable terror incident in American history because it was planned and promoted out in the open on mainstream platforms such as Facebook, Twitter, Instagram, YouTube, and Reddit, as well as fringe platforms such as Parler, Gab, 4Chan, and Telegram. As confirmed by leaked internal Facebook documents, the insurrectionists' actions were the product of weeks, months, and years of incitement, spread across the social media ecosystem that services nearly 300 million people in the U.S. and billions around the world.

A. Mainstream Social Media Platforms

Fringe platforms, despite having relatively small userbases, make use of Big Tech platforms like Twitter and Facebook to increase their reach and influence. Since Twitter's 320 million and Facebook's 2.85 billion users dwarf the hundreds of thousands of users on fringe sites, extremists leverage these mainstream platforms to ensure that the hateful philosophies which often began to germinate on message boards like Gab and 8chan (now 8kun) find a new and much larger audience. Mainstream platforms serve as a gateway for extremists to recruit curious individuals. Extremists use strategies like creating private pages and events; using coded language (called dog whistles) to imply and spread a hateful ideology on mainstream platforms; and linking to hate-filled sites to avoid content moderation.

Facebook and Twitter generally allow users to link to pages on fringe hate-filled sites, which allows visitors to mainstream sites to get to highly problematic content with little to no effort. ADL's COE found in an October 2021 study that despite Twitter's ban on external links to hate speech, extremist material and conspiracy theories, this content is frequently shared on Twitter via links from the far-right "free speech network" Gab. More than 112,000 tweets were posted containing links to Gab content between June 7 and August 22, which included antisemitism, misinformation relating to COVID-19 or the vaccines, and content promoting QAnon.

Big Tech platforms are not unwitting accomplices or merely tools for extremists to link to fringe platform content. On the contrary, platforms' algorithms amplify misinformation, extremist, and white supremacist content; connect adherents; and host and recommend anti-democratic, extremist and hate-focused groups and events. For example, last fall a single "Stop the Steal" Facebook group gained more than 300,000 members within 24 hours. Thousands of new members joined this group by the minute and some of them openly advocated for civil war.

Big tech companies know their platforms' product features are problematic. At a congressional hearing in March 2021, Twitter Chief Executive Officer Jack Dorsey admitted that his platform had "contributed to the spread of misinformation and the planning of the attack" on the U.S. Capitol on January 6, 2021. In the same hearing, Facebook's CEO Mark Zuckerberg disagreed with the assessment that Facebook had profited from the spread of disinformation and touted his platform's efforts to combat it.

Importantly, however, documents disclosed to the SEC by Facebook whistleblower Frances Haugen make it clear that Facebook was aware of both the specific role its platform played in the insurrection and the broader role the platform plays in the spread of disinformation, extremism, and hate. The SEC disclosure includes statements from Facebook's internal documents. These documents stated Facebook's role in augmenting "combustible election misinformation," noting "we amplify them and give them broader distribution." Internal Facebook documents also stated that the company had "evidence from a variety of sources that hate speech, divisive political speech and misinformation on Facebook and the family of apps are affecting societies around the world...Our core products mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish."

Over the last few years, TikTok—a social media app that allows users to create and share short videos—has also hosted hate and extremism. As ADL's COE documented in August 2020, while much of the content on TikTok is lighthearted and fun, extremists have exploited the platform to share hateful content and recruit new adherents. A recent review of the platform found that antisemitism continues to percolate across the app, including content from known antisemitic figures as well as posts perpetuating age-old antisemitic tropes and conspiracy theories. It should be noted that when alerted to the content that ADL found, TikTok took down the specific content, but they are still woefully inadequate when handling reports from ordinary users. While we appreciate their removing the specific content and their stated commitment to a zero-tolerance policy on antisemitism and hate, we are concerned that it took our notification to do so and urge them to systematically address this serious issue. Earlier this year ADL's CTS released a report that showed TikTok is still far too slow in taking down antisemitism reported by ordinary users and it still has plenty of work to do to ensure that hate is adequately remediated.

Recordings of Louis Farrakhan, Rick Wiles (founder of TruNews), and Stephen Anderson—all antisemitic individuals whose bigotry has been thoroughly documented by ADL—were readily available on TikTok in 2021. One such post, shared on May 23, 2021, showed a clip of a TruNews segment in which Rick Wiles states: "And our leaders are lowlife scum that screw little girls so the Jews can screw America...we've allowed Kabbalah practicing Jews to defile the nation." TruNews—a fundamentalist Christian streaming news and opinion platform that produces antisemitic, anti-Zionist, anti-LGBTQ+, and Islamophobic content—has been banned from YouTube and Facebook for violating the platforms' content rules.

B. Gaming Platforms

Online video games share many of the attributes of social media platforms. Games spread hate and extremism and operate at a comparable scale to social media platforms. According to the Entertainment Software Association, there are approximately 227 million gamers in the United States. Gaming analytics firm NewZoo's global market report put the gaming industry's revenue at approximately $176 billion globally. With those figures in mind, the importance of addressing hate and extremism in gaming is critical.

ADL's 2021 study of hate, harassment, and positive social experiences in online games explored players' in-game exposure to topics such as extremism and disinformation. Alarmingly, 8 percent of adult gamers (18-45) and 10 percent of teen gamers (13-17) witnessed discussions about white supremacist ideology in online multiplayer games. Seventeen percent of adult gamers saw hateful messaging linking the COVID-19 pandemic to the Asian community, and 13 percent of adult gamers saw hateful anti-immigrant messages spread in online games. The survey also showed that nearly one-in-ten online multiplayer gamers (7 percent) come across Holocaust denial discussions while playing. As we continue to pay deeper attention to the impact social media's algorithms and business model have on domestic terrorism and extremism, we must consider the way online video games have similar consequences.

IV. DOMESTIC TERRORISM AND EXTREMISM ARE GOOD FOR PLATFORMS' BUSINESS MODELS

Big Tech's fundamental business model—targeted advertising—maximizes profits by optimizing the product mechanics of the platform to increase user engagement. AI and algorithms, surveillance advertising, subscription models and product affordances work together to increase user engagement—positioning these companies as some of the most profitable businesses in the world. What is problematic, however, is that social media companies have created incentive structures that employ AI and algorithms, surveillance advertising, subscription models and product affordances to exploit people's predilection for clicking on incendiary content and sharing misinformation and divisive material.

Hate speech, conspiracy theories, and misinformation—amplified and recommended by algorithms—put corrosive and false content at the tops of personalized news feeds. Platforms benefit from the existence and spread of this content because it drives their engagement metrics by motivating users to spend as much time on the platform as possible, to increase the amount of data that can be extracted about users and, in turn, enable platforms to serve more and more targeted advertising to users—ultimately increasing revenue. In this way, social media is the most successful extraction industry the world has ever known. When critics say that the existence and viral amplification of hate content and disinformation is a feature, not a bug, of social media platforms, this is what they mean.

A. Surveillance Advertising and Political Advertising

Like other industries, social media platforms profit from delivering advertisements to users. Tech platforms are distinct from other advertising-based businesses, however, because of the specific

way these platforms collect data and target ads. As mentioned above, social media platforms are so successful because they collect and analyze troves of user data, based on user activity on the platforms, and across the internet. This user data is collected for two key purposes: first, to keep users engaged on platforms (e.g., viewing and interacting with content) for as long as possible, so users see as many advertisements as possible; and second, to deliver highly targeted advertisements to users based on what platforms know about each users' behaviors, habits, and preferences. Platforms use this data to develop highly specific advertiser-focused user segments. Then, algorithms deliver ads to specialized demographic segments through personalized content feeds.

While some user data are provided directly by users to platforms (e.g., age and location), social media companies also surveil users to gather extensive information from their profiles (e.g., friends/followers, contacts, connections, groups) as well as their online activity—both on the platform and across the internet. Platforms track "likes," shares, navigation paths, hover time, watch time, and other user engagement actions. Some platforms collect additional customer data from activities off the platform. This practice has been referred to as surveillance advertising: closely tracking and profiling individuals and groups in detail and then narrowly targeting ads at them based on behavioral history, relationships, and identity. Surveillance advertising allows platforms to dominate the digital advertising market by offering both big and small businesses an extremely efficient and effective form of advertising—far more than other options such as newspaper or local TV advertising.

One key problem with surveillance advertising is that dominant digital advertisers (namely, Facebook, which owns Instagram, and Google, which owns YouTube) curate the content each person sees on their platforms using the data collected. The goal of surveillance advertising is to keep users engaged, to serve them more ads and mine them for more data. Big Tech platforms amplify extremism, hate, and conspiracies because they know that this content generates the most engagement and, therefore, the most profit. As discussed in more detail below, platforms' algorithmic tools have significantly boosted extremist content, from white supremacist groups and Holocaust denialism to COVID-19 hoaxes and misinformation.

Surveillance advertising, which sometimes allows for microtargeting of demographic segments, can become even more problematic when used for political and "social issue" advertising. Political advertising often disseminates disinformation and fuels hate by narrowly targeting particular user segments and incensing them with outrageous, divisive content. For example, prior to Twitter banning political ads in October 2019, ADL Belfer Fellow Sam Woolley—an assistant professor at the University of Texas—conducted a study of computational propaganda, Jewish Americans, and the 2018 elections.

Woolley conducted interviews and analyzed Twitter data in order to understand both the scope of the issue on a national scale and the repercussions faced on the individual level. For the interviews, Woolley spoke to five Jewish Americans involved in politics as elected officials, policymakers, journalists, political consultants, and commentators. Woolley found that political

advertising on platforms was susceptible to being gamed by bots and taken advantage of by anonymous groups. Interviewees said tech companies seemed reluctant to remove bot-driven and harassing content and posited that the companies' reluctance came from not wanting to affect user growth—a metric used to determine company value. Alongside interviews, Woolley collected and analyzed 7,512,594 tweets related to U.S. politics from August 31, 2018 to September 17, 2018. The collected tweets showed the prevalence of political bots and highlighted political groups within the U.S. political spectrum most involved in antisemitic attacks.

Political advertising has also been considered a key source of misinformation, according to Laura Edelson, ADL Belfer Fellow and PhD candidate in computer science at New York University. Edelson and her team have specifically focused on how misinformation spreads on Facebook. Facebook has made promises to be transparent about all of the U.S. political ads on its platform—and about who paid for them. However, it routinely misidentifies political ads and also fails to disclose important information about them. Facebook does not have humans overseeing every ad that is published on the platform—even though ads must be submitted for review. Instead, the company uses a combination of artificial intelligence (AI) and machine learning (ML) models—and it also heavily relies on voluntary compliance, making it easy for bad actors to slip through enforcement gaps and also over-enforcing (and removing) legitimate ads.

Alarmingly, Edelson and her team have been able to demonstrate that extreme, unreliable news sources get more engagement on Facebook, and that the archive of political ads that Facebook makes available to researchers is missing more than 100,000 ads.

Edelson is currently working to measure misinformation and hate speech aimed at U.S. Spanish-speaking and Asian American communities by analyzing political advertising on Facebook from the platform's Ad Library and from CrowdTangle, a research and data collection tool. Notably, on August 3, 2021, after Edelson and her team started studies intended to determine whether Facebook was contributing to vaccine hesitancy and sowing distrust in elections—as well as trying to determine the role the platform may have played leading up to the January 6 insurrection—they were suspended by Facebook from accessing its data.

It's no surprise Facebook attempted to block Edelson's access to data seeking to uncover Facebook's role in the insurrection. According to reports, based on internal documents submitted to the SEC by the Facebook whistleblower, analysis of the January 6 insurrection illustrated that the company was fundamentally unprepared to manage the "Stop the Steal" movement, which turned violent and played a pivotal role in the insurrection. Facebook's internal analysis found that the policies and procedures put in place were not strong enough to prevent the growth of groups related to "Stop the Steal." The report noted that Facebook treated each piece of "Stop the Steal" content individually, rather than as part of a greater whole. The result of this decision was that only some "Stop the Steal" content or groups were taken off the platform but much of the content and many of the groups were left up and, ultimately, amplified by Facebook's own algorithms.

On September 28, 2021, Edelson testified before the House Science, Space, and Technology Committee's Investigations and Oversight Subcommittee. At the hearing, titled "The Disinformation Black Box: Researching Social Media Data," Edelson spoke about the harms caused by misinformation on social media and the difficulties researchers face in trying to study this threat to the public. Platforms like Facebook provide independent researchers little access to advertising data, so it is difficult to understand the full impact of political and "social issue" advertising. We need more transparency about Facebook and other platforms' data collection, ad targeting, and algorithmic systems.

*B. **AI and Algorithms***

AI and algorithms play a powerful role in the dissemination of extremism and online harm. As referenced in a report co-authored by ADL and other organizations fighting disinformation, "AI can be understood as machines that predict, automate, and optimize tasks in a manner that mimics human intelligence, while [machine learning] algorithms, a subset of AI, use statistics to identify patterns in data." Social media platforms use algorithms, largely fueled by AI and ML systems, to deliver, rank and moderate content, to determine what content should be recommended to a user, and to serve advertisements to users. Algorithms make these highly personalized decisions by collecting and synthesizing vast amounts of user data.

One primary reason algorithms amplify harmful online content on social media is that platforms optimize them for user engagement. They are tuned to keep eyeballs on the screen. Algorithms feed users tailored content, based on factors including browsing activity. When a user interacts with a piece of content, algorithmic systems take note of the user's behavior to find and recommend similar content to the user. For example, if someone watches a video about an election, algorithmic systems will recognize that the user may be interested in political content and will continue to recommend related content. If someone has viewed or searches for hateful content, algorithms learn to serve the same user similar or more extreme content.

In addition to personalized recommendations, algorithmic systems focus on what pieces of content are likely to attract a wide range of users. Algorithms do this by recognizing signals—including which pieces of content are forwarded, commented on, or replied to and then combine those signals to almost immediately, show that content to more users. These algorithms predict if the piece of content will increase engagement, and thus increase advertising revenue. ADL has reported on research that controversial, hateful, and polarizing information and misinformation are often more engaging than other types of content and, therefore, receive wider circulation. Platforms privilege this incendiary content, creating a stimulus–response loop. In fact, reports of a Facebook researcher who explored how the social media platforms deepened political divides illustrated the speed with which platform algorithms get to work to recommend content rife with misinformation and extremism—less than a week.

In a forthcoming peer reviewed study, Laura Edelson and a team of academic researchers consider how Facebook users interact with unreliable and partisan news sources. The team found that posts from sources known for misinformation are *six times* more likely to get engagement than factual ones. Notably, most of the misinforming content was generated by far-right sources.

In fact, in Edelson's findings, far-right engagement with misinformation made up 68 percent of total engagement. A much smaller share of publishers in other partisan categories were purveyors of misinformation. On the far right, 109 misinformation publishers accounted for almost *1.2 billion interactions*, which was more than twice the total engagement that the 154 non-misinformation news sources garnered. These findings confirm that a small number of misinformation publishers have outsized influence, generating far more interactions and audience reach than factual sources.

Extremist groups are empowered by the existence of powerful algorithms that amplify the hateful voices of a few to reach millions around the world. The persistent presence and amplification of hate, bigotry, and conspiracy theories on social media platforms has created an environment for extremism to flourish. This content, in turn, inspires individuals to commit acts of violence and domestic terrorism. While an individual who naturally engages in innocuous content (e.g., cat videos, makeup tutorials, or music videos) may not be pushed toward extremist content, individuals who engage with political content, seek to understand conspiracy theories, or have existing gender/racial resentment can quickly become trapped in a negative feedback loop.

In another example, exposure to videos from extremist or white supremacist channels on YouTube remains disturbingly common. In January 2021, Brendan Nyhan, an ADL Belfer Fellow and professor at Dartmouth College, published a report that collected comprehensive behavioral data measuring YouTube video and recommendation exposure among a diverse group of survey participants. Using browser history and activity data, the report examined exposure to extremist and white supremacist YouTube channels as well as to "alternative" channels that can serve as gateways to more extreme forms of content. Though some high-profile channels were taken down by YouTube before the study period, approximately one in ten participants viewed at least one video from an extremist channel (9.2%) and approximately two in ten (22.1%) viewed at least one video from an alternative channel. Moreover, the study found that when participants watch the videos, they were more likely to see and follow recommendations to similar videos. Consumption was concentrated among a highly engaged subset of respondents. Among those who watched at least one video of a given type, the mean numbers of videos watched were 64.2 (alternative) and 11.5 (extremist). Moreover, consumption of these videos was most frequent among people with negative racial views.

Algorithmic amplification of divisive and hateful content by Facebook, YouTube and other big tech platforms creates an environment prone to inspire those curious about extremism. ADL has reported on the clear connection between online antisemitic, extremist, and hateful images and tropes reverberating on social media and offline hate and violence directed at marginalized communities. In the United States, calls to violence in the name of white supremacy and "The Great Replacement" theory, which has proliferated online and been amplified through algorithms, correlate to fatal shootings in Poway, El Paso, Pittsburgh and more, and led to the injuries and deaths at the white supremacist attacks in Charlottesville in 2017 and on the United States Capitol on Jan 6, 2021.

*C. **Revenue Sharing and Monetization***

Revenue sharing models on social media platforms, such as subscription services and direct donations for livestreaming, allow extremist content creators to monetize the spread of hate. These revenue sharing models are designed for influencers and celebrities to earn income from the content they generate but are exploited by extremists and domestic terrorists as new sources of fundraising. Because mainstream platforms like YouTube sometimes attempt to remove violent extremist content, extremists also use niche platforms with permissive content policies, such as DLive, a video-sharing platform that makes financial transactions publicly visible, and BitChute, another video-sharing platform favored by extremists. Even if extremists are suspended from big tech platforms, they often promote their fringe channels on mainstream social media. For example, well-known antisemites E. Michael Jones (800,000 views) and Brother Nathanael (970,000 views) have been banned from YouTube, but actively promote their BitChute channels on Facebook.

ADL Belfer Fellow Dr. Megan Squire, professor of computer science at Elon University, researches monetization and de-platforming (that is, restricting or removing creators) among far-right extremists and domestic terrorists. In a 2021 study, Squire analyzed extremist monetization strategies on DLive. Squire found that a small number of "megadonors" disproportionately fund extremist content creators. These megadonors spend large amounts of money financing their favorite streamers and gain their own visibility and notoriety by doing so. Squire also analyzed content creators like Nicholas Fuentes, a well-known white supremacist who participated in the 2020 "Stop the Steal" campaign as well as the January 6 U.S. Capitol insurrection. Fuentes shrewdly optimizes his donations through his reliable livestreaming schedule.

These new forms of revenue sharing allow extremist content creators to monetize their propaganda, especially livestreamed audio content, which is more difficult to detect and remove quickly. On DLive, according to Dr. Squire's study, far-right actors can earn over $100,000 in donations in less than a year through a combination of megadonors and small donors. Extremist groups, such as the "Groypers" of America First and the Proud Boys, also earn money through revenue-sharing models. Platforms like DLive make it easy for creators to cash out funds, making it a reliable income stream for extremists.

*D. **Policies and Policy Enforcement***

As of 2021, almost every major social media platform has a stated public policy prohibiting extremism, terrorism, incitement-to-violence and hate on their platform. For instance, Facebook has a policy prohibiting dangerous individuals and organizations, while Twitter has a policy prohibiting violent organizations. The path to the creation and implementation of these policies, however, was not a direct one. Platforms are too often motivated not by harm prevention but, instead, by public perception. For example, despite repeated urging from ADL and civil society organizations to create a policy prohibiting white nationalism, Facebook only took action to implement a policy prohibiting white nationalist content following public outcry after the 2019 massacre of 51 Muslim people by a white supremacist in Christchurch, New Zealand.

In June 2020, after deep frustration with the PR-first focus of policymaking by tech platforms, a number of civil society organizations (ADL, Color of Change, Common Sense, Free Press, LULAC, Mozilla, NAACP, National Hispanic Media Coalition, Sleeping Giants) formed the Stop Hate for Profit Coalition. The coalition called on businesses who ordinarily advertise on Facebook to engage in a month-long advertising pause. Over 1,200 companies joined the July 2020 pause. Additionally, Stop Hate for Profit had a September 2020 week of action, which involved celebrities and influencers calling out hate and extremism on Facebook. Content from the September week of action had an estimated 1 billion views. In January 2021, the Stop Hate for Profit Coalition asked Facebook, Twitter, Google and other social media platforms to #BanTrumpSaveDemocracy by permanently removing Donald Trump from their platforms.

Changes long demanded by civil society around militia activity, the "boogaloo" movement, and Holocaust denial were finally made by Facebook following the Stop Hate for Profit Coalition's public pressure. The campaign's success clearly demonstrates the degree to which policymaking at social media companies is too frequently driven by public perception. Other platforms, also motivated by public pressure, took similar measures in the wake of Stop Hate for Profit. Twitter banned links to hateful content on their platform, which led to the deplatforming of noted white supremacist David Duke. Reddit released its first ever hate policy and deplatformed R/TheDonald, a forum of 800,000 users known to house hate and conspiracy theories. YouTube banned six prominent white supremacists, including Stefan Molyneux, David Duke, and Richard Spencer.

Social media companies' reactive practices of creating policies for public relations purposes in response to tragic events remained in full effect following the attack on the U.S. Capitol on January 6. Despite Twitter's July 2020 policy against content related to the hateful QAnon conspiracy, ADL was able to find numerous examples of QAnon on Twitter following the attack on the Capitol. It was only after increased public pressure—in light of the nexus between QAnon and the January 6 attack—that Twitter took more decisive action. After the insurrection, Twitter removed 70,000 QAnon accounts, which greatly reduced the spread of this hateful conspiracy theory on the platform. In fact, ADL found that immediately following the suspension of QAnon-related accounts, the use of QAnon-related hashtags plummeted by 73 percent.

The actions taken by tech companies—both to update their policies to better prohibit hate and extremism and to enforce their existing policies to remove such content from their platforms—were laudable. Ultimately, however, the fact that it took such intense public pressure for them to create policy and enforce improvements is unacceptable and, frankly, dangerous. When viewed through the lens of these companies, as primarily optimizing their business models and generating profit, these behaviors come into a much clearer focus. It's also why it is so clear that self-regulation will never work to solve this pernicious issue. What is needed is the establishment of a set of clear disincentives when platforms prioritize profit over people's safety.

*E. **Product Features***

Social media platform policies are only one part of the equation when it comes to mitigating online hate and extremism. Platform product features, like groups/pages, reporting and content moderation systems, often interact to create an environment ripe for extremists and domestic terrorists to exploit.

i. **Groups and Pages**

"Groups" is one Facebook product feature that may have had innocent origins, but for hate and extremist groups has been foundational to offline violence and domestic terrorism. Facebook claims that it is effectively addressing hate groups on its platforms. ADL and others, however, have continued to expose egregious examples of online hate, misinformation, and extremism across the company's products.

Facebook amplifying and recommending extremist groups like Boogaloo has led directly to offline violence. For example, in May 2020, Dave Patrick Underwood, a Federal Protective Services Officer, was killed in a drive-by shooting carried out by two Boogaloo adherents who were connected through Facebook and discussed the idea to commit the crime in a Facebook group. These assailants had never met prior to being connected on Facebook. This was one of many extremist-related shootouts ADL's COE tracked in 2020.

In June 2020, Facebook announced that it had taken down hundreds of groups and pages on its platform associated with the violent anti-government boogaloo movement, one of several major purges of extremist material by Facebook that year to address extremists' use of its platform. In recent months, however, several new boogaloo pages have emerged on Facebook, hiding among libertarian groups and pages that also share memes advocating for violence. One of the ways these groups have been able to remain on the platform is by using unconventional naming structures for their pages (such as "Char Broil Tru infrared grilling"). Though these new Facebook boogaloo groups typically are far smaller and produce less content than their predecessors did in 2019-20, the emergence of such pages highlights the need for Facebook to take a proactive stance to ensure that boogalooers do not successfully reestablish groups on the platform.

Perhaps most concerning, Facebook algorithms appear to be recommending these boogaloo pages and groups to like-minded users, despite the company's assertion last June that it would no longer do so. That assertion was followed by broader statements (in September 2020) that the platform would not recommend groups tied to violence, and an even broader March 2021 statement that Facebook would be ending all recommendations for "civic and political groups, as well as newly created groups." A recent review found that among groups sharing violent memes and a group simply named "Let's Overthrow the Government," Facebook was recommending groups with names like "The Hawaiian Hootenanny," "Boogaloonia," and "The Chaplain of the Redacted." In addition, after one boogaloo page was "liked," our investigation's user received suggestions of other pages with similar content, showing how opportunities are created for users to get further steeped in the ideology.

Clearly, Facebook's recommendation algorithms, filters and other detection methods for boogaloo groups and pages need upgrading. COE has found that even Facebook pages not directly associated with extremist groups are rife with violence. The public agrees: according to 2021 ADL data, 77 percent of Americans think laws need to be made to hold social media platforms accountable for recommending that users join extremist groups. And Facebook's own internal reports show that their recommendation systems are powerful ways to drive engagement and that small signals—even as small as a profile showing a woman in a southern state who liked Donald J. Trump and also Fox News, got recommendations for QAnon and other conspiracy groups within 48 hours of creating the profile, even with no other interactions on the site.

ii. Content Moderation and Reporting Systems

Today, most social media companies engage in content moderation to enforce content policies. These systems enforce the policies, sometimes called Community Guidelines or Terms of Service that determine what content, individuals, and groups are permitted on their services. Beyond having clear and comprehensive policies (which many platforms do not), platforms also communicate with their users about content management decisions. Users deserve to know that platforms will thoughtfully review their reports, especially when reporting hateful, racist or extremist content, and deserve timely and fair decisions from those systems. Generally, companies rely on a combination of human moderators and AI and ML-based tools to carry out their content moderation efforts, which include flagging, reviewing, and making determinations about content. Additionally, users report violative content to platforms. Importantly, across the industry, it is hard for users to trust that their reports are being addressed.

This year, ADL's Center for Technology and Society developed report cards on Holocaust denial and antisemitic platform content to determine the efficacy of platforms' reporting systems. Report cards have focused on a few different aspects of the reporting process. For ADL's Holocaust Denial Report Card, we first assessed a platform's response time, asking whether the platform investigated the report and promptly responded to the user. We did not include auto-generated messages to the user affirming receipt of a report. Instead, ADL considered whether any follow-up messages indicated that the platform investigated and made a content moderation decision. Second, ADL assessed whether the platform explained the reason for their decision, recording whether a user was notified about why a platform made a certain decision based on its stated policies. One noteworthy finding from this exercise is that platforms with explicit Holocaust denial policies did not necessarily do better enforcing those policies against our reported content, despite years of advocacy from civil society and researchers. Additionally, despite calls for greater transparency, another notable result is the opacity surrounding how platforms reported on the enforcement of their policies. The results of the investigation can be seen in the image below.

PLATFORM	EXPLICIT HOLOCAUST DENIAL POLICY?	GENERAL HATE POLICY?	EFFECTIVE PRODUCT LEVEL EFFORTS TO ADDRESS HOLOCAUST DENIAL?	RESPONSE WITHIN 24 HOURS?	NOTIFICATION OF POLICY REASON FOR ENFORCEMENT?	ACTION TAKEN AGAINST HOLOCAUST DENIAL?	GRADE
Twitch	Yes	Yes	Yes	Yes	No	Yes	B
Twitter	No	Yes	No	Yes	No	Yes	C
YouTube	Yes	Yes	Yes	No	No	No	C
TikTok	Yes	Yes	Yes	No	No	No	C
Roblox	Yes	Yes	Yes	No	No	No	C
Facebook (including Instagram)	Yes	Yes	No	Yes	No	No	D
Discord	No	Yes	No	Yes	No	No	D
Reddit	No	Yes	No	No	No	No	D
Steam	No	Yes	No	No	No	No	D

Note: In creating this framework for evaluating the efforts of digital social platforms, we weighted enforcement more heavily than policy and explicit policies more heavily than general policies. Additionally, because no platform had affirmative results in every category, we did not award any platform an "A."

Image: ADL Holocaust Denial Report Card

For ADL's Antisemitism Report Card, we established six categories to evaluate how well platforms responded to user reports. Four of these six categories focused on platforms' responses to reports from an ordinary user: Metric #1: Does the policy explicitly mention religion, race, or ethnicity? Metric #2: Did the platform respond within 24 to 72 hours? Metric #3: Was the user notified regarding whether the content they reported violated/did not violate a specific platform policy? Metric #4: Was the content removed or otherwise actioned as a result of the report? Two categories focused on platforms' responses to reports from trusted flaggers: Metric #5: Does the platform have a trusted flagger program? Metric #6: Did the platform take action on content reported through its trusted flagger program?

ADL investigators found that no platform performed above a B- in addressing antisemitic content reported to it. Also, no platform provided information or a policy rationale for why it did or did not remove flagged content. Users deserve more transparency and greater protection from platforms than companies are inclined to provide. Such reluctance has consequences in the form of economic, emotional, mental, political, and physical abuses that affect many people's lives, as repeatedly shown in ADL's research. It is irresponsible for platforms to take, at best, piecemeal approaches that do little to address the rapidity and depth of online hate and harassment. The results of the investigation can be seen in the image below.

Online Antisemitism Report Card

PLATFORMS	HATE POLICY THAT EXPLICITLY MENTIONS RACE, RELIGION, OR ETHNICITY?	RESPONSE WITHIN 24-72 HOURS?	NOTIFICATION OF POLICY REASON FOR ENFORCEMENT?	ACTION TAKEN AGAINST A/S?	TRUSTED FLAGGER PROGRAM?	ACTIONED UPON TRUSTED FLAGGER REPORT?	EFFECTIVE PRODUCT-LEVEL EFFORTS TO ADDRESS ANTISEMITISM?	DATA ACCESSIBILITY GRADE	TOTAL GRADE
Twitter	Yes	Yes	No	No	Yes	Yes	Yes	B	B-
YouTube	Yes	No	No	Yes	Yes	No	Yes	C	B-
Reddit	Yes	No	No	No	No	N/A	Yes	B	C
Twitch	Yes	Yes	No	Yes	No	N/A	No	C	C
TikTok	Yes	No	No	No	Yes	Yes	Yes	F	C-
Facebook (including Instagram)	Yes	No	No	No	Yes	No	No	D	C-
Discord	Yes	Yes	No	No	No	N/A	No	C	C-
Roblox	Yes	No	No	No	No	N/A	No	D	D-

Image: ADL Antisemitism Report Card

V. EXTREMISM ON SOCIAL MEDIA IS A DOMESTIC TERRORISM THREAT

Today, extremists are enmeshed in online communities where content designed to increase their propensity for hatred and violence often circulates freely. As noted above, extremist content boomerangs from fringe websites to mainstream platforms—in part because of social media's immense power, amplification of "engaging" content, and sophisticated recommendation algorithms. However, extremism and hate that start on social media do not always stay there.

A. Examining Social Media's Role in Extremist Massacres

ADL COE research fellow Joel Finkelstein conducted an in-depth examination of genocidal language and conspiracy theories pervasive on fringe platforms. Comparing the online behavior of the perpetrators of the Pittsburgh and Christchurch massacres (Robert Bowers and Brenton Tarrant) suggested that the two killers had similar ideological motivations and were subject to similar radicalization methods. Both killers announced on fringe platforms that they were about to commit violence and seemed to identify their fellow forum participants as community members who might share their propensity to commit violence. Both killers were consumed by the conspiracy of a "white genocide."

Gab and 8chan (now 8kun)—the go-to forums for Bowers and Tarrant, respectively—are rife with white supremacist, hateful, antisemitic bigotry. Examining Bowers' and Tarrant's online actions demonstrates how online propaganda can feed acts of violent terror. Additionally, violent terror can itself create online propaganda. In both Bowers' and Tarrant's cases, the shooters strongly signaled back to their fringe web communities as though they were including them as knowing co-conspirators to their criminal acts. In both cases, the participation of these fringe web communities was key to the scope, sensationalism, and ideological thrust of the act. Moreover, both shooters claimed the same twisted notion of "white genocide"—or the imminent destruction of the white race by Jews and people of color—as the motive behind their terrorist acts, suggesting a shared ideological motivation. In fringe online communities, many members indoctrinate other users based on the conspiracy propaganda of a "white genocide."

In Tarrant's 8chan manifesto, he actively and publicly sought to use his web community as co-conspirators and identified recruitment as a goal of his violence. He also told his community that he would be livestreaming his attack on Facebook. Notably, ADL's COE also reported on antisemitic, racist, and even violent content on what appeared to be Tarrant's Facebook profile.


ahem Anonymous 03/15/19 (Fri) 00:28:41 ID: c800e3 No. 12916717

Well lads, it is 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 livestream the attack via facebook
The Facebook link is below, by the time you read this I should be going live.
http://facebook.com/brenton.tarrant.9
It's been a long ride and despite all your rampant faggotry, fecklessness and degeneracy. you are all blokes
and the best bunch of cobbers a man could ask for.
I have provided links to my writings below, please do your part by spreading my messages, 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!

Image: Tarrant's post on 8chan about the New Zealand massacre showed his plans to use Facebook to livestream the attack and his desire to make his messages mainstream.

8chan was also the platform of choice for the extremists who carried out the murderous attacks in Poway and El Paso. White supremacist John Earnest, who allegedly opened fire inside a Chabad synagogue in Poway, California, killing one person and wounding three, also allegedly posted a manifesto to 8chan before his attack that admiringly referred to Tarrant and to Robert Bowers. White supremacist Patrick Crusius—who allegedly carried out a horrific attack that left 22 people dead and more injured at a Walmart in El Paso, Texas—is believed to have also posted a manifesto on 8chan. After the shooting, extremists discussed the attack across social media platforms like Twitter and Telegram.

Platforms like 8kun and Gab (which remains incredibly popular among extremists) force us to reassess our understanding of how violence may be inspired by hateful echo chambers. Even more broadly, as we have reported, mainstream platforms can push such individuals from an open community, such as Twitter, into fringe environments like Gab that foster acceptability of dangerous views. Relatedly, some studies have similarly demonstrated that ethnic hate expressed on social media can cause surges in on-the-ground hate crimes. The implications of this online-offline dynamic are highly concerning.

B. From Memes to Mobilization

White supremacists and extremists consistently co-opt innocuous or popular slogans by layering on exclusionary messaging and then using them as a call to action. They take advantage of cultural trends to infiltrate mainstream conversations and disguise their racist beliefs as irony or jest. Coded language, symbols, and narrative manipulation are key tactics white supremacists use to appeal to mainstream people, or "normies." For example, in March 2021 Chet Hanks (son of actor Tom Hanks) published a series of social media posts critiquing white men's attire and behavior, culminating a month later with the release of his song, "White Boy Summer" (WBS). A play on Megan Thee Stallion's 2019 hit song, "Hot Girl Summer," "White Boy Summer" took

the internet and meme culture by storm. While "White Boy Summer" was not initially intended to be hateful, white supremacists adopted the slogan and leveraged it for their own purposes. Some WBS memes included white supremacist/Neo-Nazi symbols, such as swastikas or references to the white supremacist 14-words slogan. Other WBS promoters, however, were tactically diluting their content disseminate their content widely across the internet and appeal to a broader audience. One Telegram account outlined this plan, encouraging users to strategically create content "without the use of Fascist/NS symbols so normies can get in the mood for white boy summer, and not get scared away."

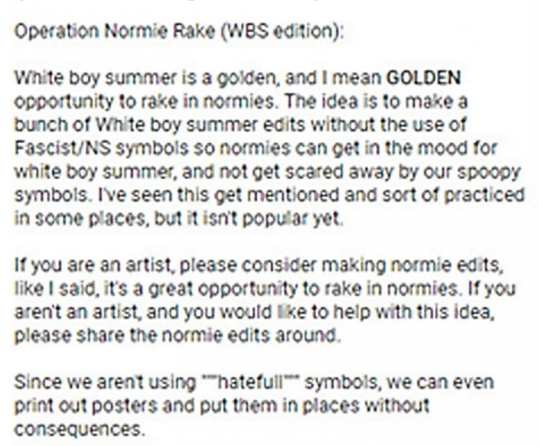


Image: Telegram post encouraging extremists to normalize posts to attract mainstream visibility.

Online users also invoked "White Boy Summer" to incite offline action. When the Derek Chauvin verdict was handed down on April 20, 2021, extremist posts encouraged people to engage in violence related to the guilty verdict, or "start celebrating #WhiteBoySummer a little early." In these cases, "White Boy Summer" was framed as a justifiable material response to "threats" against whiteness. ADL's COE tracked a variety of WBS-themed activity across the United States including merchandizing, physical propaganda distribution, extremist meetups, and proposed events. Online channels and influencers in the white supremacist movement also proposed "White Boy Summer" marches, rallies and road trips.

Social media has become a toxic ground for hosting, amplifying, and recommending corrosive content. Polarizing and bigoted language can become viral overnight, going from fringe discussions to Facebook and Twitter newsfeeds. The presence of hateful, racist and extremist content on social media thrusts bigoted ideas into the mainstream and normalizes otherwise extreme concepts and language. The leaked internal documents from Facebook have illustrated what we already knew—social media platforms cannot be trusted to make decisions that put people over profit. The result of their decisions: hate, mis- and disinformation, conspiracy theories, and extremist ideologies fester online and spread until they animate into acts of physical

violence. We cannot ignore the fact that extremism on social media is (or can be) a domestic terrorism threat.

VI. POLICY RECOMMENDATIONS

We need a whole-of-government approach to address the hate and extremism on social media—especially because it can fracture democracy and lead to offline acts of violence and domestic terrorism. ADL calls for urgent action to prevent and counter domestic violent extremism. Two frameworks that ADL has created — the REPAIR plan and the PROTECT plan— promote comprehensive strategies to mitigate the threat posed by social media's impact on domestic extremism and domestic terrorism while protecting civil rights and civil liberties. Together, these strategies can have an immediate and significant impact in stopping Big Tech from amplifying and fomenting extremism that leads to domestic terrorism. Our suggestions include:

A. ADL's Repair Plan

ADL has consistently stated that there is no single fix to the phenomenon of online hate. Whether it is in the dark corners of the internet, on the chats used by hundreds of millions of people on online multiplayer games, or a social media post that goes viral, the impact of online hate reverberates both on and offline—especially for those targeted by extremists, whom are disproportionately women and members of marginalized communities. ADL's REPAIR Plan presents an integrated agenda to fight hate online and push hate, gender-based violence and extremism back to the fringes of the digital world.

R Regulation and Reform

E Enforcement at Scale

P People Over Profit

A Access to Justice

I Interrupting Disinformation

R Research and Innovation

Congress has an important role in reducing the prevalence, impact and virality of online hate and extremism. Further, officials at all levels of government can use their bully pulpits to call for better enforcement of technology companies' policies.

Regulation and Reform

Platforms play a role in fomenting hate and violence, both by providing the means for transmitting hateful, violent, and abusive content—and, frequently, by more active enabling functions—in inciting violence, polarizing societies, spreading conspiracies, and facilitating discrimination, gender-based violence, and harassment. At the same time, tech companies are almost completely shielded from legal liability due to Section 230 of the Communications Decency Act (CDA 230) and the lack of other legislative or regulatory requirements—even where their products, actions, or omissions may aid and abet egregious civil rights abuses and

criminal activity. Because there are no third-party/independent audits of tech companies' internal systems, there is a complete lack of oversight and independent verification of the claims tech companies make, whether via Congressional testimony, in their transparency reports, or in related communications. In an absence of transparency and oversight, online spaces have been toxic for young women and a breeding ground for extremism.

- Congress must **effectively reform, not eliminate, CDA 230** to hold social media platforms accountable for their role in fomenting gender-based violence, extremist disinformation, and other forms of hate leading to harm—especially because of Big Tech's algorithmic amplification of dangerous content. Reform, however, must prioritize both civil rights and civil liberties concerns and not result in an overbroad suppression of free speech, nor unintentionally cement the monopolistic power of Big Tech by making it too costly for all but the largest platforms to ward off frivolous lawsuits and trolls.

- Section 230 reform should:
    - Stop immunizing platforms for algorithmic amplification of terrorism and discrimination. Tech companies are immune even when extremists/terrorists are recruited, radicalized, and/or are introduced to each other and plan acts of violence on their platforms. Senator Lujan's bill, the Projecting Americans from Dangerous Algorithms Act, and Reps. Eshoo and Malinowski's companion bill works to address this issue. The current broad interpretation of Section 230 means that plaintiffs alleging harm do not even get any discovery to determine what role the platform played in aiding or abetting the crime or unlawful conduct.
    - Stop immunizing tech companies from accountability for paid political and advertising content (e.g., where there is revenue-sharing or payment made from content creators to platforms). Tech companies are immune from accountability for their role in any harm caused even when they directly profit from platform-approved advertisements and other revenue-sharing agreements. Important components of Sens. Warner/Hirono's SAFE Tech Act address this issue.
    - More carefully differentiate between "conduct" and "content/speech" and eliminate immunization of the former**.** In 1996, when Section 230 was enacted, the internet was primarily text-based and noncommercial. Additionally social media platforms did not even exist. Today, however, people use the internet (and social media specifically) for far more than publishing speech. Platforms should not have wholesale immunity for everything that is produced online—especially information or conduct they create, amplify, control or profit from.
    - Ensure platforms take reasonable steps to prevent or address unlawful uses of their services. While reform cannot and should not be one-size-fits-all, every platform should do more to prevent or address unlawful content.

- Many tech policy experts have focused their efforts on reforming CDA 230 in pursuit of a non-existent one-stop solution. Importantly, reforming CDA 230 is only one essential step in a much larger process. CDA 230 reform will make platforms liable for certain unlawful third-party content; nevertheless, it is unlikely to have much impact on the "lawful but awful" hate that suffuses the internet and is often protected by the First Amendment in the United States. Therefore, policymakers must also pass laws and

undertake approaches that require regular reporting, increased transparency, and independent audits regarding content moderation, algorithms, and engagement features while looking for other incentive-based or regulatory action (e.g., potentially conditioning (narrower) Section 230 immunity on the steps platforms take to fight and mitigate egregious hate and disinformation).

- There is a strong need for **systematized, regulated, and easily accessible transparency** efforts from social media platforms. These platforms claim to have strong policies against hate, gender-based violence, and extremism, when in fact, most are unclear, hard to find, or have perplexing exceptions; enforcement is inequitable and inconsistent; and transparency reports are irregular and opaque.

- Additionally, Congress should encourage the Administration to **establish centers of expertise regarding online hate, gender-based violence, and severe harassment across agencies**. Within every agency, there should be cross-departmental task forces to help coordinate the work and support the necessary research, enforcement and plans of action. Agencies should work with Congress to develop research grant programs to comprehensively assess the links between Big Tech business models and online hate and build a more detailed knowledgebase of the industry role in online harms.

Enforcement at Scale

When something goes wrong on a major social media platform, tech companies blame scale and plead impotence. The fact that millions, even billions of pieces of content can be uploaded all over the world, shared, viewed, and commented upon by millions of viewers in a matter of seconds serves as the justification for "mistakes" in content moderation—even if those mistakes result in violence and death. But scale is not the problem here; defective policies, bad products, and subpar enforcement are the root of Big Tech's scale issue. Moreover, the ability of tech companies to comply with global privacy regulations after first arguing that scale made such compliance impossible is instructive. Equally or more significantly, when it comes to enforcement, too often platforms miss something completely, intentionally refrain from applying the rules for certain users (like elected officials), or have biased algorithms and human moderators who do not equitably apply community guidelines. Companies also make it difficult for users to effectively lodge complaints and receive redress. Indeed, existing business models make enforcement difficult, and instead, the Administration must empower and encourage "anti-hate by design" models of online product innovation.

- **Platforms need to develop a civil rights infrastructure,** so the companies mitigate harm to consumers through products, designs, algorithms, and policies that further discrimination, bias, and hate. Platforms should ensure that their design, user agreements, and policies counter the potential for bias-based discrimination and civil rights violations on the platform. To do this, platforms must regularly evaluate the way product features and policy enforcement fuel discrimination, bias, and hate and make product/policy improvements based on these evaluations. Platforms need an understanding of which populations are targeted or impacted most egregiously and why, the nature of hate content, and the path of spread; tech companies should create and maintain diverse teams

to mitigate bias when designing consumer products and services, drafting policies, and making content moderation decisions.

- Whole-of-government must exercise oversight by ensuring tech companies adopt and consistently enforce policies and community guidelines designed to identify and combat gender-based violence, hate, and harassment. While there is likely not a one-size-fits-all set of guidelines or enforcement given the force of law, incentives for effective standards and guidelines, transparency regarding them and their impact, and independent research evaluating these efforts can be imposed or supported by government. **The FTC, State AGs, and other enforcement authorities also should increase consumer protection efforts**, especially when tech companies engage in unfair and deceptive practices.

- We urge Government to consider basic consumer protection rules to product features like Facebook Groups that have amplified extremism, antisemitism, and misogyny; scaled racism and gender-based violence; and launched destructive conspiracy movements. As ADL's CEO Jonathan Greenblatt said in the Stanford Social Innovation Review, "If Mark Zuckerberg and his engineers can't improve Facebook Groups, we need to put it out to pasture permanently."

People Over Profit

The rapid and massive spread of extremism and hate on social media is a product feature, not a bug. Inflammatory mis- and disinformation and hate content generates growth and greater user engagement. Many tech company algorithms are wired to optimize for user engagement because the companies' business models are built around growing users and keeping people on the platform for as long as possible, to see as many ads as possible, which is what generates revenue. As many former and current Big Tech employees have acknowledged, platforms like Facebook build and employ algorithms designed to promote engagement, thus inevitably amplifying the most corrosive content.

- **Platforms need to adjust their algorithms** and stop recommending or otherwise amplifying organizations or content from groups associated with extremism, hate, misinformation, or conspiracies to users—even if it results in less engagement from users. Platforms must invest in both AI improvements and adequately trained and resourced human content moderators—with training focused on particular cultural contexts and languages.

- Platforms need to review and make adjustments to product features, like groups/pages, reporting, and content moderation systems, that exploit people's predilection to respond to outrage. They must consider processes that impose "friction" into product features to give users the opportunity to critically think about the content they share. Currently, split second sharing and virality are prioritized—this contributes to the amplification of highly problematic content.

- Platforms **also must put more resources toward protecting victims and targets of online harassment**, countering disinformation, and improving content moderation

instead of prioritizing the bottom line. Platforms should provide effective, expeditious resources and redress for victims of hate and harassment. For example, users should be allowed to flag multiple pieces of content within one report instead of creating a new report for each piece of content. They should be able to block multiple perpetrators of online harassment at once instead of undergoing the laborious process of blocking them individually. IP blocking, preventing users who repeatedly engage in hate and harassment from accessing a platform even if they create a new profile, helps protect victims.

- Transparency reports must evaluate success and provide evidence that independent researchers can use; such independent researchers must be granted access to data, and Congress must continue an oversight role. Companies can and should increase transparency related to their products. At present, technology companies have little to no transparency in terms of how they build, improve, and fix the products embedded into their platforms to address hate and harassment. In addition to transparency reports, technology companies should allow third-party audits of their work on content moderation on their platforms. Audits would also allow the public to verify that the company followed through on its stated actions and to assess the effectiveness of company efforts across time.

- We urge Congress and the administration to focus on how consumers—and advertisers—are impacted by a business model that optimizes for engagement. Congress must focus on how both algorithmic amplification and monopolistic power can fuel hate. **They should ensure algorithms are ethical and fair and consider regulating surveillance advertising and increasing data privacy**, so companies cannot exploit consumers' data for profit—a practice that inevitably results in greater online hate.

Access to Justice

A safer internet starts with protecting targets of harassment, not perpetrators. This means changing laws, policies, and practices that currently deny victims meaningful access to the courts and other effective avenues of redress. Victims of extremist violence, gender-based violence, hate, and harassment have no place to go in the face of physical threats, emotional injury, and financial and reputational harm when tech platforms host harassing content and enable perpetrators to abuse their targets. Victims and targets have been denied access to justice because our cyberharassment laws are outdated or don't exist at all.

According to ADL's latest data, 1 in 3 Americans who are harassed online attribute the harassment in whole or in part to their identity, referring to race, religion, gender, sexual orientation, gender identity, ethnicity, ability, and the like. More specifically, women experienced harassment disproportionately, as 35 percent of female-identified respondents felt they were targeted because of their gender. This abuse also happens in online games spaces. According to ADL's recent online gaming survey, exploring the social interactions, experiences, attitudes, and behaviors of online multiplayer gamers nationwide, for the third year in a row, gender was the most frequently cited reason for abuse.

Harassment intrudes into users' lives and hampers their ability to communicate, unfairly impacting marginalized communities' ability to work, socialize, learn, and express themselves online.

- We urge Government to provide more resources and pressure on agencies to pursue investigation and enforcement actions of bias-based cyberstalking, doxing, and swatting. Also, **Congress should update gaps and loopholes in cyber harassment laws** and the reporting of bias-based digital abuse in order to better protect victims and targets, including enacting legislation related to doxing, swatting, and non-consensual distribution of intimate imagery. One way to achieve this is by improving and passing the Online Safety Modernization Act at the federal level and focusing on passing anti-cyberharassment legislation at the state level.

- According to ADL's ethnographic study of online hate and harassment, "some of the most widely reported incidents of campaign harassment (the ability of harassers to use online networks to organize campaigns of hate) and networked harassment (the weaponization of a target's online network) have been waged against women and the LGBTQ+ community." Victims and targets of cyberhate need more resources and support. Congress and the Administration should work together to create a resource center to support targets of identity-based online harassment. This center could provide tools to victims and targets seeking to communicate with social media platforms, report unlawful behavior to law enforcement, and receive extra care. Additionally, creating a hotline for victims and targets of cyberhate and harassment and requiring the platforms to regularly report on the quantity and types of hate and harassment reported and actioned can help us tackle this issue.

Interrupting Disinformation

Hatemongers and extremists spread disinformation to harm targets and terrorize vulnerable communities; they amplify conspiracy theories to gain political aims; radicalize followers; and incite violence either intentionally as a tool to meet their goal or as a predictable outcome. Their content becomes further normalized when influential people, including high-level officeholders, spread this content further, often claiming that they are only "passing on" information they did not create for their followers to "evaluate." Hatemongers and extremists find ways to engage on mainstream social media platforms (Twitter, Facebook, YouTube), fringe platforms (Parler, Telegram, 4chan/8kun) and the Dark Web (Gab, DLive, america.win). It is a vicious cycle: this extraordinary spread is both made possible by, and helps further increase, the profound distrust of government and institutions.

The mainstreaming and normalization of hatful and extremist beliefs (including virulently misogynist, antisemitic, and racist conspiracy theories) is the foundation of much of the disinformation proliferating online. This is made evident by the fact that millions of Americans believe in QAnon conspiracies and other extremist ideologies.

Interrupting disinformation and finding/encouraging off-ramps and effective mitigation strategies to counter radicalization is no longer a marginal issue. It now requires a whole-of-government

and society approach. There is a clear connection between online extremist, antisemitic, misogynist, racist, and hateful images and tropes reverberating on social media and offline hate and violence directed at marginalized communities. Further, the deadly insurrection at the United States Capitol is a key example of the violence that can erupt when extremist disinformation spreads on social media.

- The continuing spread of baseless and dangerous conspiracy theories will continue to find fertile ground. Social media algorithms recommend content to extremist-leaning users, including related groups and pages that contain harmful content. Government must join with civil society and industry to find ways to undermine, interrupt, and mitigate disinformation without undermining civil rights and liberties. **Congress should fund research on the impact of social media platforms' recommendation systems and algorithmic amplification mechanisms** on the intersection between algorithmic amplification of disinformation, misogyny, and gender-based violence.

- Government must provide resources to civil society organizations working to counter online disinformation. **It must support widespread media literacy, digital literacy and anti-disinformation education.** Congress should investigate the nature and impact of product designs that allow hatemongers and extremists to exploit digital social platforms and spread antidemocratic, violent and hate-based disinformation and support concerted research to identify new ways of countering dangerous disinformation that leads to violence—especially gender-based violence. Government must not abuse this imperative to surveil vulnerable communities or to crack down on its non-violent critics and adversaries.

Research and Innovation

Government, civil society, and the tech sector must stay ahead of the curve as emerging threats will inevitably contribute to the impact of online hate. There must be a concerted effort to focus on technology research and innovation aimed at combating online hate. Just as privacy-by-design has been promoted, with some notable success, "anti-hate by design" must be promoted and widely incorporated into social media platforms and made a fundamental consumer expectation.

Government and platforms must focus on research and innovation to slow the spread of online hate, including, but not limited to: (1) measurement of online hate; (2) sexism, hate and extremism in online games; (3) methods of off-ramping vulnerable individuals who may be going down a path to commit extremist and gender-based violence; (4) the connection between online hate speech and hate crimes; (5) new methods of disinformation; (6) the role of internet infrastructure providers and online funding sources in supporting and facilitating the spread of hate and extremism; (7) the role of monopolistic power in spreading online hate; and (8) audio content moderation. States play a key role in this innovation, notably because our understanding of how hate impacts communities is most observable among those most familiar with their friends, neighbors, and others. Those community members are also individuals who have the most credibility in communicating with friends, family, etc. to prevent hate from taking root. States can invest in prevention, community engagement, and other tools to better understand how communities are dealing with the challenge.

B. ADL's Protect Plan

In response to the attack on the U.S. Capitol and in an effort to address the overall increase in domestic terrorism, while protecting civil liberties, ADL announced the PROTECT Plan. Domestic terrorism is a threat that impacts everyone.

P Prioritize Preventing and Countering Domestic Terrorism

R Resource According to the Threat

O Oppose Extremists in Government Service

T Take Public Health and Other Domestic Terrorism Prevention Measures

E End the Complicity of Social Media in Facilitating Extremism

C Create an Independent Clearinghouse for Online Extremist Content

T Target Foreign White Supremacist Terrorist Groups for Sanctions

Prioritize Preventing and Countering Domestic Terrorism

First, we urge Congress to adopt a whole-of-government and whole-of-society approach to preventing and countering domestic terrorism.

- In mid-June, the Biden-Harris Administration released the first-ever National Strategy to Counter Domestic Terrorism. The strategy is laudable, and a step in the right direction. However, many critical details were left unaddressed. Congress must press for further details into how the plan will be implemented, and the steps that will be taken to ensure protections for civil rights and civil liberties. Further, Departments and Agencies must create their own implementation plans for the Strategy. DHS can illuminate many of the implementation details of the Strategy by releasing its own plan. While we welcome the reinstatement of the domestic terrorism team within the Intelligence and Analysis (I&A) unit, additional initiatives and further details are needed.

- The Department of Homeland Security rightfully prioritized domestic violent extremism as a National Priority Area for the FY2021 Homeland Security Grant Program. We urge Congress to carefully oversee the effectiveness of these grants and continue the prioritization of the issue. Based on what is the most effective from this tranche of grants, the program should grow proportionate to the domestic extremist threat.

Resource According to the Threat

We must ensure that the authorities and resources the government uses to address violent threats are proportionate to the risk of lethality of those threats. In other words, allocation of resources must never be politicized, but rather, transparently based on objective security concerns.

- Congress should immediately pass the Domestic Terrorism Prevention Act (DTPA) to enhance the federal government's efforts to prevent domestic terrorism by formally authorizing offices to address domestic terrorism and requiring law enforcement agencies to regularly report on domestic terrorist threats. Congress must ensure that those offices have the resources they need and can deploy those resources in a manner proportionate to existing threats. Further, the transparency that comes with regular reporting is crucial for civil society, Congress, and the public at large to help oversee the national security process and hold leaders accountable.

- Congress must exercise careful oversight to ensure that no resources are expended on counterterrorism efforts targeting protected political speech or association. Investigations and other efforts to mitigate the threat should be data-driven and proportionate to the violent threat posed by violent extremist movements.

- The Department of Homeland Security can ensure it is resourcing proportionately by expanding data and transparency into how they see the threat and sharing with the public how the Department is aligning resources with the most lethal threats.

Oppose Extremists in Government Service

It is essential that we recognize the potential for harm when extremists gain positions of power, including in government, law enforcement, and the military.

- To the extent permitted by law and consistent with Constitutional protections, take steps to ensure that individuals engaged in violent extremist activity or associated with violent extremist movements, including violent white supremacist and unlawful militia movements, are not given security clearances or other sensitive law enforcement credentials. Appropriate steps must be taken to address any current employees, who, upon review, match these criteria. Law enforcement agencies nationwide should explore options for preventing extremists from being among their ranks.

- DHS announced that it will be vetting employees for extremist sympathies. ADL applauds this effort and welcomes any details on how the implementation of this vetting will take place, as well as any findings from the review.

Take Domestic Terrorism Prevention Measures

We must not wait until after someone has become an extremist or a terrorist attack has happened to act. Effective and promising prevention measures exist, which should be scaled.

- Congress can provide funding to civil society and academic programs that have expertise in addressing recruitment to extremist causes and radicalization, whether online or offline. By providing funding for prevention activities, including education, counseling, and off-ramping, Congress can help empower public health and civil society actors to

prevent and intervene in the radicalization process and undermine extremist narratives, particularly those that spread rapidly on the internet.

- These initiatives must be accompanied by an assurance of careful oversight and safeguards. They must also meaningfully engage communities who have been targeted by domestic terrorism and the civil society organizations embedded within them, and who have been unfairly targeted when prior anti-terrorism authorities have been misused and/or abused. They must be responsive to community concerns, publicly demonstrate careful oversight, and ensure that they do not stigmatize communities. Further, DHS should not be the only agency working on prevention; ADL urges the Department to partner with Health and Human Services and other non-security Departments whenever possible.

- While Congress has funded a small grant program for prevention measures domestically, the program is too small to have an impact at scale and, in some cases, DHS' implementation of the program has lost the confidence of communities. Now that the Administration has launched the Center for Prevention Programming and Partnerships, Congress should immediately authorize that office in statute and significantly scale its grant program; ADL has recommended a $150 million annual grant level.

End the Complicity of Social Media in Facilitating Extremism

Congress must prioritize countering online extremism and ensuring that perpetrators who engage in unlawful activity online can be held accountable. Online platforms often lack adequate policies to mitigate extremism and hate equitably and at scale. Federal and state laws and policies require significant updating to hold online platforms and individual perpetrators accountable for enabling hate, racism and extremist violence across the internet. In March 2021, ADL announced the REPAIR Plan, which offers a comprehensive framework for platforms and policymakers to take meaningful action to decrease online hate and extremism.

Create an Independent Clearinghouse for Online Extremist Content

Congress should work with the Biden-Harris Administration to create a publicly funded, independent nonprofit center to track online extremist threat information in real-time and make referrals to social media companies and law enforcement agencies when appropriate.

- This approach is needed because those empowered with law enforcement and intelligence capabilities must not be tasked with new investigative and other powers that could infringe upon civil liberties – for example, through broad internet surveillance. Scouring online sources through an independent organization will act as a buffer but will not prevent the nonprofit center from assisting law enforcement in cases where criminal behavior is suspected. This wall of separation, modeled in part on the National Center for Missing and Exploited Children (NCMEC), will help streamline national security tips and resources while preserving civil liberties. The current draft appropriations bills allocate

$500,000 toward a feasibility study for the Center; this appropriation is an excellent first step.

Target Foreign White Supremacist Terrorist Groups

Congress must recognize that white supremacist extremism is a major global threat of our era and mobilize with that mindset.

- To date, no white supremacist organization operating overseas has been designated as a Foreign Terrorist Organization. Only one has been designated as a Specially Designated Global Terrorist (SDGT). Congress should review how these designation decisions are made, whether any additional racially or ethnically motivated extremist groups outside the United States, particularly white supremacist groups, have reached the threshold for either designation, and whether such designations would help advance U.S. national interests.

- The Biden-Harris Administration must mobilize a multilateral effort to address the threat of white supremacy globally. Multilateral best practice institutions, such as the Global Counterterrorism Forum, the Global Community Engagement and Resilience Fund, and the International Institute for Justice and Rule of Law, may be helpful mechanisms through which to channel some efforts. Moreover, the Global Engagement Center should be charged with undermining the propaganda of violent extremist groups—not just designated terrorist organizations, but overseas white supremacist violent extremists as well. DHS should participate in these efforts, supporting overseas exchanges, partnerships, and best practices to engage in learning from other countries and sharing U.S. best practices, where applicable.

CONCLUSION

Thank you for the opportunity to testify before this body and for calling a hearing on this urgent topic. ADL data clearly and decisively illustrates that social media's business model directly correlates to hate rising across the United States, and fuels domestic extremism and terrorism, which continues to pose a grave threat. It is long past time to acknowledge these threats and to allocate our resources to address the threats accordingly. We must also address these threats holistically rather than piecemeal. This is precisely what ADL's **PROTECT and REPAIR** plans do, applying a whole-of-government and whole-of-society approach to push hate and extremism to the fringes of the digital world. On behalf of ADL, we look forward to working with you as you continue to devote your attention to this critical issue.

Testimony for "Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"

Senate Committee on Homeland Security and Governmental Affairs

Cathy O'Neil, CEO of ORCAA

October 28, 2021

Introduction

When we think about a user's experience on social media platforms, the raw ingredients are the pieces of content that people post, but it's the algorithms which decide who sees what and when (the "recommendation algorithm"). There are also separate algorithms (the "filter algorithms") which try to determine whether a given piece of content contains hate speech or is otherwise not allowed by policy.

The main goals of my testimony will be trying to explain, in very concrete terms, first what an algorithm is, second what a recommendation algorithm is and how recommendation algorithms lead people to become more extreme, and finally what a filter algorithm is and why we should be skeptical of their efficacy. I will conclude with an explanation of algorithmic audits and how they might be useful for understanding the harmful impact of social media platforms.

What is an algorithm?

How do you decide what to wear in the morning? You look in your closet at the available clothes and, depending on your memories and how you think your day will unfold, you decide what to wear. If you need to look professional, you'll choose a different outfit than if you're merely trying to be maximally comfortable.

You are using a "getting dressed" algorithm. Let me explain.

An algorithm needs two ingredients to do pattern matching: historical data and a definition of "success". The data can be simply memories stored in your head or digitized records stored on a computer. The definition of success can likewise be personal feelings of "being comfortable" or "being professional" or they can be mathematical formulas that computers can understand.

The algorithm sifts through the historical data provided to it, looks at examples of "success" as defined, and identifies patterns from the past that distinguished the successful cases from the

others. The algorithm learns from historical data what was successful in the past and predicts similar things to be successful in the future. In the examples of getting dressed, you could have a bad memory of wearing a particular pair of pants that ended up being uncomfortable, which would lead you to discard this choice (if the goal today is to be comfortable). If a computer is doing this pattern matching, the fancy math behind these techniques is able to find subtle, complicated patterns that people often can't. In fact, they sometimes find patterns that aren't even explainable or understandable to people.

Two important points: first, the definition of success really matters. If you want to look professional day after day, you'll end up wearing very different outfits than if you are optimizing for comfort. Second, whomever decides what "success" looks like for a powerful algorithm is actually wielding an enormous amount of power. Algorithms deployed by businesses are given a definition of success that primarily optimizes to profit. That might not be - and quite often, isn't - what is best for the rest of us; imagine an algorithm that tells us to wear uncomfortable pants every single day.

What is a recommendation engine?

Recommendation engines are a specific kind of algorithm focused on the task of making a personalized suggestion that will appeal to you -- say, a Facebook group to join, or a person to follow on twitter or Instagram. In the language of the first section, the Facebook newsfeed takes as historical data everything you've ever done on Facebook, and defines success as "keeping you on Facebook for as long as possible." Other social media platforms do very similar things, and for the same reason: the longer you are on their platform, the more you click on ads, which is how they make money. So they optimize for profit.

Behind the scenes, recommendation engines rely on hundreds of categories that represent real-life topics or interests. For instance, categories might include "baseball," or "crafts," or "cute animals," or "the stock market," or "politics." Each person gets a score for every category, and higher scores indicate a higher level of interest in that topic. So a person like me, who's a knitting fanatic but doesn't care about sports besides baseball have a high score for "crafts" and for "baseball" but a low score for "basketball" and "football."

Similarly, each piece of content gets a score in every category, representing its relevance to that topic. So a box score from last night's Red Sox game will have a high score for "baseball" and a low score for "crafts."

To a recommendation engine, each person and each piece of content is represented by a long list of scores. The algorithm then makes matches between people and content based on these lists. Basically, it serves up content whose scores closely mirror the user's scores.

These scores are constantly updated as people see and interact with (or ignore) content. If a user is shown the Red Sox box score and comments on it or shares it, their score for "baseball" will increase and their other scores will proportionately diminish. Similarly, if a knitting tutorial video gets shared and liked by many users with high "crafts" scores, then the video's "crafts" score will increase. In this way the user is "teaching" the algorithm what their scores are by every single action or inaction.

How do recommendation engines create rabbit holes?

Because people with high "baseball" scores are more likely to be served content with high "baseball" scores, they have more opportunities to push their "baseball" score *even higher* by interacting with that content. Similarly, since I love knitting already, any time a cashmere yarn store advertises to me on social media and I click on the irresistible ad, the algorithm will take note and set my "knitting" score higher, which means I'll be more likely to be shown knitting related content, especially ads, in the future.

This feedback loop gets supercharged when you introduce "viral" content that is outrageous or sensational. Whether it is happy or sad, inspiring or infuriating, the more sensational it is, the more likely it is to elicit a reaction from a user. Whether that reaction is a share/repost or a comment condemning the content, it counts as engagement and updates the scores, causing it to be shown even more. The more "viral" the content, the faster this happens. Facebook changed its algorithm to further boost viral content, and the result was more divisiveness and more extreme content.

Filter algorithms

Certain kinds of viral content -- the kinds that promote violence, or contain hate speech -- are exactly what is causing harm to individuals and society. Facebook understands this completely and has been spending the last few years trying to mitigate the harms by creating filter algorithms. Think of these as somewhat more sophisticated but similar to keyword searches you might run on your email account in order to find a specific email from a few weeks or months ago: you'd search by the name of the person who wrote to you and a word or two that was special in the conversation snippet that you remember.

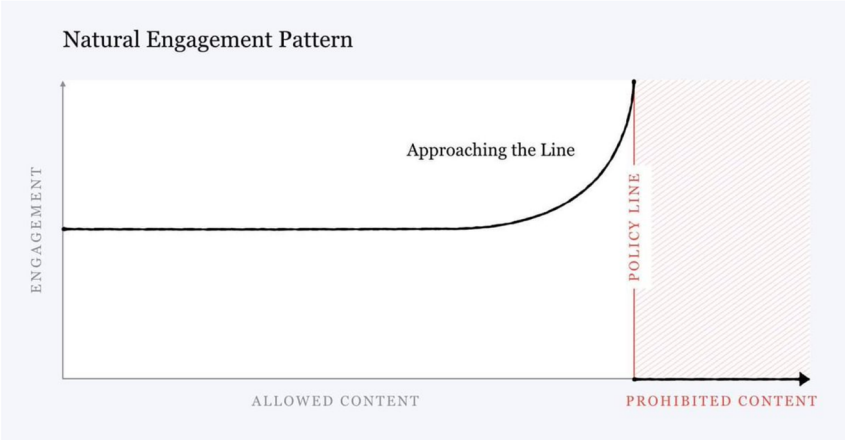
Similarly, filter algorithms look for keywords that are associated with hate speech, misinformation, or conspiracy theories. They are trained on historical pieces of content that have been labeled definitively prohibited in the past. So if someone posted the same thing again, it would get caught. If they posted something that's almost the same, it might get caught. But if

they posted something that had the same message but said in a novel way, it wouldn't get caught.

The problem for social media platforms is that there are people who get paid to bypass their filter algorithms with hateful content, or misinformation, or conspiracy theories. So it's an army of living humans, who are clever and strong willed, against an algorithm that cannot keep up.

Facebook's internal research shows its AI moderation tools successfully catch a mere 2 - 5% of all prohibited content. If you think of the filter algorithm as a net that is supposed to catch a certain type of fish, you should imagine that almost all of the fish manage to swim through the net without being threatened by the net at all.

Even in the context of 2-5% of content that is actively being scrutinised, the news is troubling. This graph from a 2018 letter by Mark Zuckerberg described the problem: content to the right of the "policy line" is bad and must be prohibited; but content just slightly to the left of the policy line is exactly what the platform wants, since it garners the most engagement.



Later in the same letter, Zuckerberg explains that Facebook will begin to identify "borderline" content and "downweight" it (i.e., show it less) so that this curve bends down instead of up. But even if they did that, it's still only going to apply to the small fraction of harmful content that they observe.

Conclusion: we need to audit these algorithms

At my company ORCAA, an algorithmic audit starts with the simple question: For whom could this fail? This centers the audit around stakeholders: the people whose lives could be impacted by the algorithm, for better or worse. Our audit process involves talking directly to stakeholders to understand their concerns.

In order to audit a social media platform's recommendation algorithm, then, we would need a complete list of stakeholders. But with billions of users, this is impossible. Nobody can anticipate all the specific subgroups of users that could have distinct concerns. I doubt even the most diligent auditor would have predicted that the Rohingya Muslims in Myanmar would need to be considered as a distinct stakeholder group.

If we can't audit the algorithm as a whole, what can we do? We can start by fixing some specific groups of stakeholders and focusing on particular harms. Then we can demand social media companies provide evidence -- or perhaps to give over raw data that allows regulators or other researchers to produce evidence -- that their products do not cause these harms to these stakeholders.

It's not enough for them to ignore the harm, or to research it in such a minimal way that they can deny it's really a problem when that research is leaked.

Questions you might want to ask me

- What is an algorithm?
- What types of algorithms are used by social media platforms?
- How algorithms push users down rabbit holes?
- What can be done?

Testimony of Professor Nathaniel Persily

James B. McClatchy Professor of Law
Co-Director of the Stanford Cyber Policy Center
Stanford Law School[1]

Before the United States Senate Committee on Homeland Security and Governmental Affairs on
"Social Media Platforms and the Amplification of Domestic Extremism and Other Harmful Content"

Submitted October 26, 2021

Thank you, Mr. Chairman and Members of the Committee, for inviting me today to testify on the role of social media in amplifying domestic extremism and harmful content. My name is Nate Persily. I am the James B. McClatchy Professor of Law at Stanford Law School and Co-Director of the Stanford Cyber Policy Center. Perhaps most notably for purposes of this hearing, I was also the cofounder of Social Science One, an effort to get internet platforms, such as Facebook, to share privacy protected data with outside researchers.

I want to use my testimony today to highlight what we know, what we need to know, and then what to do about harmful content and its relationship to social media platform policies.

But before I do that, let me begin my remarks by saying why we are here. We are here because Frances Haugen, the Facebook Whistleblower, produced thousands of pages of internal documents revealing internal research and communication at the company relating to harms Facebook investigated and steps they took or failed to take to combat them. It provided a rare glimpse of the internal workings of the company, and most impressively, the kind of data that informs the platform's assessment of harms and evaluation of potential interventions. Facebook knows an enormous amount about its users and its platform -- and most importantly, Facebook employees are the only ones with access to that information.

That equilibrium – where firm insiders know everything and the rest of us are left to guess – is unsustainable. **Facebook and the other Silicon Valley Platforms have lost their right to secrecy**. We need national transparency legislation that will allow researchers, other than those tied to the profit-maximizing mission of the firms, to get access to the data that will shed light on the most pressing questions related to the effects of social media on society.

[1] Affiliation for identification purposes only; appearing in personal capacity.

I. Harmful Speech Online: What We Know and What We Need to Know

Despite the inability to access data, researchers have learned a lot over the last decade about various online harms.[2] Of course, the harms attributed to social media have multiplied in recent years, as the platforms have been blamed for everything from human trafficking to anorexia to genocide. And as the Haugen revelations depict, even through this year, firm insiders had issued warnings about all of these issues and more.

The 2016 election represented a turning point in the way many people view social media. The disclosure by Facebook, itself, of the efforts of Russian and other foreign agents to meddle in the 2016 campaign quickly turned what had been a utopian view of the potential of social media for democracy to a dystopian one filled with amplified hate speech, disinformation, foreign election interference, and incitement to violence.

Around the same time, because of the Cambridge Analytica scandal, Facebook shut down some of the APIs and other pathways for outsiders to access its data. That scandal, as is well known, involved a researcher at Cambridge University who accessed social graph data and made it available to a political consulting firm. As a result, Facebook paid $5 billion pursuant to a consent decree with the FTC. The scandal casts a shadow over all academic efforts to access platform data.

In the years since the 2016 election and the Cambridge Analytica scandal, researchers have sought to answer the "big" questions about social media's effect on democracy.[3] I should emphasize at the outset that the study of social media's effects has been biased considerably toward the United States and Europe, despite the fact that the majority of users of platforms like Facebook now exist elsewhere. We have seen from the Haugen revelations, as well as plenty of earlier reporting, how unprepared Facebook has been in places like India, Myanmar, and Ethiopia. Especially in places where it does not have employees with the requisite language skills, Facebook often cannot enforce its community standards effectively or train classifiers to filter out problematic content. Similarly, although several important studies[4] have been published, we should be hesitant to generalize, particularly from the U.S. experience, as to how social media is affecting democracies around the world. Indeed, as "bad" as things might be here, where a disproportionate share of the "integrity" resources of the firms are directed, it is quite likely that the problems elsewhere are much worse, as the recently disclosed documents suggest.

---

[2] A disproportionate share of our knowledge is based off of Twitter data, because tweets are public and the platform has been the most open and welcoming of outside researchers. As a result, much of what we know about the internet and social media may be a bit warped by the unique affordances of Twitter, which is much smaller than Facebook or YouTube.

[3] See generally Nathaniel Persily & Joshua A.Tucker, eds., 2020. *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge University Press); Joshua A. Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal & Brendan Nyhan, *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature* (March 19, 2018), https://ssrn.com/abstract=3144139 or http://dx.doi.org/10.2139/ssrn.3144139.

[4] See, e.g., Payal Arora, The Next Billion Users: Digital Life Beyond the West (2019).

A. Amount of Content v. Rates of Exposure

The most difficult challenge in assessing the scale of harmful content online is that no one outside the firms knows how often users are exposed to such content. Whether the category is domestic extremism, terrorist content, election interference, incitement, disinformation or hate speech – those outside the firms do not know the true scale of the problem. Indeed, even those at the platforms have a myopic view because they can only see their slice of social media, and do not know about exposure through other online platforms, let alone cable and broadcast news, talk radio, and other legacy media.[5]

We know there is *a lot* of harmful content online, but there is a lot of all kinds of content online.[6] The key question is who produces, sees, and engages with this content? And how much does the average user, as well as large groups of users at the tails of the distribution, see such content?[7]

For most users of these platforms, their online lives are very similar to their offline lives. The types of people they talk to and the types of information they consume are similar to what they see on television and other media. For many people, in fact, their offline lives lead to exposure to more heterogenous content, given that our weak ties (e.g., high school friends and distant relatives) may be more politically diverse than the people in our immediate vicinity or the news sources we choose to read.[8]

Scholars now realize that, when assessing the scale of harmful online content, focusing on the average user leads to a warped assessment of the problem.[9] Although we

---

[5] Cf. Jennifer Allen et al., Evaluating the fake news problem at the scale of the information ecosystem, Science Advances, April 3, 2020, https://www.science.org/doi/pdf/10.1126/sciadv.aay3539.

[6] One issue relating to assessing the quantity of online content – harmful or otherwise – concerns the use of automation ("bots") in flooding the information ecosystem. See Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini & Filippo Menczer, The spread of low-credibility content by social bots, Nature Communications (2018), https://www.nature.com/articles/s41467-018-06930-7; Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer, Uncovering Coordinated Networks on Social Media: Methods and Case Studies, Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (2021), https://ojs.aaai.org/index.php/ICWSM/article/view/18075/17878 .

[7] For a sophisticated analysis that tries to grapple with both the "denominator problem" and the definition of hate speech, see Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, & Joshua A. Tucker. "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath." Quarterly Journal of Political Science 16, no. 1 (2021): 71-104.

[8] See Matthew Gentzkow & Jesse M. Shapiro, Ideological Segregation Online and Offline, Quarterly Journal of Economics (2011), Vol. 126, Issue 4, 1799-1839, https://academic.oup.com/qje/article-abstract/126/4/1799/1924154?redirectedFrom=fulltext ; Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, &Joshua A. Tucker. "How many people live in political bubbles on social media? Evidence from linked survey and Twitter data." Sage Open 9, no. 1 (2019): 2158244019832705.

[9] See generally Dimitar Nikolov, Alessandro Flammini & Filippo Menczer, Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. HKS Misinformation Review, 1(7). 2021. http://doi.org/10.37016/mr-2020-55.

all have heard stories about relatives who began by joining a knitting club on Facebook and ended up as QAnon adherents, that pathway is less frequented than conspiracy-curious users becoming incrementally more involved with communities defined by their conspiracy of choice or, even more ubiquitous, diehard adherents using the platforms as forums for strengthening their communities and providing a space for organizing.[10]

By all accounts, the producers and consumers of harmful speech – disinformation, hate speech, incitement, etc. – represent a small, but active, dedicated, and sometimes dangerous, share of users.[11] This appears to be true for election[12] and vaccine disinformation,[13] as well as hate speech or extremist racist content.[14] Even a small share of Facebook or YouTube users, though, can still equal millions of people.

This is also why it is difficult to study these problems from the outside. Studies of small shares of radicalized users require very large sample sizes. It is sometimes difficult to recruit people from these communities to be research subjects and random samples might not capture them. The platforms, however, know how large these communities are and the nature of the content they produce and consume. If outsiders had access to the same kinds of data disclosed in the studies from the Haugen revelations,

---

[10] Shruti Phadke, Mattia Samory, & Tanushree Mitra, What Makes People Join Conspiracy Communities?: Role of Social Factors in Conspiracy Engagement, Proceedings of the ACM on Human-Computer Interaction, Volume 4, Issue CSCW3, December 2020, https://dl.acm.org/doi/abs/10.1145/3432922; Shruti Phadke and Tanushree Mitra, Educators, Solicitors, Flamers, Motivators, Sympathizers: Characterizing Roles in Online Extremist Movements. Proc. ACM Meas. Anal. Comput. Syst. 37, 4, Article 111 (August 2018), https://doi.org/10.1145/1122445.1122456.

[11] See Shruti Phadke & Tanushree Mitra, Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups, http://faculty.washington.edu/tmitra/public/papers/hategroups-chi2020.pdf .

[12] One study of online election disinformation on Twitter in 2016, for example, finds exposure to be concentrated on a small slice of the population (1%), and sharing of that content occurs among an even tinier slice (0.1%). Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake news on Twitter during the 2016 US presidential election." Science 363, no. 6425 (2019), 374-378. See also Guess, Andrew, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science advances* 5, no. 1 (2019); Nicolas Berlinski, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2021. The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections. Journal of Experimental Political Science. https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/effects-of-unsubstantiated-claims-of-voter-fraud-on-confidence-in-elections/9B4CE6DF2F573955071948B9F649DF7A ("[E]xposure to claims of voter fraud reduces confidence in electoral integrity, though not support for democracy itself. . . . Worryingly, corrective messages from mainstream sources do not measurably reduce the damage these accusations inflict. These results suggest that unsubstantiated voter-fraud claims undermine confidence in elections, particularly when the claims are politically congenial, and that their effects cannot easily be mitigated by fact-checking.").

[13] Andrew M. Guess, Brendan Nyhan, Zachary O'Keeffe, and Jason Reifler. 2020. "The sources and correlates of exposure to vaccine-related (mis)information online." Vaccine 38(49): 7799-7805, https://www.sciencedirect.com/science/article/pii/S0264410X20313116 ; Francesco Pierri, Brea Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, John Bryden, The Impact of online misinformation on U.S. COVID-19 vaccinations, May 2021, https://arxiv.org/abs/2104.10635 .

[14] Annie Y. Chen et al., Exposure to Alternative and Extremist Content on YouTube (2020) at https://www.adl.org/media/15868/download.

we could all better understand the scope of the problem related to harmful speech online and what might be done about it.

B. Algorithmic Delivery v. Search, Subscription, and Selection

How users arrive at harmful content represents a critical question to which the literature is only beginning to provide an answer. Is the most problematic content served to users through algorithms or is it sought out by users interested in the content, who will often subscribe to channels, follow accounts, or become members of groups? The answer, of course, is that users arrive at such content through both paths – algorithmic curation based on past viewing habits, as well as through search, subscription and selection. However, understanding which is a more dominant pathway for radicalization is critical to understanding whether certain types of interventions, such as direct regulation of algorithms, will prove fruitful in reducing the reach of harmful content.

The Facebook Files present damning evidence of internal research describing how the Newsfeed algorithm was delivering harmful content to users around the world. Although Facebook took certain measures to reduce the reach of election-related disinformation, incitement, and hate speech, it still appears their efforts could not catch a large share of the election and COVID-related disinformation flowing over the platform. As several of the Facebook disclosures demonstrate, the machine learning classifiers developed to deal with a lot of harmful content are still in their infancy, and they are better at flagging some community standards violations (such as nudity) than others (such as hate speech).

Understanding the role algorithms play in amplifying harmful content is critical to evaluating the platforms' independent responsibility for the content they themselves are making popular.[15] It has become fashionable to describe Facebook, Google and Twitter as the "new public square." But the online spaces they control are very different than the Boston Commons or a street corner with a soap box. Based on the behavioral history of the user and predictions based on data gathered from everyone on the platform, the algorithms prioritize certain communication over others; they do not allow every speaker to have access to every willing listener at every given time. Although they are not publishers in the traditional sense, the platforms' decisions to prioritize some content over others -- and therefore give it "reach" -- creates greater responsibility than if the platform were merely hosting all comers or prioritized information on a first-come-first-served basis.

The experience with Facebook groups and the platform's recommendations of them demonstrates this dynamic. The most recent revelations tell of how internal researchers created fake accounts to demonstrate that Facebook's algorithm recommended groups and pages that would lead some users down rabbit holes of racist

[15] See Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer & Alessandro Flammini, How algorithmic popularity bias hinders or promotes quality, Scientific Reports (2018), https://www.nature.com/articles/s41598-018-34203-2.

extremism and conspiracy thinking. Based on these and other internal studies, Facebook banned political group recommendations in the period preceding the 2020 election. But overtly political groups are just the tip of the iceberg; the same dynamics have been seen with respect to anti-vax and QAnon content, as well as other kinds of conspiracies, both political and otherwise. As the disclosures and other independent research has confirmed, these groups have proven especially nimble in adapting to and circumventing the measures platforms use to take them down or demote them.[16]

Perhaps on no other issue is there such a gap between the contentions of the platforms and those of their critics (which includes both whistle blowers and conventional wisdom). Critics say the algorithms play an outsized role in surfacing extremist and sometimes dangerous content.[17] In particular, algorithms that myopically prioritize engagement will keep delivering to users the kind of content they engaged with previously that is likely to keep them on the platform. A user that is curious about conspiracies or other fringe content could become deeply enmeshed in a community dedicated to that issue, the argument goes, if the algorithm keeps recommending fringe pages, channels, or groups. Given that much incendiary and potentially harmful material may generate engagement because of its salacious or emotional appeal (rather than appeal to reason), the algorithm that prioritizes engagement necessarily feeds users more of the problematic content that they have signaled through their watch history would keep them on the platform.[18]

The platforms reject this characterization.[19] They point to measures they take to demote and takedown, rather than amplify, harmful content. They also maintain that they

---

[16] See generally the work of the Virality Project of the Stanford Internet Observatory at the Stanford Cyber Policy Center, noting the way anti-vax groups visually block key words, include disinformation in comments instead of posts, and share screenshots instead of links to problematic content. See Rachel Moran et al., Content Moderation Avoidance Strategies, July 29, 2021, https://www.viralityproject.org/rapid-response/content-moderation-avoidance-strategies-used-to-promote-vaccine-hesitant-content .

[17] See, e.g., Eslam Hussein, Prerna Juneja, & Tanushree Mitra, Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube, Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 48 (May 2020), https://doi.org/10.1145/3392854 .

[18] Much has been written trying to test these hypotheses, especially as they relate to YouTube. See Annie Y. Chen et al., Exposure to Alternative and Extremist Content on YouTube (2020) at https://www.adl.org/media/15868/download ; Manoel Horta Ribiero et al., "Auditing Radicalization Pathways on YouTube," 2019, https://arxiv.org/abs/7908.08373; Mark Ledwich and Anna Zaitsev, "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization," 2019, https://arxiv.org/abs/7912.11211; Kevin Munger and Joseph Phillips, "Right-Wing YouTube: A Supply and Demand Perspective," The International Journal of Press/Politics, October 21, 2020; Buntain, Cody et al., "YouTube Recommendations and Effects on Sharing Across Online Social Platforms," ArXiv:2003.00970 [Cs), July 20, 2020, https://arxiv.org/abs/2003.00970; Faddoul, Chaslot, and Farid, "A longitudinal analysis of YouTube's promotion of conspiracy videos."; Eslam Hussein, Prerna Juneja, and Tanushree Mitra, "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube," Proceedings of the ACM on Human-Computer Interaction 4, no. CSCW1 (May 28, 2020): 1-27; 24.Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, David M. Rothschild, Markus Mobius, and Duncan J. Watts, "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube," 2020, https://arxiv.org/pdf/2011.12843.pdf.

[19] See, e.g., Nick Clegg, You and the Algorithm: It Takes Two to Tango, Mar. 31, 2021 at https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2 ; The YouTube

abandoned a singular focus on raw engagement (which paradoxically led users to spend less time on the platform because engaging content was often low quality) and replaced it with measures of healthy engagement, "Meaningful Social Interactions," and "Valued Watch Time." Moreover, they contend that the recommendation algorithms direct people more often to mainstream news and information and, in the case of Facebook Newsfeed, more often toward content produced by friends and families.

Moreover, the existence of conspiracy-mongering, hate speech, and incitement on peer-to-peer messaging platforms complicates the story about the algorithm being the principal source of radicalization. Platforms such as WhatsApp (owned by Facebook) are known, particularly in the developing world, to be founts of the same problematic content found in the Facebook Newsfeed. [20] Yet, WhatsApp has no algorithm and no advertising. Rather, both grassroots and elite influencers use WhatsApp to incite violence, propagate disinformation, and promote hate speech – without the platform even knowing about it since the messages are encrypted and unseeable by the platform.

This debate between the platforms and their critics on the role of recommendations and algorithmic curation in the propagation of dangerous content can only be resolved with outside access to platform data and auditing of the algorithms. The Facebook Files reveal internal researchers' concerns consistent with those of the critics – namely, that the Newsfeed algorithm and group recommendations amplified COVID and election-related disinformation, content related to the January 6th insurrection, and hate speech in the U.S. and abroad. This debate over the algorithm is also a debate about social media itself – whether a platform can organize information based on user behavioral signals in a way that does not reinforce "unhealthy" choices previously made. Moreover, can they do so in a way that also does not also open them up to claims of censorship and shadow-banning based on ideological or partisan bias?

C. The Poorly Understood Role of Advertising in Propagation of Online Harms

If we array platform "affordances" along a continuum of responsibility, advertising would seem to be the area where platforms have the greatest responsibility for content delivered to users. Whereas the platform is understandably shielded (by section 230 of the CDA) from liability for user-generated content, when the platform takes money for content that it delivers in microtargeted fashion to users calculated to be persuaded by it, the platform's responsibility for that content is at its apex. Google and Facebook are online advertising monopolies, in the end, so their power over ads above all else requires the greatest scrutiny.

Team, Continuing our work to improve recommendations on YouTube, Jan. 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

[20] See Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, & Scott A. Hale, Tiplines to Combat Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian Election on WhatsApp, July 23, 2021, https://arxiv.org/abs/2106.04726; Punyajoy Saha, Binny Mathew, Kiran Garimella, Animesh Mukherjee, "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups, Feb. 7, 2021, https://arxiv.org/abs/2102.03870.

Nevertheless, we know precious little about the role that advertising plays in promoting hate, disinformation, incitement and other illegal activity online. We have examples, most notably from the 2016 election, of Russian agents buying ads on all major platforms to sow division and fan the flames of ethnic, racial, and religious hatred, as well as meddle in the election itself. Moreover, Facebook, at times, has relaxed some of its rules on disinformation and some community standards when it comes to political ads, not wanting to referee "truth in advertising" during an election campaign. However, one non-peer reviewed study found that when Facebook initiated a ban on advertising from fake news websites following the 2016 election that sharing of fake news decreased substantially.[21]

To understand the relationship of advertising to dissemination of harmful content, it is important to recognize the differences between online and legacy media advertising. It may be more appropriate to consider not advertising per se, but paid versus organic content and the interaction between the two. To be sure, we have many examples of inciting and dangerous content purchased through traditional advertising. Indeed, the audacity of the Russian intervention in 2016 can be seen in the exploitation of the traditional ad platforms to send simple, if polarizing, messages (sometimes purchased with rubles) to large numbers of users, as well as organize events and recruit followers.

Beyond traditional and familiar forms of advertising, though, amplification, itself, is for sale by the platforms. Traditional media entities realize this, which is why many of the most prominent legacy publications pay to boost their reporting on both Twitter, Google, and Facebook. However, these same opportunities are available to fringe groups, foreign actors, individuals, and truly fake news sites. The platforms identify content as promoted or sponsored (sometimes in small type), but it is very difficult to distinguish advertisements from organic content. Indeed, sometimes the posts can be identical -- the same story might arrive into a user's newsfeed because a friend has forwarded it or because the entity behind it paid to place it there. The distinctive feature about all paid content on modern social media platforms, though, is the microtargeting that the platforms make for sale. Therefore, not only can a publication (or bad actor) amplify content for a broader audience, they can also target a narrow one either by providing their own lists of individuals to create a custom audience or using the tools the platforms make available to target based on a myriad of characteristics.

Because Russian advertising efforts earned such infamy after 2016, the platforms developed political ad libraries to provide for greater transparency.[22] In addition, they adopted verification procedures for political ads (involving postcards sent to actual domestic addresses), to ensure election-related advertisements are not purchased by

---

[21] Lesley Chiou & Catherine Tucker, 2018. "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement," NBER Working Papers 25223, https://ideas.repec.org/p/nbr/nberwo/25223.html

[22] See Erika Franklin Fowler, Michael M. Franz, & Travis N. Ridout, Online Political Advertising in the United States, in Social Media and Democracy: The State of the Field and Prospects for Reform Ch. 6 (Persily, N. and J. Tucker eds., Cambridge University Press, 2020).

foreign actors. For the 2020 election, moreover, Twitter banned election ads and Facebook banned new ads for the last week of the campaign.

Outside efforts to evaluate the completeness of the ad libraries have run into problems, however. Most notoriously, Facebook this past summer suspended the accounts of researchers at NYU who were studying political ads.[23] The researchers developed a browser plug-in (the Ad Observer) that users could install to scrape the advertising information they were seeing when logged into Facebook and other sites. They discovered that many political ads never made it into the ad library[24] and that ads violating Facebook ad policies would find ways to avoid detection.[25] The NYU researchers remain suspended, even though the Federal Trade Commission took the extraordinary step of issuing a letter saying the plug-in would not run afoul of the FTC's privacy-related consent decree with Facebook. As it now stands, outsiders are severely handicapped in their efforts to evaluate the relationship of advertising and other boosted content to a variety of online harms.

II. What to do about it? Regulation and Transparency Relating to Harmful Content and Social Media

The fiasco involving the NYU Ad Observatory, as well as my own experience trying to facilitate outsider access to social media data with Social Science One, has convinced me that only federal legislation will open up these platforms to outside scrutiny. Researcher access to platform data (discussed in the Appendix) is only one aspect of the necessary transparency, and transparency is only one component of a larger legislative agenda relating to harmful online content. Transparency legislation may be the constitutionally safest form of regulation, though.

To level set, most direct regulation of harmful, but legal, online content would violate the First Amendment. With a few notable exceptions,[26] the regulation of hate speech, disinformation and (most forms of) incitement, cannot be done through outright bans or takedown mandates. For all the talk about repealing or amending Section 230 of the Communications Decency Act, no change in that law will get at most of the problems discussed here. It might expose the platforms (and potentially all platforms, including small startups trying to challenge established players) to liability for defamation and some examples where the connection between online speech and offline violence is most clear. But most forms of election or COVID-related disinformation, let alone hate speech

[23] Laura Edelson & Damon McCoy, We Research Misinformation on Facebook. It Just Disabled Our Accounts, NY Times, Aug. 10, 2021 (noting that their study suggested 100,000 ads that were not included in the ad library).
[24] See Tony Romm & Isaac Stanley-Becker, Tens of thousands of political ads on Facebook lacked key details about who paid for them, new report finds, Washington Post, Mar. 8, 2020.
[25] Jeff Horwitz, Political Groups Elude Facebook's Election Controls, Repost False Ads, Wall Street Journal, Nov. 1, 2020.
[26] Those exceptions might concern foreign agents intervening in elections, as well as some forms of child protection.

and incitement that does not satisfy the high standard the Court has set for direct advocacy of (likely) violence, cannot be regulated outright.

All of that said, there is still a lot that Congress can do. A case in point – certain forms of regulation of online advertising, in general, and microtargeting, in particular, should pass constitutional muster. Taxation of online advertising revenue might hit these firms the hardest. Even transparency (that is, compelled disclosure) as to purchaser identity, ad content, targeting, and exposure would help prevent the purchased amplification for harmful content. These measures should be applied to all ads, not just political ones.

Second, content neutral forms of regulation such as privacy and competition (antitrust) protections should not face high First Amendment hurdles. It might seem that these types of interventions, while necessary to serve other interests, are unrelated to the propagation of online harmful content. But restricting the kinds of data that large platforms collect about their users is one more way to inhibit microtargeting of ads and other content. Moreover, antitrust scrutiny, even if it falls short of breaking up these firms or treating them like quasi-public utilities, might lead to measures that allow for entry of new social media companies. In particular, if data portability and interoperability become part of the antitrust agenda,[27] the power of Google and Facebook over the information ecosystem (as well as the importance of their community standards and their enforcement) will be muted.[28]

Finally, the transparency agenda must be broad and multifaceted. It should include the kind of researcher access described in the Platform Transparency and Accountability Act below. But it should also require production of data in publicly available (but privacy-protected) interfaces that would allow journalists and the general public to understand in broad terms user exposure and engagement with content.[29] Similarly, algorithms for newsfeeds and recommendation engines must be subject to independent audits to unearth hidden biases and vulnerabilities. Algorithmic transparency might be too much to ask for, since these algorithms change almost daily and public disclosure of the algorithm may enable bad actors to game it. Moreover, there is almost no single person at the firms who understands the code, which has been built and rebuilt for the last twenty years or more. But just as we should be able to measure

---

[27] See Francis Fukuyama, Barak Richman, Ashish Goel, Roberta R. Katz, A. Douglas Melamed, Marietje Schaake, Report of the Working Group on Platform Scale 2020, https://pacscenter.stanford.edu/wp-content/uploads/2020/11/platform_scale_whitepaper_-cpc-pacs.pdf.

[28] It should be noted, however, that having a dozen smaller Facebooks and YouTubes is not necessarily an easier environment from a security perspective. Those two umbrella firms are able to spend untold sums on integrity and security because they are bundled together. So, Instagram and WhatsApp, for example, freeride off of the security teams at Facebook and would not have as developed security if they were to fend for themselves. Similarly, moves for portability and interoperability must reconcile with the privacy tradeoffs inherent in any system that either allows one to "take one's data" to a different platform or forces platforms to make their data treasure troves available to a set of outside competitors.

[29] I have in mind here richer versions of products like Google Trends and Crowdtangle, which would provide aggregated data that extends beyond mere engagement to include actual exposure.

emissions from a car's tailpipe, we should be able to evaluate exactly what the algorithms spit out under different conditions.

The revelations contained in the Facebook Files represent a turning point for our understanding of one of the most powerful corporations ever to exist. The picture they paint is one of an institution that is simply incapable of managing the technology that it has unleashed on the world. Even with the best of intentions and proper allocation of resources by these powerful firms, though, the "social" aspect of social media will inevitably create harms that elite actors cannot contain. Social media in an unhealthy and polarized society will reflect the underlying fissures that are tearing communities apart. This is why outsiders' understanding of what is happening online is so important. All stakeholders – civil society, governments, the firms themselves and users – have a role to play in counteracting the downsides and preserving the upside of this new technology. Only if some outside, independent entity untethered to the profit maximizing mission of the firm can regularly access and interpret the data revealed in these new disclosures can we come to grips with the measures that might be necessary prevent the harms to users and society that the Facebook Files have revealed.

A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act

Nathaniel Persily

The disclosures of whistleblower Frances Haugen have provided a unique glimpse into Facebook's internal research and the ways that the company evaluates and addresses different harms on the platform. As explosive as the content contained in Haugen's revelations may have been, most of the reaction may have arisen from the mere fact that outsiders got an opportunity to see what Facebook knows (or could know) about its users and the information ecosystem it controls. Every inadvertent disclosure that comes out of Facebook gains such notoriety because most of what the public normally sees is subjected to rigorous vetting, corporate-speak and spin.

We should not need to wait for whistleblowers to blow their whistles, however, before we can understand what is actually happening on these extremely powerful digital platforms. Congress needs to act immediately to ensure that a steady stream of rigorous research reaches the public on the most pressing issues concerning digital technology. No one trusts the representations made by the platforms themselves, though, given their conflict of interest and understandable caution in releasing information that might spook shareholders. We need to develop an unprecedented system of corporate data-sharing, mandated by government for independent research in the public interest.

This is easier said than done. Not only do the details matter, they are the only thing that matters. It is all well and good to call for "transparency" or "data sharing," as an uncountable number of academics have, but the way government might set up this unprecedented regime will determine whether it can serve the grandiose purposes tech critics hope it will.

As with so many areas of tech regulation, transparency laws come with tradeoffs. In some cases, for instance, transparency might inhibit necessary security or harm prevention measures, as public disclosures about platform standards' enforcement might lead to gamesmanship by bad actors. When it comes to data access for research, the chief risk that needs to be addressed is user privacy. The shadow of Cambridge Analytica is cast over any academic access to user data, as that scandal involved a university researcher mishandling user data for the benefit of a private political consulting firm. If user data cannot be protected, then the public will not have faith in any government-mandated data-sharing program.

It is critical to understand at the outset, though, that user data is already collected and analyzed – but only by employees at the firms themselves. The threshold question when it comes to outside researcher access is whether the firms (and their employees who are tied to their profit maximizing mission) should have a monopoly on the insights that access to such data guarantees. Perhaps the firms should be prevented from gathering so much user data, but once they do, the public needs to be aware of it and to benefit from the insights that independent analysis will provide.

These benefits will be substantial. Most importantly, the mere fact that outsiders will have access to platform data will affect platform policies and behavior. Digital platforms, like any other association, institution or individual, will alter their behavior if they know they are being watched. Second, researcher access will enable evaluation and auditing of platform rules and interventions to gauge the responsibility of firms for problems that occur on their platforms. In other words, researcher access can enable outside auditing of actions taken by platform against users and

doi:10.1017/xxxxxxx xxxx

content. Third, such access will inform policy makers seeking to regulate the platforms: only if they understand what is actually going on online might they be able to craft the appropriate regulations related to antitrust, privacy, advertising, child safety, content moderation or anything else. Finally, research on digital trace data is absolutely critical to understanding the sociology of the online information ecosystem, irrespective of potential links to policy. A large share of the human experience is taking place online. To understand it we need access to the relevant data.

The proposed legislation that follows – the Platform Transparency and Accountability Act – intends to design a data sharing program that protects user privacy to the extent possible while ensuring outside independent research on platform data. There are many ways to craft such a regime, and I hope this proposal sparks alternative approaches. The key features of any such system, though, must be (1) access by researchers not chosen by the firm to (2) the same data that the firms' own data analysts can analyze but (3) in a secure environment that minimizes any risks of disclosure of user private data.

Any proposal for outside access to platform data must wrestle with several questions (and this list is necessarily underinclusive). First, to which companies or platforms should such a regulatory regime apply? Second, who should have access? Third, to what data should they have access? Fourth and most important, how shall such access be regulated to protect both user privacy and research integrity?

1 Which Platforms?

Google and Facebook are first among (un)equals when it comes to the sheer volume of social media and digital trace data the firms possess. Any regulatory regime aimed at researcher access should be reverse engineered to capture those two firms in particular, as well as TikTok, which is quickly becoming a real competitor to YouTube. Twitter, which already provides more data than any other firm for researcher access, could also be added to the list, if the focus of the regulation is social media, per se.

But what about Amazon, Apple, and Microsoft? Researchers could gain enormous insight from access to those firms' data. Amazon, in particular, represents a monopoly of a different sort with data on users that could be extremely helpful to understanding the digital economy. Moreover, if the communications ecosystem is the target for research, what about the cable and cell phone companies, such as Comcast and Verizon? Surely, they possess data farther down the stack that could be helpful in assessing some relevant problems. A similar argument could be made for traditional media companies, e.g., Fox, or "new media" companies, such as Netflix.

To some extent, the universe of firms to which a data access regime would be applicable depends on the range of phenomena one considers worthy of study and the inability of researchers to gain insights from the outside. For those (like me) for whom the principal concern is the health of the information ecosystem and its impact on democracy, Google, Facebook, and Twitter reign supreme. The identification of the relevant firms, then, would include a definition of social media or search firms meeting some threshold of daily or monthly active users.

The Honest Ads Act[1] took a stab at such a definition in its attempt to force a disclosure regime on online political advertising. That bill defined an "online platform" as "any public-facing website, web application, or digital application (including a social network, ad network, or search engine) which...has 50,000,000 or more unique monthly United States visitors or users for a majority of months during the preceding 12 months." The legislative proposal that follows here lowers the bar to 40,000,000 monthly active users in order to capture TikTok as well.

1. https://www.scribd.com/document/409188376/Mcg-19321

2 Which Researchers?

Deciding which researchers shall have access is one of the biggest challenges to legislation in this area. "Researchers" come in many forms and a wide variety of civil society actors have an interest in the data held by internet platforms. However, some quality control must exist lest political operatives and propagandists repurpose themselves as "researchers" to gain access to platform data. It may also be that a separate regime for platform data access could be erected for think tanks or journalists, many of whom (such as Pew, ProPublica, the Markup, Buzzfeed or the Guardian) have done foundational research on these types of topics. Although categories such as journalists or think tanks may be difficult to cabin and enforce, transparency legislation should have as its goal making as much information available to as many watchdog groups, consistent with the privacy interests of users.

Focusing a data access regime on university-affiliated researchers has several advantages, however. First, a university is an identifiable "thing," and while low quality academic institutions exist, regulations can more easily specify the type of institutions that house the academics that should be granted access. Second, universities can be signatories to data access agreements with the platforms so as to add another layer of security (and retribution) against researcher malfeasance. Third, universities have Institutional Review Boards (IRBs) that can provide ethics and Human Subjects review for research proposals. Admittedly, IRBs have many well-known problems, but they are existing institutions that are in the business of evaluating research projects and the implications for human subjects. Fourth, in the wake of the Cambridge Analytica scandal, which involved an academic operating outside of his academic capacity, involving universities directly in the process of vetting and vouching for their researchers will make clear to the platforms which researchers are nested in a larger regulatory, contractual, and employment framework. Fifth, the National Science Foundation, which would play a role in vetting researchers, has established procedures in place to vet research projects and researchers from universities.

3 What Data?

In some settings, it is quite easy to define the data that should be made available for research. For instance, when drug trial data are made available for outside review, there are settled and familiar expectations for what kind of information the pharmaceutical company will provide. For Google and Facebook, though, the volume and variety of data they possess are so vast that any legally defined data access regime cannot simply say "turn over all available data to researchers." Some kind of principle should specify the range of data that should be available for research, or at least a process for deciding what data should be made available.

At a minimum, researchers should be allowed to analyze any data that is otherwise for sale to commercial entities or advertisers. If the datasets are available for a price, then they can be made available for academic analysis. Similarly, any data that goes into the preparation of government or other reports, such as those relating to enforcement of community standards (e.g., how many pieces of content were designated as hate speech and taken down) should be made available.

Beyond that, the key types of datasets that should be made available relate to "who" viewed/engaged with "what" content "when" and "how." In other words, to answer the most pressing questions relating to social media, we need data that can assess which types of people (though not individuals themselves) were seeing certain online content at certain times. The platforms already collect data of that nature. As part of the regulatory process, the platforms should be forced to identify datasets already in their possession, as well as data that are regularly collected. Then, the FTC, working with the NSF, should establish an application process for projects targeting those datasets. In addition, in order to prevent platforms from suddenly changing their data retention practices now that they are subject to oversight, the enforcement authority (here, the FTC) should have the

authority to require the production of datasets deemed reasonably necessary for providing answers to questions researchers ask.

Moreover, the FTC should require the platforms to produce the code necessary to describe how the data were gathered and assembled, and to describe the chain of custody of the dataset. Researchers need to understand how the platform came up with the dataset. The platforms should also be fined if they misrepresent the origins of the data or otherwise produce a dataset inconsistent with what was requested.

All such data must be anonymized or pseudonomized. Moreover, if it can be done without degrading the quality of research, technologies such as differential privacy or the construction of synthetic datasets should be encouraged. In other words, user data must be presented in a format that protects user privacy as much as possible while maintaining utility for the research project.

4 How Shall the Data be Analyzed While Protecting User Privacy?

One of the reasons that the legislative proposal presented here vests enforcement authority in the FTC is that the FTC has been on the frontlines of enforcing privacy promises (to the extent that it is authorized to do so). The consent decree with Facebook following the Cambridge Analytica scandal, for which Facebook was required to pay a $5 billion fine, was negotiated and enforced by the FTC. In an ideal world, the United States, like Europe, might have a cabinet level position that is responsible for digital services, but if any progress on researcher access is to be made in the next two years, it will need to work with existing agencies. The FTC, working with the National Science Foundation, is the logical choice. That agency, then, will be responsible for vetting researchers and research projects and specifying the conditions under which research shall be conducted.

Although the government will be heavily involved in enforcing the program of researcher access, the datasets themselves should never be placed in government hands. It is absolutely critical that there be no risk of government surveillance or privacy intrusions as a result of this program. Alternative models of access would place the datasets in a government-controlled researcher sandbox, which would allow the government to control directly the environment in which data are analyzed. Doing so would necessarily run the risk that at some point in the future, government officials would see this research environment as a honey pot for intelligence and law enforcement activities.

Under the proposal that follows, the data reside at the firm, which is responsible for maintaining security of the research environment and monitoring all research conducted therein. Researchers need to be monitored whenever they are in touch with the data. Every keystroke must be recorded as the data analysis is conducted. Researchers may not take any data out of the research environment without a privacy review being conducted. That includes immediately prior to publication – all publication drafts must be given a privacy review to ensure no data leakage. And in the event that a researcher engages in malfeasance both the researcher and the affiliated university shall be legally liable (even criminally liable) for any privacy violation. We need to make sure measures are in place that reassure the public that no individual's data is of interest to the research project, just the aggregated findings derived from them.

If the platform follows all applicable regulations concerning protecting privacy in the research environment, then it will be immune from suit for the fact that it made such data available under this program. To be clear, this does not immunize them from harms identified by the researchers. If the platform is discovered to be acting fraudulently or contributing to offline harm, then that information might later end up in a lawsuit or even a criminal prosecution. The point about legal immunity here is that the platforms cannot simultaneously be forced by the law to provide data to researchers and then be subject, for example, to a state tort law claim for violations of privacy.

5 Conclusion

Researcher access is only one component of transparency regulation, and transparency legislation is only one component of tech regulation. Nothing in this proposal should be seen as preventing broader reporting obligations for the platforms or construction of public facing APIs. Indeed, we should strive for a system in which any data on issues of public concern relating to the online information ecosystem should be available to the public, if it can be done in a privacy-protective way without other security risks.

One provision in the proposed legislation goes in that direction by dealing with the problem of scraping data from public-facing platforms. It would shield researchers from criminal or civil liability for scraping of public data from large platforms, like Facebook and YouTube. Of course, people disagree about what data, in fact, are "public" on these platforms. However, for researchers who scrape, they cannot be subject to money damages or criminal liability. This would not solve the problem faced by the NYU Ad Observatory, which had its accounts taken down by Facebook since it promoted a plug-in that allowed users to scrape their Facebook. But it would shield them from further actions, such as lawsuits that the platforms might initiate to get damages for terms of service violations arising from scraping.

A similar impulse underlies "Aaron's Law"[2] introduced by Representative Zoe Lofgren and Senator Ron Wyden. In a now famous and tragic episode, Aaron Swartz downloaded a large number of articles from the digital repository, JSTOR. In doing so, he breached the applicable terms of service for the website. Swartz was later arrested and prosecuted under the CFAA, which could have led to a penalty of 35 years in prison and up to $1 million in fines. However, he committed suicide before he was brought to trial. Aaron's Law would remove the threat of a felony prosecution for breaching terms of service in actions like this, if they do not cause significant economic or physical damage.

Just as researcher access is not coterminous with transparency, transparency does not address all problems that tech regulation seeks to solve. Nothing in this proposal should be seen as taking the place of proposals to address competition and antitrust, child safety, advertising, content moderation, cybersecurity and privacy. Indeed, a proposal like the one that follows should be bundled together with federal privacy legislation or other broad regulations of the tech industry.

Researcher access, however, is a condition precedent to effective tech regulation. Right now, we do not know what we do not know. There are fundamental inconsistencies between platform's public representations and those made by whistleblowers, let alone those that feed conventional wisdom. For example, on the critical question of whether algorithms and recommendation systems are leading users toward extremism or promoting disinformation, the defenders and critics of platforms fundamentally degree on basic facts. Policy makers need and deserve answers to these kinds of questions. Only if the government develops and mandates outside researcher access might we be able to get the answers necessary to make effective policy. Otherwise, we will be left with whatever studies the platforms choose to release or whatever research whistleblowers take with them on the way out the door.

Author

Nathaniel Persily is the James B. McClatchy Professor of Law at Stanford University and Co-director of the Stanford Cyber Policy Center.

2. https://www.congress.gov/bill/113th-congress/senate-bill/1196

Conflict of Interest

Not applicable.

Nate Persily
(in private capacity)

An Act

To support research about the impact of digital communication platforms on society by providing privacy-protected, secure pathways for independent research on data held by large internet companies.

SEC. 1. Short Title

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, that this Act may be cited as the "Platform Transparency and Accountability Act."

SEC. 2. Congressional Findings and Purpose

(a) The Congress finds that certain of the Nation's largest internet platforms exert unprecedented control over the speech marketplace.

(b) Exploitation of the affordances of these platforms has threatened the safety and integrity of our electoral processes, has increased our vulnerability to propaganda attacks by hostile nation-states and domestic extremists, has led to promotion of off-line violence, and has misled the public as to critical facts necessary to promote public health and well-being.

(c) Because of the unprecedented control these platforms exercise over massive amounts of user data and the speech marketplace, Congress finds it necessary to promote independent research on those platforms in order to reveal and help address societal and individual harms caused or exacerbated by these new technologies.

(d) The Congress declares it to be its purpose and policy, through the exercise of its powers to regulate commerce among the several States and with foreign nations and to provide for the general welfare, to assure so far as possible free and fair elections in this Nation so as to preserve our republican form of government, guard against foreign propaganda, and ensure the free flow of information in interstate commerce –

(1) by providing that Qualified Platforms and Qualified Researchers have separate but dependent responsibilities, interests, and rights with respect to obtaining data and information for Qualified Research Projects that will benefit the public good while protecting the privacy rights of the individual user;

(2) by authorizing the Federal Trade Commission to set mandatory data and information sharing requirements applicable to Qualified Platforms affecting interstate commerce, and by creating a Platform Transparency and Accountability Division for carrying out adjudicatory functions under the Act;

(3) by providing the groundwork for understanding the prevalence and character of disinformation, hate speech, and harmful and illegal content spreading by way of large Qualified Platforms, as well as potential political bias in content moderation practices and in algorithmic prioritization of content on those platforms;

(4) by investigating the exploitation of platform affordances by domestic and foreign actors seeking to undermine United States democracy and confidence in the election infrastructure;

(5) to ensure that the market power of certain Qualified Platforms does not pose anticompetitive effects that restrain the information economy; and

(6) to help inform policy makers and regulatory agencies by promoting an accurate understanding of the practices of Qualified Platforms and the dynamics of social media.

SEC. 3. Definitions

For the purposes of this Act --

(1) The term "commerce" means trade, traffic, commerce, transportation, transmission, or communication among the several States, or between any foreign country and any State, or between any State and any place outside thereof.

(2) The term "Commission" or "FTC" means the Federal Trade Commission established under the FTC Act.

(3) The term "Chair" means the Chair of the Federal Trade Commission.

(4) The term "Director" means the Director of the Platform Transparency and Accountability Division appointed by the Chair of the Federal Trade Commission.

(5) The term "Division" means the Platform Transparency and Accountability Division within the Federal Trade Commission.

(6) The term "Personal Information" means any information that is reasonably capable of being associated with a particular individual.

(7) The term "Qualified Platform" means a large, consumer-facing, online or internet-accessible business that meets the criteria for the same established by the Division or its appropriate delegate. Qualified Platforms must have over forty million active monthly Users of their service in the United States and shall include, but are not limited to, any provider of a large online platform, including an online social media service, which, at the request of a recipient of the service, stores and disseminates information to the public.

(8) The term "Qualified Data and Information" means information from a Qualified Platform that meets the criteria for the same established by the Division or its appropriate delegate. Qualified Data and Information may include information about User exposure, engagement, and other behaviors; data about content producers and content production policies; information that the Qualified Platform otherwise makes available for sale to commercial entities or advertisers; information that goes into the preparation of reports that Qualified Platforms provide to the government or other entities, such as those relating to enforcement of community standards; and metadata related to any of the preceding categories.

(9) The term "Qualified Researcher" means a university-affiliated researcher conducting research according to a research plan that has been approved by the Division or its appropriate delegate. No employee of a state of federal law enforcement agency or any government employee except for a university-affiliated researcher shall be considered a Qualified Researcher.

(10) The term "Qualified Research Project" means a research plan that has been approved by the Division or its appropriate delegate.

(11) The term "State" includes any state within the United States, as well as the District of Columbia, Puerto Rico, the Virgin Islands, American Samoa, and Guam.

(12) The term "Scrape" or "Scraping" refers to the act of electronically collecting data that Platforms make available via the user interface, either manually or through an automated process.

(13) The term "User" means a person or entity that uses a social media platform or online marketplace for any purpose, including advertisers and sellers, regardless of whether that person has an account or is otherwise registered with the platform.

SEC. 4. Obligations and Immunity for Qualified Platforms

(a) Each Qualified Platform shall comply with applicable federal, state, and local information sharing and privacy laws and regulations as well as all rules, standards, regulations, and orders issued by the FTC pursuant to this Act which are applicable to their own actions and conduct.

(b) In order to meet its obligations under this Act, a Qualified Platform must provide reasonable privacy and cybersecurity safeguards for the Qualified Data and Information that the Platform shares with Qualified Researchers. Such safeguards, at minimum, shall include

(1) encryption of the data in transit and at rest;

(2) delivery of data in a format determined by the Division that is not reasonably capable of being associated or linked with a particular individual;

(3) use and monitoring of a secure environment to facilitate delivery of the Qualified Data and Information to Qualified Researchers while protecting against unauthorized use of such data;

(4) evaluation by the Qualified Platform of any results garnered by Qualified Researchers before submission for publication but only to prevent public release of Personal Information or other violations of law.

(c) No cause of action under state or federal law relating to or arising solely from the release of data to Qualified Researchers may be brought against any Qualified Platform that complies with this Act and the privacy and cybersecurity provisions described herein.

(d) The legal immunity provided by subsection (c) shall extend only to the fact that data was made accessible to outside researchers and shall not extend to liabilities arising from findings discovered as a result of such research.

SEC. 5. Obligations and Immunity for Qualified Researchers

(a) Qualified Researchers shall be actively engaged in conducting research under a research plan, which was approved by the Division or its appropriate delegate based on its assessment of (1) the intellectual merit of the project (i.e. its potential to advance understanding the impact of

large digital communication platforms on society); and (2) its broader impacts (i.e. the project's benefit to society).

(b) Each Qualified Researcher shall comply with

(1) applicable federal, state, and local information sharing and privacy laws and regulations as well as all rules, standards, regulations, and orders issued by the FTC pursuant to this Act which are applicable to their own actions and conduct; and

(2) a prohibition on any attempt to reidentify, access, or publish Personal Information based on Qualified Data and Information that a Qualified Researcher may receive.

(c) No cause of action arising solely from Qualified Researchers' access and use of Qualified Data and Information may be brought against Qualified Researchers who conduct Qualified Research Projects in compliance with this Act and abide by all information sharing and privacy standards described in (a). This immunity includes immunity from potential liability under applicable federal, state, and local laws, as well as any potential liability for a violation of a Platform's Terms of Service that arises solely from the Qualified Researchers' access and use of Qualified Data and Information.

SEC. 6. Sharing of Qualified Data and Information by Qualified Platforms

(a) The Commission shall prescribe regulations requiring that Qualified Platforms maintain and provide Qualified Researchers access to Qualified Data and Information and accurate records of Users' interactions with or exposure to Qualified Data and Information.

(b) Qualified Platforms will be required to provide Qualified Researchers access to Qualified Data and Information as prescribed by regulation;

(c) Qualified Platforms will be required to provide a data codebook that outlines the structure, contents, and layout of the data and must provide the Qualified Researchers with methodological details on how data were collected, cleaned, or manipulated.

(d) Qualified Platforms must enable Qualified Researchers to preserve access to Qualified Data and Information as necessary to carry out and replicate Qualified Research Projects.

(e) The Commission shall also issue regulations requiring that Qualified Platforms, through posting of notices or other appropriate means, keep Users informed of their privacy protections and the information that the Qualified Platform is required to share with Qualified Researchers under this Act.

(f) Any Qualified Data and Information obtained or provided under this Act shall be obtained with a minimum burden upon the Qualified Platform, although Qualified Platforms may not shift the cost of their compliance with this Act to the Qualified Researchers. Unnecessary duplication of efforts in obtaining information shall be reduced to the extent feasible.

(g) Qualified Researchers are authorized to compile and analyze Qualified Data and Information under this section and are required to make all reports created from that analysis freely available to the public in both summary and detailed form.

(h) Twenty (20) business days prior to public release of an analysis by a Qualified Researcher based on Qualified Data and Information or at a time designated by the Division, the Qualified Researcher shall submit a pre-publication version of their research to the Qualified Platforms and the Division for evaluation to confirm that the analysis does not expose Personal Information.

(1) Qualified Platforms may object to the publication or release of any analysis that will necessarily expose Personal Information or otherwise violate federal, state, and local information sharing and privacy laws and regulations or any applicable rules, standards, regulations, and orders issued by the FTC. Such objections must be made in writing to the Division or its delegate within ten (10) business days of the date that the Qualified Researcher submitted the pre-publication version of the research or at a time specified by the Division. Such objections shall include proposed changes to the publication to address the legal problems identified.

(2) If no objection is timely made by a Qualified Platform or the Division, the research may be published.

(3) If objection is timely made by a Qualified Platform, the Qualified Researcher will have ten (10) business days to modify the publication and re-submit it to the Division, which shall decide within ten (10) business days whether the publication complies with the Act. If the Division finds that the publication does not comply with the Act, the Qualified Researcher may appeal such finding to the U.S. Court of Appeals for the Federal Circuit.

(i) The Commission shall have the authority to make, amend, and rescind, in the manner prescribed by 5 U.S.C. § 553, such rules and regulations as it may deem necessary to carry out its responsibilities under this Act.

(j) Access to Qualified Data and Information shall not be granted to any Qualified Research Project pursuant to this statute if it has not been approved or deemed exempt by an Institutional Review Board at the researcher's affiliated university.

SEC. 7. Scraping of Data from Qualified Platforms for University-Affiliated Research

(a) Any university-affiliated researcher conducting research that has been approved or deemed exempt by an Institutional Review Board at the researcher's affiliated university shall be immune from civil or criminal liability for the scraping of data made available through the user interface on Qualified Platforms, regardless of whether the researcher is a Qualified Researcher or the research project is a Qualified Research Project.

(b) Researchers who meet the requirements in (a) are not required to notify Qualified Platforms about Scraping practices, and no cause of action may be brought against Qualified Researchers arising from Scraping practices that comply with this section and such conduct shall be deemed authorized conduct for purposes of the Computer Fraud and Abuse Act (CFAA), 18 U.S.C. 1030.

SEC. 8. Platform Transparency and Accountability Division

(a) It is the purpose of this section to enable and approve Qualified Researchers to carry out the types of Qualified Research Projects set forth in this Act.

(b) There is hereby established within the Commission a Platform Transparency and Accountability Division. The Division shall be headed by a Director who shall be appointed by the Commission Chair, with the approval of a majority of the Commissioners, and who shall serve for a term of four years unless previously removed by the Chair.

(c) The Division is authorized to develop and establish recommended standards, criteria, and approval process for Qualified Researchers, Qualified Research Projects, Qualified Data and Information, and Qualified Platforms under the processes for notice and comment rulemaking in 5 U.S.C. § 553.

(d) The Division shall publish within six months of enactment of this Act and thereafter as needed but at least annually a list of its criteria for identifying Qualified Researchers, Qualified Research Projects, Qualified Data and Information, and Qualified Platforms. Qualified Researchers may suggest platforms for inclusion. Criteria for qualified researchers shall not include consideration of political views, race, gender, gender identity, ethnicity, sexual orientation, age, or disability, although they may express preference for projects proposed by residents of the United States. No person may be qualified as a Qualified Researcher if they act as an Agent of a foreign power as defined by 50 U.S.C. § 1801.

(e) The Division is authorized to inspect data and question Qualified Platforms about the Qualified Data and Information they are making available to the Commission and to Qualified Researchers.

(f) The Director may issue formal written guidance to persons subject to the Act, provided that the Director shall publish all such guidance within six months of its issuance, with the names of the parties and any trade secret or other confidential information redacted.

(g) In addition to any authority vested in the Division by other provisions of this section, the Director, in carrying out the functions of the Division, is authorized to

(1) prescribe such regulations as the Director deems necessary governing the manner in which its functions shall be carried out;

(2) convene an advisory board from relevant Qualified Platforms and Qualified Researchers;

(3) receive money and other property donated, bequeathed, or devised, without condition or restriction other than that it be used for the purposes of the Division and to use, sell, or otherwise dispose of such property for the purpose of carrying out its functions;

(4) in accordance with the civil service laws, appoint and fix the compensation of such personnel as may be necessary to carry out the provisions of this section;

(5) obtain the services of experts and consultants in accordance with the provisions of section 3109 of title 5, United States Code;

(6) delegate an appropriate entity or independent agency, such as the National Science Foundation (NSF), to assist the Division with carrying out its obligations to appraise Qualified Platforms, Qualified Data and Information, Qualified Researchers, and Qualified Research Projects;

(7) accept and utilize the services of voluntary and non-compensated personnel and reimburse them for travel expenses, including per diem, as authorized by section 5703 of title 5, United States Code;

(8) enter into contracts, grants or other arrangements, or modifications thereof to carry out the provisions of this section, and such contracts or modifications thereof may be entered into without performance or other bonds, and without regard to section 3709 of the Revised Statutes, as amended (41 U.S.C. 5), or any other provision of law relating to competitive bidding;

(9) make advance, progress, and other payments which the Director deems necessary under this title without regard to the provisions of section 3324 (a) and (b) of Title 31; and

(10) make other necessary expenditures.

(h) The Director shall submit to the Chair, to the President, and to the Congress an annual report of the operations of the Division under this Act, which shall include a detailed statement of all private and public funds received and expended by it, and such recommendations as the Director deems appropriate.

SEC. 8. Enforcement

(a) Qualified Researchers who intentionally violate information sharing and privacy standards described in (a) shall be subject to both civil and criminal enforcement, under applicable federal, state, and local laws.

(b) The Commission is hereby empowered and directed to enforce the provisions of this Act, and violations of this Act by a Qualified Platform shall be deemed an unfair trade practice within the meaning of 15 U.S.C. § 45(a)(4).

(c) Whenever the Commission shall have reason to believe that a Qualified Platform has been or is in violation of any provision of this Act, the Commission may commence a civil action in a district court of the United States for an injunction against the Qualified Platform that the Commission believes has violated this Act. Remedies in an injunctive action brought by the Commission are limited to an order enjoining, restraining, or preventing any act or practice that constitutes a violation of this Act and imposing a civil penalty of up to [$10,000] for each violation, which shall accrue to the United States and may be recovered in a civil action brought by the Attorney General of the United States.

(d) In the event any enforcement action is appealed, the prevailing party in the action may, in the discretion of the court, recover the costs of the action including reasonable investigative costs and attorneys' fees.

Hearing Before the U.S. Senate Committee on Homeland Security & Governmental Affairs

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"

Testimony of Dr. Mary Anne Franks
Michael R. Klein Distinguished Scholar Chair and Professor of Law, University of Miami
President and Legislative & Tech Policy Director, Cyber Civil Rights Initiative

Oct. 26, 2021

On October 14, 2021, Facebook announced a new Artificial Intelligence project called Ego4D. The name derives from the project's focus on "egocentric," or first-person, perception, and Facebook plans to use the resulting dataset to, among other things, equip augmented reality glasses and virtual reality headsets with the capacity to transcribe and recall audio and visual recordings of individuals around the user. Asked whether Facebook had implemented measures to address privacy concerns regarding these capabilities, a spokesperson replied that the company "expected that privacy safeguards would be introduced further down the line."

As underscored by multiple internal documents recently released by whistleblower Frances Haugen, this approach is characteristic of Facebook: aggressively push new, untested, and potentially dangerous products into the public realm and worry about the consequences later, if at all. Documents that Haugen shared with the Securities and Exchange Commission reveal the "'asymmetrical' burden placed on employees to 'demonstrate legitimacy and user value' before launching any harm-mitigation tactics—a burden not shared by those developing new features or algorithm changes with growth and engagement in mind." While it may have been abandoned as an official motto, "move fast and break things" still seems to be an accurate description of Facebook's philosophy.

It is even more remarkable that Facebook should choose to announce such a highly controversial new project as the company faces a storm of criticism and scrutiny over documented evidence that it knowingly allowed violent extremism, dangerous misinformation, and sexual exploitation to flourish on its platforms. One might have expected Facebook would be more circumspect about drastically increasing the capacity of individuals to record people around them without consent in light of the revelation, for example, that it allowed nude images of an alleged rape victim to be viewed 56 million times, simply because the man she accused of raping her was a famous soccer star.

Is it arrogance? Is it callousness? Or is it merely – confidence? Confidence that no matter what is revealed about Facebook's role in the disintegration of our shared reality or the dissolution of our democracy – not its acceleration of conspiracy theories from PizzaGate to QAnon to Stop the Steal, its amplification of deadly disinformation about COVID-19, its endangerment of the mental health of teenagers, its preferential treatment of powerful elites, or its promotion of violently racist and sexist propaganda – it will face no real consequences? After all, that seems to be the lesson that Facebook and other dominant tech companies like Twitter and Google have learned every time they have been implicated in scandal. The media attention will be intense for

a while, they might be called before Congress to answer some uncomfortable questions, they may face some fines, some bills attempting to regulate the tech industry might be introduced – but the companies will reassure the public that their purpose was never to cause harm, they will promise to do better, and nothing will really change.

Debates over tech companies' "intentions" tend to serve as a distraction from substantive reform efforts. Moral and legal responsibility is not limited only to those who act with the express purpose of causing harm. We also hold entities accountable when they know their actions will cause harm, when they are reckless about the possibility of harm, and even sometimes when they are negligent about harm. Facebook and other tech companies have known for years that a business model focused on what is euphemistically called "engagement" is ripe for exploitation and abuse. These companies have, at a minimum, consciously disregarded substantial and unjustified risks to individual privacy, equality, and autonomy.

The dominant tech companies are also aware that these risks are not politically neutral. Contrary to oft-repeated claims that social media is biased against conservatives, the algorithms of major social media sites disproportionately amplify right-wing content. The lopsided political amplification of social media is all the more troubling given the disproportionate rate of right-wing violence: "Since 2015, right-wing extremists have been involved in 267 plots or attacks and 91 fatalities," more than four times the number of plots and attacks associated with left-wing ideology.

Facebook repeatedly and deliberately promotes conservative sites on its platform, even changing its algorithm to reduce the visibility of left-leaning news sites and allowing right-wing sites to "skirt the company's fact-checking rules, publish untrustworthy and offensive content and harm the tech giant's relationship with advertisers," despite the efforts of Facebook employees to convince the company to consistently apply its own policies. Internal Facebook research titled "Carol's Journey to QAnon," demonstrated how quickly Facebook's algorithm recommended extremist conspiracy theories to an account set up for an imaginary woman with interests in Fox News and Sinclair Broadcasting. The day after the 2020 election, 10% of all views of political content on Facebook in the U.S. were of posts that falsely claimed that the vote was fraudulent. As one Facebook employee wrote in an internal document, "If [the civic integrity team] takes a hands-off stance for these problems, whether for technical (precision) or philosophical reasons, then the net result is that Facebook, taken as a whole, will be actively (if not necessarily consciously) promoting these types of activities. The mechanics of our platform are not neutral."

Other social media platforms demonstrate partisan patterns as well. Twitter recently released internal research demonstrating that its algorithms also amplify right-wing content more than left-wing content. Research by the Tech Transparency Project found that YouTube algorithms create a much more robust filter bubble for right-wing content than left-wing content, and that Fox is by far the most recommended information channel on YouTube.

Researchers have suggested that the Fox News channel dominates YouTube because it traffics in conspiracy theories and employs more polarizing and inflammatory language than left-leaning channels like MSNBC. The influence of Fox News illustrates that the ecosystem of extremism and disinformation is not limited to social media. Indeed, it would not be much of an

exaggeration to say that Fox News pioneered the strategies of outrage, engagement, and virality that now characterize social media. Scholars have noted that Fox News, more so than Facebook or any other social media platform, "is by far the most influential outlet on the American right," and that television (especially cable news) is more influential than social media as a source of political news for Americans.

To be clear, the object of concern here is not conservative content as such. Rather, it is content that encourages the dehumanization of human beings; targets individuals for violence, threats, and harassment; traffics in dangerous disinformation; and promotes baseless conspiracy theories that undermine our democratic institutions.

The security of America is under attack by those who fear equality and resent the loss of unearned privileges. Social media plays a large role in amplifying these antidemocratic forces, but mainstream media also plays a role, as do elected officials and other figures with influential platforms. No industry and no individual should be considered above the law when it comes to the reckless endangerment of democracy.

Hearing Before the U.S. Senate Committee on Homeland Security & Governmental Affairs
"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"

Supplemental Statement with Policy Recommendations

Dr. Mary Anne Franks
Michael R. Klein Distinguished Scholar Chair and Professor of Law, University of Miami
President and Legislative & Tech Policy Director, Cyber Civil Rights Initiative

November 12, 2021

In my October 26, 2021, written testimony, I offered a summary of how the business model of dominant tech platforms – prioritizing "engagement" for the purposes of extracting and selling increasing amounts of personal information for advertisers and data brokers – inevitably amplifies extreme and harmful content. I further noted that a disproportionate amount of online extremist content that leads to offline violence involves far-right conspiracy theories and propaganda, including false statements about the 2020 election that fueled the violent attack on the Capitol on January 6, 2021. I offer this supplemental statement to respond in more detail to questions raised about the First Amendment during the hearing and to outline my specific policy recommendations for addressing the threat that online extremism poses to democracy.

I. First Amendment and Free Speech Concerns

Several Senators raised concerns during the hearing about how tech industry regulation, including Section 230 reform, might impact the First Amendment and free speech. This is a significant question, as social media platforms have become increasingly important sites of democratic discourse and debate. That is why it is vital to reject any attempt by the government to force social media platforms to carry certain speech or demand that they provide access to certain speakers, as such measures would violate the First Amendment rights of speech and association of these private companies.

Respecting the First Amendment means respecting for tech companies' right to fact-check, label, remove, ban, and make other interventions as they see fit about the content on their sites. Providing additional or alternative information to false or misleading posts is classic "counterspeech," a treasured First Amendment value famously identified by Justice Brandeis in *Whitney v. California*, a landmark free-speech case: "If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence."[1]

The First Amendment also protects the right to refuse to host content altogether, as the right to free speech includes both the right to speak and the right *not* to speak. As the Supreme Court held in *West Virginia State Board of Education v. Barnette*, "If there is any fixed star in our constitutional constellation, it is that no official, high or petty, can prescribe what shall be orthodox in politics, nationalism, religion, or other matters of opinion, or force citizens to confess by word or act their

1. *Whitney v. California*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring).

faith therein."[2] The First Amendment also protects the right of association, including the right of private actors to choose with whom they wish to associate.[3] And the Supreme Court has long recognized that private-property owners generally have the right to exclude individuals from their property as they see fit.[4]

Some Senators expressed concern that allowing tech companies to ban, de-emphasize, label, or otherwise make editorial decisions about speech on their platforms means granting them the power to decide what "truth" is. There are several responses to this. One, as emphasized above, respect for the First Amendment rights of private entities to speak and associate as they see fit cannot be discarded simply because those entities happen to be very powerful or because their decisions inspire disagreement. Two, if the concern is that social media platforms have too much influence over public opinion, this should be a concern that applies equally to what the platforms leave up as much as it applies to what they take down. Finally, we should be careful not to attribute more power to tech companies than they actually have. However ubiquitous social media platforms are, they do not have exclusive control in shaping public opinion or determining access to information. As I noted briefly in my testimony, mainstream media, including cable and network news, radio stations, magazines, and newspapers, still exist and exert considerable influence over many Americans. So do libraries, schools, universities, churches, community centers, town halls, courthouses, and legislatures. The fact that something does not appear on a social media site does not mean it has disappeared from the world.

That being said, the tech industry does enjoy privileges far beyond those of other institutions and industries, which contributes to its outsized influence over public opinion. During the hearing, Senator Romney raised an important question about what distinguishes social media from other media, and one answer is its almost complete lack of accountability either to regulators or to private citizens. News outlets and radio stations can be sued for defamation; property owners can be held liable for injuries sustained on their premises; workplaces must comply with anti-discrimination regulations. But courts' broad interpretation of Section 230 of the Communications Decency Act has preemptively absolved online intermediaries of nearly all liability for harmful content and conduct on their sites and services.

II. Policy Recommendations

Accordingly, the first and most detailed policy recommendation outlined below focuses on Section 230 reform, which is the single most important step toward changing the incentives of the tech industry and thus reducing its harmful influence. Calls for increased transparency, due process, and responsible platform policies are all important, but none of them will be effective without first puncturing the industry's immunity shield. The time for trusting the tech industry to regulate itself is long past; it must finally be forced to internalize the costs of the harms it currently forces the public to bear.

A. Section 230 Reform.

2. 319 U.S. 624, 642 (1943).
3. *Boy Scouts of Am. v. Dale*, 530 U.S. 640, 647-48 (2000).
4. *PruneYard Shopping Ctr. v. Robins*, 447 U.S. 74, 82 (1980) ("[O]ne of the essential sticks in the bundle of property rights is the right to exclude others.").

Section 230 has three key provisions: Section 230 (e)(3) sets out the principle of broad immunity for tech companies, and other two sections detail the two situations in which this principle is applied: when a company leaves harmful content up—(c)(1)—and when it takes it down or restricts it—(c)(2).

While it could fairly be said that Section 230(c)(2), which protects online intermediaries from suit when they remove content they find objectionable, provides incentives for the tech industry to engage in responsible regulation, that effect is undone by Section 230 (c)(1), which has been interpreted to grant them the same protection if they do nothing. Rather than encouraging the innovation and development of measures to fight online abuse and harassment, Section 230 (c)(1) removes incentives for online intermediaries to deter or address harmful practices no matter how easily they could do so. It effectively grants powerful corporations a super-immunity, encouraging them to pursue profit without internalizing any costs of that this pursuit. It eliminates real incentives for tech corporations to design safer platforms or more secure products. Section 230 (c)(1)'s preemptive immunity ensures that no duty of care ever emerges in a vast range of online scenarios and eliminates the incentives for the best positioned party to develop responses to avoid foreseeable risks of harm.[5]

Not only does Section 230 (c)(1) fail to incentivize safer tech products and practices, but it also denies members of the public access to the courts to seek redress for injuries. Private individuals are left to deal with the fallout of a reckless tech industry moving fast and breaking things – including life-destroying harassment, publicized sexual exploitation, and ubiquitous surveillance – on their own. Section 230 preempts plaintiffs from ever bringing suit in many cases and makes it difficult for any suit that is brought to survive a motion to dismiss.

Some Section 230 defenders argue that many suits against online intermediaries will ultimately fail on the merits. But whether a plaintiff's claim will ultimately succeed in any given case is always indeterminate. The value of the right to bring the claim does not turn on whether the claim is vindicated in the end. Moreover, in many cases, the discovery process will provide significant value not just to the plaintiff in the case at hand, but to legislators, regulators, future plaintiffs, and the public.

Congress should amend Section 230 as detailed below to allow people who been injured by online harms to have their day in court. Unless the harmful content or conduct in question is clearly speech protected by the First Amendment, plaintiffs should not be barred from suing online intermediaries, and online intermediaries that demonstrate deliberate indifference to harmful content unprotected by the First Amendment should not be able to take advantage of Section 230's protections.

1. Limit Section 230's protections to speech protected by the First Amendment.

Both critics and defenders of Section 230 agree that the statute provides online intermediaries broad immunity from liability for a wide range of Internet activity. While critics of Section 230 point to the extensive range of harmful activity that the law's deregulatory stance effectively allows to flourish, Section 230 defenders argue that an unfettered Internet is vital to a robust online

[5] Michael L. Rustad, Thomas H. Koenig, *Rebooting Cybertort Law*, 80 WASH. L. REV. 335, 382 (2005)

marketplace of ideas. The marketplace of ideas is a familiar and powerful concept in First Amendment doctrine, serving as a justification for a laissez-faire approach to speech. Its central claim is that the best approach to bad or harmful speech is to let it circulate freely, because letting ideas compete in the market is the best way to sort truth from falsity and good speech from bad speech, and because government cannot be trusted to make such decisions wisely or fairly.

The Internet-as-marketplace-of-ideas presumes that the Internet is primarily, if not exclusively, a medium of speech. The text of Section 230 reinforces this characterization with the terms "publish," "publishers," "speech," and "speakers" in 230(c), as well as the finding that the "Internet and other interactive computer services offer a forum for a true diversity of political discourse, unique opportunities for cultural development, and myriad avenues for intellectual activity."

When Section 230 was passed, it may have made sense to think of the Internet primarily as a speech machine. In 1996, the Internet was text-based and predominantly noncommercial. Only 20 million American adults had Internet access, and these users spent less than half an hour a month online. But by 2019, 293 million Americans were using the Internet, and they were using it not only to communicate, but also to buy and sell merchandise, find dates, make restaurant reservations, watch television, read books, stream music, and look for jobs. Many of these activities have very little to do with speech, and many of their offline cognates would not be considered speech for First Amendment purposes. If the broad immunity afforded online intermediaries is justified on First Amendment principles, then it should apply only to online activity that can plausibly be characterized as speech protected by the First Amendment. What is more, it should only apply to third-party protected speech for which platforms serve as true intermediaries, not speech that the platform itself creates, controls, or profits from.

To accomplish this, the word "information" in Section 230 (c)(1) should be replaced with the word "speech protected by the First Amendment." This revision would put all parties in litigation on notice that the classification of content as protected speech is not a given, but a fact to be demonstrated. If a platform cannot make a showing that the content or information at issue is speech, then it should not be able to take advantage of Section 230 immunity.

2. Incorporate the longstanding principle of collective responsibility.

Many harmful acts are only possible with the participation of multiple actors with various motivations. The doctrines of aiding and abetting, complicity, and conspiracy all reflect the insight that third parties who assist, encourage, ignore, or contribute to the illegal actions of another person can and should be held responsible for their contributions to the harms that result, particularly if those third parties benefited in some material way from that contribution. While U.S. law, unlike the law of some countries, does not impose a general duty to aid, it does recognize the concept of collective responsibility. Third parties can be held both criminally and civilly liable for the actions of other people for harmful acts they did not cause but did not do enough to prevent.

Among the justifications for third-party liability in criminal and civil law is that this liability incentivizes responsible behavior. Bartenders who serve alcohol to obviously inebriated patrons can be sued if those patrons go on to cause car accidents; grocery stores can be held accountable for failing to clean up spills that lead to slip and falls; employers can be liable for failing to respond

to reports of sexual harassment. Such entities are often said to have breached a "duty of care," and imposing liability is intended to give them incentive to be more careful in the future. It is a central tenet of tort law that the possibility of such liability incentivizes individuals and industries to act responsibly and reasonably.

Conversely, grants of immunity from such liability risk encouraging negligent and reckless behavior. The immunity granted by Section 230 does just that, despite the evocative title of its operative clause, "Protection for 'Good Samaritan' blocking and screening of offensive material." This title suggests that Section 230 is meant to provide "Good Samaritan" immunity in much the same sense as "Good Samaritan" laws in physical space. Such laws do not create a duty to aid, but instead provide immunity to those who attempt in good faith and without legal obligation to aid others in distress. While Good Samaritan laws generally do not require people to offer assistance, they encourage people to assist others in need by removing the threat of liability for doing so.

Subsection (c)(2) of Section 230 is a Good Samaritan law in a straightforward sense: it assures providers and users of interactive computer services that they will not be held liable with regard to any action "voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable" or "taken to enable or make available to information content providers or others the technical means to restrict access" to such material. Importantly, because most interactive computer service providers are private entities, their right to choose whether to carry, promote, or associate themselves with speech is not created by Section 230, but by the First Amendment. Subsection (c)(2) merely reinforces this right by making it procedurally easier to avoid specious lawsuits.

On the other hand, Subsection 230(c)(1)'s broad statement that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider," has been interpreted in ways directly at odds with Good Samaritan laws, as well as with a host of other legal principles and settled law. Where (c)(2) offers immunity to interactive computer service providers in exchange for intervening in situations where they have no duty of care, (c)(1) has been read to provide the same immunity to providers who do nothing at all to stop harmful conduct – and, even more perversely, extends that same immunity to providers who actively profit from or solicit harmful conduct. For example, Section 230(c)(1) has been invoked to protect message boards like 8chan (now 8kun), which provide a platform for mass shooters to spread terrorist propaganda, online firearms marketplaces such as Armslist, which facilitate the illegal sale of weapons used to murder domestic violence victims, and to classifieds services like the now-defunct site Backpage, which was routinely used by sex traffickers to advertise underage girls for sex.

In subsidizing platforms that directly benefit from illegal and harmful conduct, Section 230(c)(1) creates a classic "moral hazard," ensuring that the multibillion-dollar corporations that exert near-monopoly control of the Internet are protected from the costs of their risky ventures even as they reap the benefits. Given that the dominant business model of websites and social media services is based on advertising revenue, they have no natural incentive to discourage abusive or harmful conduct: "abusive posts still bring in considerable ad revenue… the more content that is posted, good or bad, the more ad money goes into their coffers."

Online intermediaries who do not voluntarily intervene to prevent or alleviate harm inflicted by another person are in no sense "Good Samaritans." They are at best passive bystanders who do nothing to intervene against harm, and at worst, they are accomplices who encourage and profit from harm. Providing them with immunity flies in the face of the longstanding legal principle of collective responsibility that governs conduct in the physical world. In physical spaces, individuals or businesses that fail to "take care" that their products, services or premises are not used to commit wrongdoing can be held accountable for that failure. There is no justification for abandoning this principle simply because the conduct occurs online. In fact, there are more compelling reasons for recognizing collective responsibility online, because online interaction provides so many opportunities for direct tortfeasors to escape detection or identification.

Creating a two-track system of liability for offline and online conduct not only encourages illegality to move online, but also erodes the rule of law offline. Offline entities can plausibly complain that the differential treatment afforded by broad interpretations of Section 230 violates principles of fairness and equal protection, or to put it more bluntly: if they can do it, why can't we? There is a real risk that Section 230's abandonment of the concept of collective responsibility will become the law offline as well as on.

To undo this, Section 230 (c)1 should be further amended to clarify that providers or users of interactive computer services cannot be treated as the publisher or speaker of protected speech *wholly provided by* another information content provider, *unless such provider or user intentionally encourages, solicits, or generates revenue from this speech.* In addition, a new subsection should be added to Section 230 to explicitly exclude from immunity intermediaries who exhibit deliberate indifference to unlawful content or conduct.

The revised version of Section 230(c) would read:

> (1) **Treatment of publisher or speaker**
> No provider or user of an interactive computer service shall be treated as the publisher or speaker of any ~~information~~ speech protected by the First Amendment wholly provided by another information content provider, unless such provider or user intentionally encourages, solicits, or generates revenue from this speech.
>
> (2) **Civil liability**
> No provider or user of an interactive computer service shall be held liable on account of-
> (**A**) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or
> (**B**) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1);
>
> (3) **Limitations.** The protections of this section shall not be available to a provider or user who manifests deliberate indifference to unlawful material or conduct.

B. Targeted Federal Criminal Legislation.

Congress should enact federal criminal legislation addressing new and highly destructive forms of technology-facilitated abuse, especially those disproportionately targeted at vulnerable groups, including nonconsensual pornography, sexual extortion, doxing, and digital forgeries ("deep fakes"). As Section 230 immunity does not apply to violations of federal criminal law, the creation of these laws will ensure that victims of these abuses will have a path to justice with or without Section 230 reform.

The anonymity, amplification, and aggregation possibilities offered by the Internet have allowed private actors to discriminate, harass, and threaten vulnerable groups on a massive scale. There is empirical evidence showing that the Internet has been used to further chill the intimate, artistic, and professional expression of individuals whose rights were already under assault offline. Even as the Internet has multiplied the possibilities of expression, it has multiplied the possibilities of repression. The new forms of communication offered by the Internet have been used to unleash a regressive and censorious backlash against women, racial minorities, and sexual minorities.

The Internet lowers the costs of engaging in abuse by providing abusers with anonymity and social validation, while providing new ways to increase the range and impact of that abuse. The online abuse of women amplifies sexist stereotyping and discrimination, compromising gender equality online and off. Victims of online abuse do not feel safe on or offline. They experience anxiety and severe emotional distress. They suffer damage to their reputations and intimate relationships as well as their employment and educational opportunities. Some victims are forced to relocate, change jobs, or even change their names. Because the abuse so often appears in Internet searches of their names, victims have difficulty finding employment or keeping their jobs.

Failing to address online abuse does not just inflict economic, physical, and psychological harms on victims—it also jeopardizes their right to free speech. Online abuse silences victims. Targeted individuals often shut down social media profiles and e-mail accounts and withdraw from public discourse. Those with political ambitions are deterred from running for office. Journalists refrain from reporting on controversial topics. Victims of sexual extortion are coerced into silence with threats of violence, insulating perpetrators from accountability.

Technology-facilitated abuses that have proven particularly destructive include nonconsensual pornography (also known as "revenge porn"), sexual extortion (also called "sextortion," a form of blackmail in which sexual information or images are used to extort sexual acts and/or money from the victim), doxing (the publication of private or personally identifying information, often with malicious intent), and so-called "deep fakes" (the use of technology to create false visual and audio material indistinguishable from authentic visual and audio representations of individuals). Fortunately, strong federal legislation has already been drafted to address the first three issues, and there are strong efforts to address the fourth in process.

The Senate should pass the Stopping Harmful Image Exploitation and Limiting Distribution (SHIELD) Act of 2019, now included in the Violence Against Women Reauthorization Act of 2021, which would make it a crime to knowingly distribute or threaten to distribute private, sexually explicit visual material of an individual with knowledge of or reckless disregard for the depicted individual's lack of consent to the distribution and reasonable expectation of privacy and without a reasonable belief that distributing the depiction touches a matter of public concern.

Congress should also pass a measure similar to the Online Safety Modernization Act of 2017 (H.R.3067), sponsored by Congresswoman Katherine Clark, which would prohibit multiple forms of "cybercrimes against individuals" including both sextortion and doxing.

Congress should also draft and enact a statute criminalizing so-called "deep fakes." The statute should target a narrow category of digital forgeries, defined as audiovisual material that has been created or materially altered to falsely appear to a reasonable observer to be an actual record of actual speech, conduct, appearance, or absence of an individual, that are created, distributed, or reproduced with the intent to seriously harm or with reckless disregard for whether serious harm would result to a falsely depicted individual, or with the intent to incite violence or interfere with official proceedings.

SPLC | ACTION

October 27, 2021

The Honorable Gary Peters
Chair
Homeland Security & Governmental Affairs Committee
U.S. Senate
Washington, DC 20510

The Honorable Rob Portman
Ranking Member
Homeland Security & Governmental Affairs Committee
U.S. Senate
Washington, DC 20510

Dear Chair Peters and Ranking Member Porter:

On behalf of the Southern Poverty Law Center (SPLC) Action Fund, we write to provide our insights for your hearings on "Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content." We appreciate the opportunity to share our expertise as you investigate the connection between tech platforms and hate and extremism and to offer several policy recommendations to address this problem. We would ask that this statement be included as part of the official hearing record.

The SPLC is the premier U.S. nonprofit advocacy organization working to serve as a catalyst for racial justice in the South and beyond. We work in partnership with communities to dismantle white supremacy, strengthen intersectional movements, and advance human rights of all people. We have deep expertise in monitoring the activities of domestic hate groups and other extremists – including the Ku Klux Klan, the neo-Nazi movement, neo-Confederates, racist skinheads, anti-racist skinheads, antigovernment militias, and others. We currently track more than 1,600 extremist groups operating across the country and publish investigative reports, share key intelligence, and offer expert analysis to the media and public. SPLC employs a three-pronged strategy: litigation, public education, and policy advocacy. The organization works in the courts to win systemic reforms on behalf of victims of bigotry and discrimination. Through "Learning for Justice" the organization provides free resources to caregivers and educators to help advance human rights and inclusive democracy.

Related, the SPLC Action Fund advocates for the implementation of policies and laws to eliminate the structural racism and inequalities that fuel oppression of people of color, immigrants, young people, women, low-income people, and the LGBTQ+ community. The Action Fund is dedicated to fighting for racial justice alongside impacted communities in pursuit of equity and opportunity for all. We work primarily in the Southeast United States where we have offices in Alabama, Georgia, Florida, Louisiana, Mississippi, and Washington, D.C.

Our organization has watched for years as social media companies failed to uphold their own terms and conditions, enabling the expansion and radicalization of the far right, and, also, violence. Repeatedly, we have alerted companies like Twitter, YouTube, and Facebook to how far-right extremists organize on their platforms. These companies have often responded with half-measures, or in some cases, inaction, to what we flagged. Importantly, we have seen in cases

where social media companies do take steps to remove extremists from their platform, as they did in 2018 with Alex Jones and in 2020 with Canadian white supremacist Stefan Molyneux, these figures become limited in their ability to spread hateful propaganda and lies. These powerful, highly trafficked platforms have the power to reduce harm if they want to do it. To do so, they would have to choose responsibility to democracy and the public good over private profit.

Twitter and Facebook present themselves as being non-ideological resources for communication but have demonstrable ties to the far right and long histories of promoting hate to their consumers. Facebook has partnered on fact checking with a group called Check Your Fact, which is linked to far-right *The Daily Caller*.[1] Founded by FOX News' Tucker Carlson, *The Daily Caller* has in the past employed white nationalist activists,[2] including Jason Kessler,[3] who helped stage the deadly August 2017 Unite the Right event in Charlottesville, Virginia, and Peter Brimelow, the founder of the hate group VDARE.[4] Pro-Trump billionaire Peter Thiel, who has promoted and funded anti-immigrant political campaigns,[5] and is aligned with anti-democratic figures like Curtis Yarvin, serves on Facebook's board.[6] Facebook included the low-standard publication *Breitbart News* as part of its news section,[7] despite its history of publishing racist, anti-immigrant posts[8] and blogs authored by extremists.[9]

Facebook's role in facilitating extremist activity is well known. In advance of the deadly January 6 attack on the Capitol, anti-government militia activists organized openly on social media platforms including Facebook and Twitter. In charging documents[10] for January 6th rioters, Kelly and Connie Meggs, leaders of the Florida Oath Keepers chapter, prosecutors say the militia group used Facebook to discuss forming an "alliance" and coordinating plans with another extremist group, the Proud Boys, ahead of the riot at the Capitol. However, this is not a new problem that is limited to the events exposed in the wake of the January 6th attacks. Anti-government militia organizations were well documented by watchdog organizations and federal investigators organizing a "Call to Action" on Facebook for the armed stand-offs led by Cliven and Ammon Bundy and their associates at the Malheur National Wildlife Refuge in Oregon in 2016 and in Nevada in 2014. Hundreds of thousands of Facebook posts and messages were revealed in the trial for the 41-day occupation in Burns, Oregon. By the fall of 2020 Bundy's People's Rights had amassed some

---

[1] Scott Waldman, *Science Insider,* "Facebook fact checker has ties to news outlet that promotes climate doubt" https://www.science.org/content/article/facebook-fact-checker-has-ties-news-outlet-promotes-climate-doubt, April 25, 2019.
[2] Matt Gertz, *Media Matters,* "The Daily Caller has published white supremacists, anti-Semites, and bigots. Here are the ones we know about," https://www.mediamatters.org/maga-trolls/daily-caller-has-published-white-supremacists-anti-semites-and-bigots-here-are-ones-we, September 5, 2018.
[3] https://www.splcenter.org/fighting-hate/extremist-files/individual/peter-brimelow
[4] https://www.splcenter.org/fighting-hate/extremist-files/individual/jason-kessler
[5] Hannah Gais, SPLC *Hatchwatch,* "White Nationalist Who Met With Peter Thiel Admired Terroristic Literature," https://www.splcenter.org/hatewatch/2021/03/18/white-nationalist-who-met-peter-thiel-admired-terroristic-literature, March 18, 2021.
[6] Benjamin Wallace-Wells, *The New Yorker,* "The Rise of the Thielists: Has the Republican Party found its post-Trump ideology?", https://www.newyorker.com/news/annals-of-populism/the-rise-of-the-thielists, May 13, 2021.
[7] Francis Augustin, *Business Insider,* "Facebook chose to keep Breitbart on News Tab and gave it special treatment — even after employees warned of its embellished and hyper-partisan coverage of events like the George Floyd protests," https://www.businessinsider.com/facebook-files-breitbart-news-tab-employee-objections-2021-10
[8] *SPLC Hatewatch,* "Stephen Miller: The Breitbart Emails," https://www.splcenter.org/stephen-miller-breitbart-emails, October 24, 2021.
[9] Joseph Bernstein, *Buzzfeed News,* "Here's How Breitbart And Milo Smuggled White Nationalism Into The Mainstream," https://www.buzzfeednews.com/article/josephbernstein/heres-how-breitbart-and-milo-smuggled-white-nationalism, October 5, 2017.
[10] *United States of America v. Kelly Meggs and Connie Meggs,* https://www.documentcloud.org/documents/20527411-supplemental-meggs-motion

20,000 members in sixteen states, primarily through organizing on Facebook, before the removal of many pages by the company.

Twitter CEO Jack Dorsey has acknowledged a connection to Ali Alexander, the extremist who led the "Stop the Steal" movement.[11] He "follows" on his own website far-right figures including Mike Cernovich, who is known primarily for spreading disinformation on it, including "Pizzagate," as well as lies about the outcome of the 2020 presidential election.[12] Dorsey also follows Tim Pool, a Pro-Trump YouTube performer, who claims to chat with the tech mogul.[13] Election Integrity Partnership, a non-partisan group that includes researchers from Stanford University and the University of Washington, listed Pool among a group of verified Twitter "superspreaders" who pushed disinformation to Twitter following the 2020 election.[14] Pool told us in an interview that Dorsey "seems very adamant that far-right figures be given unrestrained platforms," based upon his private conversations with him.

Although news outlets have focused acutely on Facebook in recent weeks and the way its business model transports right-leaning people into more extreme spaces, the site hardly stands alone in its capacity to feed consumers a diet of radical, far-right propaganda.[15] In July, our organization published a longform analysis demonstrating how Twitter enabled the attack on the U.S. Capitol by "verifying" extremists with no discernible credentials as public figures and allowing them to amass huge platforms on the back of spreading politically-charged disinformation and hate.[16] Extremists first started pushing the #StoptheSteal hashtag on Twitter during the 2016 election, and for over five years, made it synonymous with discrediting the votes of people who live in predominantly Black neighborhoods of swing states. In the aftermath of the 2020 election, Twitter allowed many of the same extremists who pushed the hashtag in previous years to do so again. In the aftermath of the election, right-wing users sometimes shared the #StoptheSteal hashtag dozens of times per minute.[17]

The Southern Poverty Law Center has repeatedly flagged to Twitter the account of Jack Posobiec, a "Stop the Steal" extremist who they verified in April 2017, at a time when the public primarily knew him for promoting Pizzagate[18] and other disinformation campaigns.[19] In the months leading up to Twitter verifying Posobiec's account, he labeled himself as a "fmr CBS News Journalist" in his bio on the site, even though CBS News told the Southern Poverty Law Center that he never

---

[11] Jesselyn Cook, "Jack Dorsey Thinks Fringe Figure Ali Akbar Makes 'Interesting Points'," https://www.huffpost.com/entry/ali-akbar-jack-dorsey-twitter_n_5c40cb9ce4b027c3bbbf3a0c, January 18, 2019.
[12] Michael Edison Hayden, "'We Make Mistakes': Twitter's Embrace of the Extreme Far Right," https://www.splcenter.org/hatewatch/2021/07/07/we-make-mistakes-twitters-embrace-extreme-far-right, July 7, 2021.
[13] Ibid.
[14] Kari Paul, *The Guardian,* "A few rightwing 'super-spreaders' fueled bulk of election falsehoods, study says," https://www.theguardian.com/us-news/2021/mar/05/election-misinformation-trump-rightwing-super-spreader-study, March 5, 2021.
[15] Brandy Zadrozny, *NBC News,* "'Carol's Journey': What Facebook knew about how it radicalized users," https://www.nbcnews.com/tech/tech-news/facebook-knew-radicalized-users-rcna3581, October 22, 2021.
[16] Michael Edison Hayden, *SPLC Hatewatch,* "'We Make Mistakes': Twitter's Embrace of the Extreme Far Right," https://www.splcenter.org/hatewatch/2021/07/07/we-make-mistakes-twitters-embrace-extreme-far-right, July 7, 2021
[17] Michael Edison Hayden, *SPLC Hatewatch,* "Far Right Resurrects Roger Stone's #StopTheSteal During Vote Count," https://www.splcenter.org/hatewatch/2020/11/06/far-right-resurrects-roger-stones-stopthesteal-during-vote-count, November 6, 2021.
[18] Michael Edison Hayden, *SPLC Hatewatch,* "SPLC Investigation: Far-Right OANN Anchor Jack Posobiec's Rise Tied to White Supremacist Movement," https://www.splcenter.org/splc-investigation-far-right-oann-anchor-jack-posobiecs-rise-tied-white-supremacist-movement, July 8, 2020
[19] Joseph Bernstein, *Buzzfeed News,* "Inside The Alt-Right's Campaign To Smear Trump Protesters As Anarchists," https://www.buzzfeednews.com/article/josephbernstein/inside-the-alt-rights-campaign-to-smear-trump-protesters-as, January 11, 2017.

worked for them.[20] He has targeted Jewish journalists with hate,[21] hyped to his followers a Polish neofascist group that during the 1930s bombed Jewish homes,[22] and repeatedly promoted politically charged lies on his way to reaching over one million followers.[23] Posobiec has also promoted to Twitter multiple Russian military intelligence-led social media operations,[24] including Macron Leaks, which sought to disrupt the results of the 2017 elections in France. Multiple analysts have pointed out the likelihood that automation stemming from Russia buoyed his account during that campaign.[25] Posobiec posted to his followers about "Stop the Steal" as early as September 7, 2020 – nearly two months before election day.[26]

Twitter has also verified other extremists involved in pushing Pizzagate,[27] including Cernovich and Gateway Pundit blogger Cassandra Fairbanks,[28] who has read excerpts from the Unabomber manifesto to her followers on that site and misled them to believe "antifa" attacked her while raising $25,000 in handouts off of that lie.[29] Like Posobiec, both extremists also pushed lies about the 2020 election four years after Pizzagate, suggesting that if Twitter had removed them in 2016, the company may have reduced the spread of additional lies that inspired violence on January 6. Other verified "Stop the Steal" extremists, including Brandon Straka and Scott Presler, built followings in the hundreds of thousands after being promoted by the company's business model.[30] Neither man had any claim to being a public figure outside of the popularity they gained through their verified Twitter handles. The FBI notes in their indictment of Straka, relative to his actions on January 6, over a dozen different references to his Twitter account.[31] Analysts previously noted the degree to which automated account or "bots" appear to have buoyed his tweets, making him more visible on the site. [32]

---

[20] Michael Edison Hayden, *SPLC Hatewatch,* "Jack Posobiec's Rise Tied to White Supremacist Movement," https://www.splcenter.org/hatewatch/2020/07/08/jack-posobiecs-rise-tied-white-supremacist-movement, July 8, 2020.
[21] Michael Edison Hayden, *SPLC Hatewatch,* "Twitter Gave Free Rein for Jack Posobiec To Publish Antisemitic Hate and Disinformation," https://www.splcenter.org/hatewatch/2020/07/08/twitter-gave-free-rein-jack-posobiec-publish-antisemitic-hate-and-disinformation, July 8, 2020.
[22] Michael Edison Hayden, *SPLC Hatewatch,* "OANN's Posobiec Met With Polish Neo-Fascists and Amplified Their Messages Online," https://www.splcenter.org/hatewatch/2020/07/22/oanns-posobiec-met-polish-neo-fascists-and-amplified-their-messages-online. July 22, 2020.
[23] Michael Edison Hayden, *SPLC Hatewatch,* "Neo-Nazi Collaborator Jack Posobiec Crosses One Million Twitter Followers," https://www.splcenter.org/hatewatch/2020/09/09/neo-nazi-collaborator-jack-posobiec-crosses-one-million-twitter-followers, September 9, 2020.
[24] Michael Edison Hayden, *SPLC Hatewatch,* "Jack Posobiec Links to Russian Intelligence-Backed Website," https://www.splcenter.org/hatewatch/2021/04/28/jack-posobiec-links-russian-intelligence-backed-website, April 28, 2021.
[25] Michael Edison Hayden, *SPLC Hatewatch,* "Jack Posobiec Central in Spreading Russian Intelligence-Led #MacronLeaks Hack," https://www.splcenter.org/hatewatch/2021/01/29/jack-posobiec-central-spreading-russian-intelligence-led-macronleaks-hack, January 29, 2021.
[26] Atlantic Council's DFRDLab, *Just Security,* "#StopTheSteal: Timeline of Social Media and Extremist Activities Leading to 1/6 Insurrection," https://www.justsecurity.org/74622/stopthesteal-timeline-of-social-media-and-extremist-activities-leading-to-1-6-insurrection/, February 10, 2021.
[27] Michael Edison Hayden, *SPLC Hatewatch,* "'There's nothing you can do': The Legacy of #PizzaGate," https://www.splcenter.org/hatewatch/2021/07/07/theres-nothing-you-can-do-legacy-pizzagate, July 7, 2021.
[28] Jared Holt, *Right Wing Watch,* "Gateway Pundit Reporter Lavishes Praise on Unabomber: 'A Brilliant Man'," https://www.rightwingwatch.org/post/gateway-pundit-reporter-lavishes-praise-on-unabomber-a-brilliant-man/, May 6, 2020.
[29] Jared Holt, *Right Wing Watch,* "Cassandra Fairbanks Claims Antifa Attacked Her. Police Reports and Neighbors Say Otherwise," https://www.rightwingwatch.org/post/cassandra-fairbanks-%e2%80%8bclaims-antifa-attacked-her%e2%80%8b-police-reports-and-neighbors-say-otherwise/, June 17, 2020.
[30] Michael Edison Hayden, "'We Make Mistakes': Twitter's Embrace of the Extreme Far Right," https://www.splcenter.org/hatewatch/2021/07/07/we-make-mistakes-twitters-embrace-extreme-far-right, July 7, 2021
[31] Ibid.
[32] Tim Fitzsimons, *NBC News,* "Meet Brandon Straka, a gay former liberal encouraging others to #WalkAway from Democrats," https://www.nbcnews.com/feature/nbc-out/meet-brandon-straka-gay-former-liberal-encouraging-others-walkaway-democrats-n902316, August 21, 2018.

Beyond these figures, Twitter provides the only mainstream social media home for white nationalist leaders like Peter Brimelow,[33] Richard Spencer, [34] and Jason Kessler, [35] who operate on the site under their own names and faces. Pseudonymous white supremacists and serial harassers also commonly cycle through handles on Twitter after the site suspends them, suggesting that ban evasion represents a pervasive problem for the company.[36] Twitter's algorithm frequently recommends that users follow white supremacists, sometimes in groups of five or more accounts at one time. [37] Suspended U.S. State Department official Matt Gebert, who started posting to Twitter in more conventional right-wing circles during the Obama era, began sharing overt white supremacist material only a few years after joining the site, we found in a review of his accounts.[38] By the Trump era, Gebert hosted clandestine meetings of white supremacists at his home. He obtained a security clearance, and the State Department briefed the Senate Foreign Relations committee about our findings about him in August 2019.[39]

In recent years, we have seen the emergence of a disturbing new trend of voter suppression against Black and Latinx people through the use of targeted mis/disinformation campaigns spread on social media platforms by both foreign and domestic sources. These efforts involve outside groups posing as Black or Latinx community leaders and influencers, building large followings, and then spreading false information about how, when, and where to vote.[40] A Brennan Center report revealed that most online voter suppression campaigns were targeted at "specific segments of the population in terms of race, gender, and income, potentially leading to discriminatory effects."[41] That report also documented that these campaigns were most often sponsored by undisclosed groups. While it may be unclear who is funding these misleading and intentionally polarizing campaigns, the intent is unmistakable: "to depress turnout among people of color by fueling cynicism and distrust in the political process."[42]

One example, documented by an October 2019 Senate Select Committee on Intelligence report,[43] revealed Russian interference in the 2016 presidential campaign. According to the report, Russian operatives, under the direction of the Kremlin-backed Internet Research Agency, worked to manipulate and distort facts, create and expand racial divisions, and discourage Black and Latinx voters from showing up at the polls.

---

[33] https://twitter.com/peterbrimelow?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
[34] https://twitter.com/RichardBSpencer
[35] https://twitter.com/TheMadDimension
[36] Michael Edison Hayden, *SPLC Hatewatch,* "Spectre Unmasked: Racist 'Alt-Right' Podcaster Used To Be Local Reporter," https://www.splcenter.org/hatewatch/2019/01/18/spectre-unmasked-racist-alt-right-podcaster-used-be-local-reporter, January 18, 2019.
[37] https://twitter.com/MichaelEHayden/status/1441556029926428673?s=20
[38] Michael Edison Hayden, *SPLC Hatewatch,* "U.S. State Department Official Involved in White Nationalist Movement, Hatewatch Determines," https://www.splcenter.org/gebert, August 7, 2019.
[39] Jennifer Hansler, *CNN News,* "State Department official on leave after civil rights group accuses him of being involved in white nationalist movement, source says," https://www.cnn.com/2019/08/09/politics/matthew-gebert-splc/index.html, August 9, 2019.
[40] Ashley Bryant, *Campaigns and Elections,* "Combatting the Disinformation Campaign Targeting Black and Latinx Voters," https://campaignsandelections.com/industry-news/combatting-the-disinformation-campaign-targeting-black-and-latinx-voters/, July 20, 2020
[41] Young Mie Kim, Brennan Center, "Voter Suppression Has Gone Digital" https://www.brennancenter.org/our-work/analysis-opinion/voter-suppression-has-gone-digital, November 20, 2018
[42] Shannon Bond, *National Public Radio,* "Black And Latino Voters Flooded With Disinformation In Election's Final Days," https://www.npr.org/2020/10/30/929248146/black-and-latino-voters-flooded-with-disinformation-in-elections-final-days, October 30, 2020
[43] Senate Select Committee on Intelligence, "Russian Active Measures: Campaigns and Interference in the 2016 Elections, Volume 2, Russia's Use of Social Media," https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf, October 8, 2019

Finally, video-sharing websites like YouTube and TikTok also show how a manipulative business model mixed with extreme far-right content can intensify the radicalization of users. YouTube's recommendation algorithm has nudged consumers of more mainstream content toward increasingly radical material, according to multiple analysts and the statements of former far-right extremists who have left behind that movement. Researcher Becca Lewis posited in her 2018 report "Alternative Influence," that YouTube's creators "built [it] to incentivize the behavior of [far-right] political influencers."[44] In a report, commissioned by the New Zealand government in the aftermath of the March 2019 terrorist attack in Christchurch, the man who gunned down 51 Muslims, Brenton Tarrant, told authorities that YouTube provided a "significant source of information and inspiration" for his racist views. Researchers have flagged that TikTok has had a similar impact[45] – driving right-leaning consumers to QAnon content and videos promoting violent extremist groups like the Oath Keepers.[46]

Policy Recommendation

Recommendations for tech companies

- Most tech companies have their own Terms of Service, essentially rules of the road. The SPLC encourages corporations to create – and enforce – policies and terms of service to ensure that social media platforms, payment service providers, and other internet-based services do not provide platforms where hateful activities and extremism can grow. When tech companies do decide to act against hate, it is too often only after a violent attack has occurred. They need to proactively address the problem of extremist content on their platforms rather than simply react after people have been killed.

- Tech companies of all sizes should coordinate to understand how the impact of one company's actions affects others. If a right-wing extremist loses their platform at Facebook and YouTube, but remains on Twitter, they can continue to reach a wide audience. Companies must coordinate to anticipate and measure this phenomenon as it happens.

- Tech companies must understand and appreciate their unique position as the single largest storehouses of human knowledge about complex socio-technical problems. Tech companies store vast quantities of invaluable, irreplaceable data, and currently they are the sole arbiters of who gets to use that data and for what purpose. Instead of locking out researchers, tech companies should proactively engage outside partners to put their vast quantities of data to work. These efforts must go beyond just providing pre-curated data sets. Furthermore, companies should not limit researcher access to individuals associated with a university. Journalists, lawyers, Congressional staff, and other interested parties should be allowed access to data without such unnecessary gatekeeping measures.

- Tech companies must redesign their "Trust and Safety" systems to reward people who want to do the right thing. At the same time, services must not solely rely on users to report content. Small companies should be financially rewarded for building powerful, proactive teams that keep their services safe and abuse-free.

---

[44] Rebecca Lewis, *Data & Society,* "Alternative Influence: Broadcasting the Reactionary Right on YouTube," https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf, September 18, 2018.
[45] New Zealand Royal Commission of Inquiry, "Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019," https://christchurchattack.royalcommission.nz/the-report/firearms-licensing/general-life-in-new-zealand/, December 21, 2020.
[46] Matt Binder, *Mashable,* "TikTok's algorithm is sending users down a far-right extremist rabbit hole," https://mashable.com/article/tiktok-recommendations-far-right-wing, March 28, 2020.

- Tech companies must not treat white supremacist or anti-government militias groups with a light touch out of fear of appearing to be politically biased. As a recent leak of Facebook's Dangerous Individuals and Organizations list[47] revealed, Facebook's tiered approach to handling groups engaged in promoting violence and hateful views online places looser prohibitions on predominantly white anti-government militias than on groups and individuals listed as terrorists or as violent criminal enterprises, the majority of whom are non-white. Social media companies must protect their most vulnerable users, rather than perpetuate the injustices leveled against Muslim communities in particular through America's "war on terror."

Recommendations for the Biden administration and Congress

- Require public transparency and accountability with respect to the harmful content proliferated on media platforms.

- Require disclosure about who is paying for specific online political advertisements.

- Require regular, mandatory reporting by technology service providers to document abuse of their systems including financial support of violence, harassment, and terrorism.

- Invest in basic and applied research. Many thorny issues involving these social media platforms – such as how financial exploitation can be tracked on encrypted platforms, for example, or how cryptocurrency transactions can be tracked at scale – may have technology-based solutions.
    - Prioritize funding programs for research into technologies that can be used to detect and prevent online financial harms while preserving human rights. This is especially critical as we anticipate extremists' gradual move to a more decentralized technology landscape.
    - Ensure improved training at the state and local levels to better enable detection of fraud that could have federal implications.

- Promote anti-bias education and resilience programs that help steer individuals away from hate and extremism. The law is a blunt instrument to confront hate and extremism; it is much better to prevent these criminal acts in the first place. Since it is not possible to legislate, regulate, or tabulate racism, hatred, or extremism out of existence, we need federal and state government leadership to promote anti-bias, antihate, and democracy-building education programs – such as the SPLC's Learning for Justice resources[48] – in our nation's schools. Especially in these divided and polarized times, every elementary and secondary school should promote an inclusive school climate and activities that celebrate our nation's diversity.

- Promote programs and processes that intervene ethically in the lives of individuals – often called "deradicalization" efforts. The SPLC has partnered with American University's Polarization and Extremism Research and Innovation Lab (PERIL)[49] to help parents and caregivers understand how extremists exploit online spaces by targeting children and young adults with propaganda. In our guide – "Building Resilience & Confronting Risk in

---

[47] Sam Biddle, *The Intercept,* "Revealed: Facebook's Secret Blacklist of Dangerous Individuals and Organizations," October 12, 2021.
[48] https://www.learningforjustice.org/
[49] Polarization and Extremism Research and Innovation Lab (PERIL), American University, https://www.american.edu/centers/university-excellence/peril.cfm.

the COVID-19 Era"[50] – we provide tangible steps to counter the threat of online radicalization, including information on new risks, how to recognize warning signs, and how to get help and engage a radicalized child or young adult.

Conclusion

Mainstream social media companies, among the most trafficked websites on Earth – reaching billions of people – demonstrably tilt vulnerable consumers to embrace extreme far-right views.[51] These billion-dollar corporations shaped the atmosphere on January 6, 2021, when Trump supporters launched an unprecedented attack on the Capitol building in Washington D.C. Twitter, which has plucked far-right disinformation posters from obscurity and turned them into celebrities through its verification system and amplification algorithm, must enforce its own terms of service. Twitter should not refrain from doing so out of fear of public blowback. Lawmakers must also encourage platforms to act and hold them accountable. Threats to profitability alone will not change things. Despite public outrage over Facebook's failure to protect its users, the company's profits rose seventeen percent between July and September.[52] If we fail to adapt after a wakeup call like January 6, America risks plunging into the civil war sought by white nationalists and antigovernment extremists, threatening our democratic institutions. The future of our democracy depends upon how we meet the moment, a moment that demands both regulation and a substantial change in culture.

Thank you for holding this important hearing. We deeply appreciate the Committee's continued leadership in working to address domestic extremism in a constitutional and effective manner. We look forward to working with you as you continue your focus on this issue. If you have questions about this statement or need additional information, please contact Michael Lieberman, SPLC Senior Policy Counsel, at Michael.Lieberman@splcenter.org.

Sincerely,

Susan Corke

Susan Corke
Director
Intelligence Project

Michael Edison Hayden
Senior Investigative Reporter/Spokesperson
Intelligence Project

[50] "Building Resilience & Confronting Risk in the COVID-19 Era: A Parents & Caregivers Guide to Online Radicalization," Southern Poverty Law Center/American University's Polarization and Extremism Research and Innovation Lab (PERIL), https://www.splcenter.org/peril

[51] Semrush, "Top 100: The Most Visited Websites in the US," https://www.semrush.com/blog/most-visited-websites/, September, 2021.

[52] Mike Isaac, *New York Times*, "Facebook's profit jumps 17 percent," https://www.nytimes.com/2021/10/25/technology/facebook-profits-earnings-q3-2021.html, October 25, 2021.

Post-Hearing Questions for the Record
Submitted to the Honorable Karen Kornbluh
From Senator Kyrsten Sinema

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
October 28, 2021

1. A recent study showed that online misinformation is six times more likely to see user engagement than other content. What steps should users be taking now to not contribute to misinformation sharing and to protect themselves against falling prey to it?
    a. Certain groups, such as veterans, servicemembers and law enforcement, can be especially targeted for recruitment by extremist organizations. Are there additional steps that need to be taken for specific populations who might be more often be targeted to help them identify and not fall prey to disinformation and extremist or harmful content?

The fact that online misinformation receives increased engagement when compared to other types of content underscores the urgent need to scrutinize social media amplification. The burden lies with platforms to alter their processes—including how their algorithms amplify incendiary content—but also to provide more information to users so that they better understand who is paying for an ad or funding an aggregator outlet. In the meantime, there are steps that users can take now to avoid sharing or engaging with such content.

First, users can scrutinize the source of the information they are consuming. While this information may not be available as a label on the content or in a database on the social media site, third-party organizations, such as news rating service NewsGuard, make information available on outlets including whether they repeatedly publish verifiably false information or gather and present information responsibly. This process of active examination is key: academic research suggests that individuals are more likely to believe misinformation when they fail to carefully evaluate the material, regardless of whether it aligns with an individual's political preferences.[1]

Second, according to a study in the magazine *Nature*, six "degrees of manipulation"—impersonation, conspiracy, emotion, polarization, discrediting, and trolling—are used to spread misinformation and disinformation.[2] Resources that help individual users defend themselves from such misinformation tactics exist online, and are backed by empirical results. For example, researchers developed "Bad News," a gamified intervention that educates users on the six common misinformation techniques and demonstrates ways to distinguish between real and fake

[1] Bence Bago, David G. Rand, and Gordon Pennycook, "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines," Journal of Experimental Psychology: General 149, no. 8 (2020): 1608-1613.
[2] Jon Roozenbeek and Sander van der Linden, "Fake news game confers psychological resistance against online misinformation," Palgrave Communications 5, no. 65 (2019).

news headlines. Research reveals that playing the game once can increase users' ability to identify misinformation, although the immunizing effect diminishes after a set period of time.[3]

Third, psychological research suggests that "prebunking"—making users aware of false information before they are exposed to it online—can be especially effective. Recent studies have demonstrated that prebunking can counteract misinformation on such polarizing topics as climate change and vaccines.[4]

Finally, the Department of Veterans' Affairs, the congressionally-chartered Veterans Service Organizations, and police departments can implement proactive education about extremist targeting of veterans, servicemembers and law enforcement and warn them about sharing personal data and taking information they read online at face value—just as they should educate them about cybersecurity hygiene. The initiative of the Secretary of Defense to conduct a one-day stand-down to address extremism across the military is an important first step, but the need is ongoing.[5]

2. Your testimony highlighted the need for a digital code of conduct, one that is agreed to by the companies and enforced through Congressional oversight. Is this something social media companies can do now, or do they need an action from Congress to spur this effort?

A digital code of conduct would commit the platforms to make the design changes needed to enforce their terms of service. It might include using a "circuit breaker" mechanism like the one used on Wall Street to pause trading—to halt the spread of viral content while content moderators evaluate it or implementing know your customer rules for advertisers.[6] While regulatory oversight and enforcement seem warranted at this point, the platforms could commit to third-party monitoring with reports released to the public and providing data to researchers—so that the public and the Federal Trade Commission could hold them to whether they are measuring up to their public commitments.

---

[3] Rakoen Maertens, Jon Roozenbeek et al., "Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments," Journal of Experimental Psychology: Applied 27, no. 1 (2021): 1–16.
[4] Sander van der Linden, Anthony Leiserowitz et al., "Inoculating the Public against Misinformation about Climate Change," Global Challenges 1, no. 2 (2017); Daniel Jolly and Karen M. Douglas, "Prevention is better than cure: Addressing anti-vaccine conspiracy theories," Journal of Applied Psychology 47, no. 8 (2017): 459-469.
[5] Department of Defense, "Stand-Down to Address Extremism in the Ranks," February 5, 2021.
[6] Ellen Goodman, "Digital Information Fidelity and Friction," Knight First Amendment Institute, February 26, 2020.

Post-Hearing Questions for the Record
Submitted to David Sifry
From Senator Kyrsten Sinema

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
October 28, 2021

1. According to internal Facebook documents recently brought to light, we know that, after the election, Facebook ended a number of practices it had put in place to protect against election-related disinformation. We quickly saw the Stop the Steal hashtag take root. After January 6th, many of those protective policies were reinstated. It seems this is one example where more transparency and giving academics access to company data could help inform these companies and the public as to responsible practices that should be standard for social media companies. Can you expand on how granting access to researchers could improve company policies and practice?
    a. In the case of the example above, what specific information would you want to see to better understand how these policies impacted the spread of election-related disinformation?

There are a number of reasons why researcher access to anonymized company data would be beneficial. First, it allows for the better public understanding of what's actually going on, and the effects—both intended and unintended—of the algorithms, policies, and enforcement decisions made. This allows policy makers to make informed decisions about policy and regulation. Second, we believe that sunlight is the best disinfectant. When companies know that their data will be analyzed and reviewed, they will make different, more responsible decisions. For example, Facebook had a system called XCheck (pronounced "cross check") that created a pathway for important and influential users to avoid automated detection and removal decisions, and to ensure that their content was only reviewed by human moderators. While this may have been a temporary solution for Facebook to deal with error-prone AI systems, rather than fixing the problem, the system grew to encompass over 5.7 million users—far more than even the human reviewers could cope with. So, this meant that Facebook's entire content moderation system became subverted—and worse yet, that it would allow its most influential and popular users to skirt its rules with impunity, as there weren't enough human moderators to review reports around rule violations.

Crosscheck was never publicly discussed or disclosed by anyone at Facebook until internal documents were leaked a few months ago. We would have never known about this "shadow" policy and enforcement system if it weren't for a courageous whistleblower. But if we had regulated transparency reporting and researcher access, this would come to light—and Facebook would have likely made different resourcing decisions.

We need ***data about content policies and enforcement.*** *We need specific data about* ***high visibility content*** *(source accounts, reach, impressions, e.g., through a public archive). We*

also need data about ads and other paid content as well as around harmful algorithmic practices.

Ultimately, however, we need comprehensive data. ***Platforms are currently giving researchers and the public tiny slivers of data****, if any at all, which do not allow for a full picture of what's happening on their platforms.*

2. When I think of positive examples of granting researchers access to private company databases, I immediately think of like vulnerability detection or bug bounty programs for companies concerned about their cybersecurity. These are an industry best practice and broadly used. What models can we look to as we discuss granting researchers access to company data?

In August 2021, ADL Belfer Fellow Laura Edelson, an NYU PhD candidate and researcher, who was analyzing political advertising on Facebook from the platform's Ad Library and from CrowdTangle, a research and data collection tool, was deplatformed from Facebook hours after the platform learned she and her team were studying the company's role in spreading disinformation related to the January 6th insurrection. We need laws to ensure researchers like Ms. Edelson have the opportunity to engage in complex research and inform the public about the way social media advertising impacts society. We desperately need more transparency from platforms.

ADL supports the Social Media DATA Act and other reform measures that:

- *Increase social media data access to academic researchers*
- *Set uniform standards for what data must be accessible*
- *Authorize the FTC to oversee and ensure research is consistent with consumer privacy.*

3. Certain platforms have indicated an interest in creating a platform for kids. The Facebook Files highlighted internal studies about the negative impact of Instagram on kids, particularly teen girls. Given that we don't have visibility to the algorithms guiding the current platforms, I am concerned about the impact new products targeted at kids could have. What actions should Congress and other leaders take now to protect kids from the potential negative impact of these platforms?

   The leaked documents from whistleblower Frances Haugen show that Instagram has a negative impact on kids and teenagers—especially teenage girls and their body image. The platform knew this information but continued to build Instagram Kids, only shelving the project when the public found out about the alarming findings.

   It's important to consider social media platforms beyond Instagram. ADL has spent a lot of time focusing on online gaming spaces, as they are incredibly popular among both adults and young people. In September 2021, ADL's Center for Technology and Society released a first-of-its-kind survey on the online gaming experiences for youth:

- *60% of children ages 13-17 have experienced harassment while playing games online*
- *10% of young people ages 13-17 reported being exposed to discussion in online multiplayer games around white supremacist ideology*

We need better protections and more guardrails on how social media platforms target young people…especially those individuals who are harassed, harmed or susceptible to extremist recruitment.

For example, Omegle is a platform used by many kids/ teens, but ADL has identified numerous extremists who were using Omegle and similar sites, such as Monkey.cool and Chatroulette to take advantage of kids, troll and harass people of color, and other marginalized communities; or to try to introduce users to their extremism beliefs. ADL released an educational explainer that covered "what young people, parents and families should know about Omegle."

Congress needs to empower the FTC to hold these platforms accountable. Congress also needs to pass laws that increase protections for young people. Additionally, we need transparency measures to better understand and mitigate social media's harmful impact on kids.

The FTC should (1) initiate rulemaking and take other appropriate actions to regulate unfair and deceptive commercial data practices; (2) create an Office of Civil Rights; and (3) commit greater resources to aggressively enforce against unfair and deceptive practices, including those that involve children. The FTC must use all tools at its disposal to mitigate the threat of online hate.

ADL and over 20 national advocacy organizations sent a letter to the FTC urging the above. FTC must engage in increased action to regulate unfair and deceptive practices and increase enforcement against Big Tech.

4. During the COVID-19 pandemic, we have seen multiple examples of the spread of misinformation on social media platforms about COVID vaccines and also unproven treatments. And while companies like Facebook tried to address this, they weren't very successful. Particularly, much of this content spread through comments. What lessons should these challenges teach us and the companies about how to prevent the spreading of misinformation?

There are design practices that the industry can and should adopt to reduce the spread of misinformation and hate. Earlier this year, ADL's Center for Technology and Society released the Social Pattern Library, a collection of design principles and patterns that takes an anti-hate by design approach to reduce the levels of spread of mis- and disinformation.

For example: it recommends platforms to put in product features like Nudges which encourage users to pause before posting.

Of note, YouTube and Twitter have implemented Nudges after ADL's recommendation, and seen decreases in hate across the board.

Research shows that adding this kind of friction slows the hate and misinformation spread, while maintaining users' civil liberties.

We would see a much healthier and safer internet if platforms implemented Anti-Hate-By - Design practices, but ultimately without incentives to do so, we will never see change on a big enough scale.

We're glad to see efforts like the DETOUR Act consider the harmful way dark patterns exploit peoples' natural instincts so platforms can get more of our data or push users to engage with viral content.

5. A recent study showed that online misinformation is six times more likely to see user engagement than other content. What steps should users be taking now to not contribute to misinformation sharing and to protect themselves against falling prey to it?
    a. Certain groups, such as veterans, servicemembers and law enforcement, can be especially targeted for recruitment by extremist organizations. Are there additional steps that need to be taken for specific populations who might be more often be targeted to help them identify and not fall prey to disinformation and extremist or harmful content?

The persistent presence and amplification of hate, bigotry, and conspiracy theories on social media platforms has created an environment for extremism to flourish. Today, extremists are enmeshed in online communities where content designed to increase their propensity for hatred and violence often circulates freely. Extremist content boomerangs from fringe websites to mainstream platforms—in part because of social media's immense power, amplification of "engaging" content, and sophisticated recommendation algorithms.

We need look no further than the deadly insurrection at our Capitol which civil society organizations like ADL have called "the most predictable terror attack in American history," because it was planned and promoted mainstream platforms such as Facebook, Twitter, Instagram, YouTube, and Reddit, as well as fringe platforms such as Parler, Gab, 4Chan and Telegram. What was once on the fringes of our society has now been normalized in our newsfeeds, and incited people to action. It's not just members of extremist groups that commit acts of violence. According to ADL research over 70% of those arrested for storming the Capitol on January 6th were not card-carrying extremists.

There is no one-size-fits-all solution, and we recommend a whole-of-government, whole-of-society approach to solving this problem. Along with the legislative and regulatory proposals I've already described in previous answers, there are some things that everyday folks like you and I can do to contribute to reducing the spread of hate and misinformation.

First, we must have resources to increase our media literacy- when we see an inflammatory or shocking piece of content or a claim that seems like it could be too good to be true, pause, and engage critical thinking skills. We should ask ourselves, "Who created this post or message, and why? Does this come from a media source that I trust? Is this backed by research or an established journalistic organization, or is it anecdotal? Is this an ad? Is it being promoted to me? Who is the intended audience, and what is the context?"

Even taking a few moments to ask ourselves these kinds of questions before sharing, commenting, or reacting will reduce the spread of misinformation and other inflammatory content, and, along with reducing the incentives for platforms to promote and amplify this kind of content, will help to reduce its reach and spread.

Post-Hearing Questions for the Record
Submitted to Dr. Cathy O'Neil
From Senator Kyrsten Sinema

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
October 28, 2021

1. According to internal Facebook documents recently brought to light, we know that, after the election, Facebook ended a number of practices it had put in place to protect against election-related disinformation. We quickly saw the Stop the Steal hashtag take root. After January 6th, many of those protective policies were reinstated. It seems this is one example where more transparency and giving academics access to company data could help inform these companies and the public as to responsible practices that should be standard for social media companies. Can you expand on how granting access to researchers could improve company policies and practice?
    a. In the case of the example above, what specific information would you want to see to better understand how these policies impacted the spread of election-related disinformation?

2. When I think of positive examples of granting researchers access to private company databases, I immediately think of like vulnerability detection or bug bounty programs for companies concerned about their cybersecurity. These are an industry best practice and broadly used. What models can we look to as we discuss granting researchers access to company data?

3. Certain platforms, such as Instagram, have indicated an interest in creating a platform for kids. The Facebook Files highlighted internal studies about the negative impact of Instagram on kids, particularly teen girls. Given that we don't have visibility to the algorithms guiding the current platforms, I am concerned about the impact new products targeted at kids could have. What actions should Congress and other leaders take now to protect kids from the potential negative impact of these platforms?

4. During the COVID-19 pandemic, we have seen multiple examples of the spread of misinformation on social media platforms about COVID vaccines and also unproven treatments. And while companies like Facebook tried to address this, they weren't very successful. Particularly, much of this content spread through comments. How did the spread of COVID misinformation, particularly through likes and comments, happen, and what should it teach us about how these companies deploy these algorithms moving forward?

The witness failed to respond to these questions at time of printing. If responses are received, they will be on file in the committee offices for public inspection.

Supplemental Testimony of Professor Nathaniel Persily[1]

Responses to Questions Submitted by Senate Committee on Homeland Security and Governmental Affairs Following Hearing on "Social Media Platforms and the Amplification of Domestic Extremism and Other Harmful Content"

Submitted December 16, 2021

[1] James B. McClatchy Professor of Law, Co-Director of the Stanford Cyber Policy Center, Stanford Law School. Affiliation for identification purposes only; appearing in personal capacity.

Post-Hearing Questions for the Record
Submitted to Dr. Nathaniel Persily
From Senator Kyrsten Sinema

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
October 28, 2021

1. *According to internal Facebook documents recently brought to light, we know that, after the election, Facebook ended a number of practices it had put in place to protect against election-related disinformation. We quickly saw the Stop the Steal hashtag take root. After January 6th, many of those protective policies were reinstated. It seems this is one example where more transparency and giving academics access to company data could help inform these companies and the public as to responsible practices that should be standard for social media companies. Can you expand on how granting access to researchers could improve company policies and practice?*
    a. *In the case of the example above, what specific information would you want to see to better understand how these policies impacted the spread of election-related disinformation?*

Researcher access to data can help the platforms improve their polices, inform governments seeking to craft regulations, and answer critical questions for the public about the nature of the online information ecosystem. As the Frances Haugen revelations depict, the platforms often undertake intensive research in order to improve their products and to investigate safety challenges. The questions the platforms ask are not necessarily the same as those of interest to outsiders, however. Indeed, sometimes the platforms may have a disincentive to study certain problems, for fear of what they might learn. In the wake of the recent revelations, moreover, platforms may worry about conducting open-ended research, if they need to avert to the risk that a whistleblower might later leak preliminary or unpopular findings.

With that all said, there are a number of areas in which outside research might help improve platform policies. Most importantly, if the Congress passes the Platform Accountability and Transparency Act, outside researchers would be able to conduct cross platform research in ways that they have not been able to heretofore. Outside researchers will then be in a better position than internal data scientists to evaluate harms and interventions, as they will be able to compare different measures that have been tried by different platforms. As it stands, researchers at individual platforms are often siloed in their perspective as their insights are often limited to those garnered by analysis of their own platform's data.

Even with respect to a single platform, such as Facebook, outsider access can help investigate and facilitate greater understanding of online harms, such as the Stop the Steal movement described in the question. As a general rule when it comes to social media research, researchers need to know "who" saw/engaged with "what", "when" and "why." In other words, we need to understand which categories of users are seeing which content at a given time and then, if possible, why they might be seeing it.

With the Stop the Steal movement, platform data could help outside researchers understand the pathway to radicalization for some users. One of the fundamental questions with respect to many radical online groups – whether Stop the Steal, anti-vaccine, or QAnon groups – concerns the degree to which the platforms' affordances facilitate persuasion versus radicalization. Do people come to the platform searching for problematic content and groups? Or rather, do they fall into those groups because the recommendation algorithms lead them there or their friend networks repeatedly feed them recruitment content from these groups? Of course, some users progress along each pathway, but understanding which pathway predominates is critical to designing appropriate interventions to thwart the growth of these groups. Finally, data from the platform would allow us to measure the popularity over time of a given group committed to disinformation and to assess whether events, leadership, or other factors were most successful in promoting the group and its content.

2. When I think of positive examples of granting researchers access to private company databases, I immediately think of like vulnerability detection or bug bounty programs for companies concerned about their cybersecurity. These are an industry best practice and broadly used. What models can we look to as we discuss granting researchers access to company data?

Several different examples of researcher data access can serve as models for a program of access to social media data. First, we can look to existing attempts to work with the companies. Second, we can examine models for access to sensitive data held by the government. Third, we can model such efforts on other examples from industry of successful data-sharing programs.

Social Science One, which I founded with Harvard Professor Gary King, represented the most ambitious effort to date to try to develop a data-sharing system with one of the social media companies (in this case, Facebook). We worked with the Social Science Research Council to develop a system in which researchers would apply for both data access and funds. They would then be given access to datasets that we helped develop with Facebook researchers. The effort produced one of the largest social media datasets in existence, and was particularly successful in developing a "Research Data Agreement" between universities and Facebook, which can serve as a model for other companies going forward. However, the project was hobbled from its inception by Facebook's interpretation of the relevant privacy restrictions (either GDPR or the FTC Consent Decree under which they had been operating since the Cambridge Analytica Scandal). These interpretations restricted the kinds of data Facebook would make available and then led to the addition of statistical noise (through differential privacy) that made the resulting datasets difficult to analyze. Finally, the URLs dataset was limited to URLs that had been shared about 100 times and, we later learned, that through an error a large share of the U.S. population was left out of the dataset.

With respect to government datasets, there are often established procedures for researcher access. For individual level census data, for example, the Census Bureau has created a system of "Special Sworn Researchers"[2] that other agencies also employ for controlling which outsiders

[2] See U.S. Census Bureau, Frequently Asked Questions about Accessing Data in the Data Linkage Infrastructure, Sept. 8, 2016, https://www2.census.gov/about/linkage/faqs.pdf .

get access to government held data. In addition, through a network of Federal Statistical Data Research Centers, outside researchers can access census and other data in secure environments.

Finally, with respect to other industry-held datasets, there are some models from health care and the pharmaceutical industry, as well as financial institutions and credit card companies. On health care I would note, as an example, the Health Care Cost Institute (healthcostinstitute.org) which gathers together both public and private firm data to generate datasets that outside researchers can analyze. Mastercard, as another example, has provided Harvard researchers with controlled access to anonymized and aggregated data.[3] Both the Carnegie Endowment for International Peace[4] and the NetGain Partnership[5] have produced important documents reviewing available models for researcher access to industry data. A panel of the National Academy of Sciences has recently investigated these other models of industry-academic partnerships for its multi-workshop project: "Towards a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century." Their most recent hearing is available at https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a#sectionEventMaterials .

3. *A recent study showed that online misinformation is six times more likely to see user engagement than other content. What steps should users be taking now to not contribute to misinformation sharing and to protect themselves against falling prey to it?*

With rising concern about the health of the information ecosystem, experts have been exploring how to build resilience in the user population. Digital literacy proposals with that goal take many forms, from K-12 curricula to civil society proposals for the general population. The Stanford History Education Group, under the leadership of Professor Sam Wineburg, has developed a curriculum for "Civic Online Reasoning."[6] The tools they have developed attempt to teach students of all ages how to fact check and to think critically about the information they encounter online.

The platforms, too, have tried different tactics to slow down the spread of misinformation. Reliance on professional fact checkers (and labeling and demotion of misinformation) are the

[3] See Michelle Coscia, Frank Neffke, & Ricardo Hausmann, Center for Inclusive Growth, Datasets Access Request, https://www.mastercardcenter.org/forms/harvard-data-set .
[4] Jacob N. Shapiro, Natalie Thompson, & Alicia Wanless, Research Collaboration on Influence Operations Between Industry and Academia: A Way Forward, Dec. 2020, https://carnegieendowment.org/files/Shapiro_Thompson_Wanless_Instantiating_Models_final.pdf ; Jacob N. Shapiro, Michelle Nedashkovskaya, and Jan Oledan, "Collaborative Models for Understanding Influence Operations: Lessons From Defense Research," Carnegie Endowment for International Peace, June 25, 2020, https://carnegieendowment.org/2020/06/25/collaborative-models-for-understandinginfluence-operations-lessons-from-defense-research-pub-82150 .
[5] Elizabeth Hansen Shapiro, Michael Sugarman, Fernando Bermejo, Ethan Zuckerman, New Approaches to Platform Data Research, Feb. 2021, https://drive.google.com/file/d/1bPsMbaBXAROUYVesaN3dCtfaZpXZgI0x/view .
[6] See Stanford History Education Group, Civic Online Reasoning, at cor.stanford.edu.

most frequently cited examples, but the platforms have also tried to trigger critical thinking among users as they encounter disinformation. Facebook and Twitter, for example, have limited (or added friction) to the ability to forward content without having read it oneself, on the theory that mindless, reflexive forwarding is often how low-quality content gets spread. They have also added warnings or priming messages in certain circumstances to encourage users to think critically about the information they encounter on the platform.

I should add one word of caution about all such efforts described here. As dangerous as disinformation is, unwarranted skepticism can be equally pernicious. The more we caution people about the unreliability of information they encounter online, the greater the risk that they will increasingly distrust true information. Even under the best of circumstances, most users will not do research for every news item they come across online. By continually hammering on the theme that nothing one sees online should be trusted, we play into the strategy of bad actors (both foreign and domestic) who seek to undermine all legitimate media institutions and to paralyze the body politic with skepticism of all online information. One of the reasons we need greater access to platform data is to get a better sense of the scale and character of the disinformation problem, and to evaluate platform interventions that have been tested to confront it.

a. Certain groups, such as veterans, servicemembers and law enforcement, can be especially targeted for recruitment by extremist organizations. Are there additional steps that need to be taken for specific populations who might be more often targeted to help them identify and not fall prey to disinformation and extremist or harmful content?

The problem of harmful online content is not a single problem. Different measures should be taken to deal with disinformation, hate speech, and incitement, let alone bullying, graphic depictions of violence or encouragement of self-harm. Strategies that might work for adults might not work for kids; strategies that might work in English or in Western countries, might not work in other languages or other cultures. Moreover, people seeking out harmful content are in a different position than those who see the content because an algorithm fed it to them.

The Redirect Method,[7] pioneered by Jigsaw, represents one effective way to intervene to disrupt the pathway to radicalization. The basic strategy of the Redirect Method is to shift people's attention away from the problematic content they are seeking, but to do so in a way that does not produce a backfire effect that might reinforce their extremist beliefs. The method need not be applied to the disinformation or harmful speech problem, in general. First used to deal with terrorist recruitment, the method can be targeted to particular types of content and particular communities to shift them off the pathway down a potentially dangerous rabbit hole.

With respect to particular communities like the ones mentioned in the question, one key strategy is to get trusted members of those communities to speak out and provide content that confronts the lies or hate that others are promoting. Generic preaching is unlikely to persuade someone away from problematic content. Rather, trusted advocates from the user's own "tribe" will be

[7] See Moonshot, The Redirect Method, https://moonshotteam.com/redirect-method/ .

most effective in countering content received from other members of the tribe or elites seeking to manipulate them.

4. *While extremist organizations may be recruiting on large social media platforms, they are retreating to private servers to plan, coordinate and share propaganda. These private sites feature encrypted chat and resources on tactical training and firearm drills similar to what is taught in our military. What can be done to deter the use of these sites and servers for such purposes?*

Indeed, the rise of encrypted messaging presents a "wicked problem." There is a point at which the modes of encrypted digital communication begin to look more like telephone conversations, rather than television broadcasts or newspaper publications. As with any law enforcement effort to get access to and disrupt plots formulated through conversations using some technology, though, established procedures exist for wiretapping (or its digital analog) under conditions when there is probable cause to believe a crime is occurring. The rise of encryption will make it even more difficult to disrupt the kinds of networks described in the question, though. There may be some interventions that could be designed at different levels of the internet stack to ascertain when such activity may be occurring. However, unless we want the government or the platforms to have access to all such conversations, it is difficult to create a process that would allow for the dissolution of encryption just in cases of conspiracies like those mentioned.

Post-Hearing Questions for the Record
Submitted to Dr. Nathaniel Persily
From Senator Maggie Hassan

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"

October 28, 2021

1.) Louis Barclay, a United Kingdom-based developer, wrote about his interaction with Facebook following the company's threat of legal action against him over a software tool that he wrote.

The tool was designed to allow users to remove the News Feed, the scrolling list of updates, pictures, and ads, from their Facebook user experience. The curated and targeted content Facebook serves up through the News Feed is designed to extend a user's engagement with the platform as long as possible, and is a key revenue generator for Facebook. Naturally, Facebook has a significant financial interest in ensuring an uninterrupted flow of content, interspersed with ads, is consumed by a user for as long as possible.

We have seen research that extended exposure to the types of echo chambers pushed by Facebook and others harms individuals, society, and our democracy. Curated content, by algorithm or otherwise, is designed to keep users engaged, including by eliciting a strong emotional response. For that reason, research into how to mitigate those harms, protect users, and develop safer, less manipulative social media systems is important.

Mr. Barclay shared the tool that he developed with academic researchers and the public, and Facebook responded with the threat of litigation and shut down the tool on the grounds that it violated the company's terms of service.

a.) How have you seen companies build their terms of service to, in effect, deny academic or other inquiries into the safety of their platforms?

b.) How are companies using terms of service to shut down research into their product's algorithms, effects on customers, business models, customer networks, propagation of information among users, or other lines of research?

Excerpt from Elizabeth Hansen Shapiro, Michael Sugarman, Fernando Bermejo, Ethan Zuckerman, New Approaches to Platform Data Research, Feb. 2021, https://drive.google.com/file/d/1bPsMbaBXAROUYVesaN3dCtfaZpXZgI0x/view, p. 43:

Platforms' Terms of Service agreements are the focal point for controversies over the appropriate uses of platform data. Terms of service for products like platform API's stipulate the uses to which platform data can and cannot be put. These terms apply as well to issues like the quantity of data that can be collected, the number of authorized accounts that can access such data, the types of analyses that can be run, the handling of potentially identifiable information, and the handling of illegal or deleted content.

Variety of Researcher Orientations. The researchers, journalists, and activists we spoke with varied in their approaches to respecting platform terms of service in their work. Some (especially those who employ scraping techniques) take a directly adversarial approach, circumventing TOS entirely and making arguments for the social and civic value of the knowledge they generate with the data. Other researchers acknowledge the limitations set by platform terms of service and shape their research questions and data collection to fit both what data are legitimately available and the uses to which such data can be put according to terms of services. Some subcommunities of researchers, particularly those in the field of internet studies, have developed their own ethical frameworks for the collection and use of social media data which shape their choices about when, whether, and how to break specific terms of service.

Lax Enforcement Creates Grey Areas. We found an intriguing and problematic dynamic operating between platform companies and researchers when it comes to enforcement of violations of terms of service. On the one hand, some platform terms of service — for example the Twitter's terms of service (TOS) — explicitly prohibit the kinds of analysis that are routinely performed in the study of certain phenomena (particularly bot detection) linked to misinformation. We heard that while Twitter will sometimes reach out to researchers to inform them that their work has violated TOS, Twitter has not systematically pursued violations in a way that would set particular precedents or give any teeth to the parts of its TOS that researchers routinely break. This has created a dynamic in which researchers feel free to respect some terms while breaking others, knowing Twitter will mostly turn a blind eye. From a field-level perspective, this grey area of rule-breaking and lax enforcement has prohibited a more consolidated approach on both sides to carving out research exemptions in the context of Terms of Service agreements themselves. Explained one researcher,

> "Now, everybody knows that nothing is really going to happen, but it creates a dynamic where, instead of agreeing that we will respect TOS, there's this murky taboo where

> people have an understanding that some TOS are really important to respect — those that have to do with security and privacy — and there are some TOS that are kind of a joke. I don't think that's a good dynamic because I do think that terms of services should chart the path forward for how to do this type of research ethically and responsibly."

(For assistance with answers to this question I have relied on Laura Edelson, a Ph.D. candidate at NYU Tandon School of Engineering, whose Facebook account (famously) was suspended because of her distribution of a research-related plug-in that Facebook deemed a violation of its terms of service. She generously contributed what is written below.)

Many of the companies' terms of service forbid the use of the basic tools of digital investigation — namely, collecting information in an automated fashion and using temporary research accounts. These tools are critical to independent research into the platforms, because they enable the collection of information at a scale that allows meaningful insights to be drawn, and because they enable researchers to probe how the companies' algorithms respond to different prompts.

There are good reasons for preventing the use of these tools by data aggregators, commercial actors, and intelligence services, but the companies tend not to distinguish between public-interest uses of these tools and nefarious ones. For example, Facebook issued cease-and-desist letters in an effort to shut down Gizmodo's tool to study the People You May Know recommendation algorithm[8]; Louis Barclay's tool mentioned above, which was being used by a university to study the addictiveness of Facebook's News Feed;[9] and of course the Ad Observer plugin, which enlists Facebook volunteers to study the flow of disinformation through advertisements.[10] In 2018, the Knight First Amendment Institute proposed to Facebook that it amend its terms of service to authorize public-interest investigations that respect user privacy, but the company has not moved forward in that direction.[11]

The saga involving the NYU Ad Observer has shined a light on the problem of platform terms of service inhibiting research. In 2020, NYU's Ad Observer browser plug-in project, led by Damon McCoy and Laura Edelson, proceeded unimpeded by Facebook. The Ad Observer plug-in is a single-purpose tool whose "citizen scientist" volunteers have explicitly opted in to sharing advertising data with McCoy and Edelson to facilitate their research, and the data that Ad Observer enables volunteers to submit is stripped of all personally-identifying information.

---

[8] Surya Mattu & Kashmir Hill, Facebook Wanted Us to Kill This Investigative Tool, Gizmodo, Aug. 7, 2018, https://gizmodo.com/facebook-wanted-us-to-kill-this-investigative-tool-1826620111 .
[9] Louis Barclay, Facebook Banned Me for Life Because I Help People Use It Less, Slate, Oct. 7, 2021, https://slate.com/technology/2021/10/facebook-unfollow-everything-cease-desist.html .
[10] Laura Edelson & Damon McCoy, We Research Misinformation on Facebook. It Just Disabled Our Accounts, New York Times, Aug. 10, 2021 https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html .
[11] Knight First Amendment Institute at Columbia University, More than 200 Researchers Support Knighty Institute Call to Facilitate Research of Facebook's Platform, June 12, 2019, https://knightcolumbia.org/content/more-than-200-researchers-support-knight-institute-call-to-facilitate-research-of-facebooks-platform .

Weeks before the November 2020 election, Facebook demanded that NYU researchers McCoy and Edelson halt their work, stop publishing their data for other researchers to use, and destroy the records of data they had collected. Facebook made these demands four days after McCoy and Edelson, and their colleagues, published a lengthy critique concluding that Facebook's attempts at political advertising transparency were "theater." In their demand that McCoy and Edelson cease their work, Facebook asserted that McCoy's and Edelson's research violates Facebook's terms of service, and it threatened "enforcement action." Facebook did not take action at that time but ultimately terminated the Facebook accounts of McCoy, Edelson, and their colleague Paul Duke on August 3rd, 2021. In doing so, Facebook has shut down a critical tool needed to assess the spread of disinformation, especially as the 2022 election approaches.

[The Article that follows appeared as a report for the Brookings Institution and is available on its website at https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research/ . For a longer discussion of the Platform Accountability and Transparency Act, please also see the Lawfare podcast at the following link: https://www.lawfareblog.com/lawfare-podcast-free-data .]

BROOKINGS

Report

How to fix social media? Start with independent research.

Nathaniel Persily and Joshua A. Tucker Wednesday, December 1, 2021

We appear to have reached an inflection point when it comes to concern about the harms of social media and the willingness of governments to do something about them. The recent revelations by Facebook Whistleblower Frances Haugen have set off alarms around the world concerning everything from Instagram's effect on teen mental health to Facebook's responsibility for political violence. The revelations were explosive in their own right, but the reaction to them demonstrates how little outsiders know about what is happening inside these companies. The furor that has followed in the wake of these unprecedented disclosures makes it clear that outsider access to the data held by social media companies represents the critical first step to understanding the effect of these platforms on society and how regulation should be crafted to prevent harm.

When social media platforms first launched nearly two decades ago, they promised to bring people together and give the average person a megaphone to speak to the world. Social media enabled millions, and then billions, of people to connect and form communities online. They allowed masses of people to organize protest movements against the powerful, from the Arab Spring to Occupy Wall Street.[2] For the academic research community, these new platforms—which automatically generated a digital record of their users' behaviors—promised untold new opportunities to observe human behavior and attitudes, and, combined with technological advances around machine learning and AI, the opportunity for seemingly unlimited advancement of scientific knowledge.

This narrative in the United States shifted about five years ago, however, with growing awareness of the prevalence of mis- and dis-information online, as well as revelations of Russian attempts to use social media to manipulate the 2016 U.S. election and incite

social conflict. In the wake of the 2016 election, the media and Congress investigated how Russian actors exploited the affordances of social media platforms—everything from their advertising interfaces to closed groups to normal organic posts—to attack candidate Hillary Clinton in the U.S. presidential election and to propagate divisive messages concerning topics such as immigration, Muslim rights, religion, and gun policy. In addition, profit-oriented groups from Macedonia to California found ways to make money on these platforms by spreading disinformation and division. The pathologies attributed to social media have only multiplied since 2016, as Facebook, Twitter, and YouTube have been blamed for everything from polarization to COVID disinformation to anorexia to genocide.

As the earlier utopian prediction for social media turned decidedly pessimistic, research on these new technologies developed into a field of its own. However, because the platforms tightly controlled the data necessary to study these phenomena, academic researchers were limited in their efforts to get a handle on the scale, character, and causes of the various phenomena attributed to the rise of social media. To generate this broad new literature, researchers turned to surveys, experiments, browser plugins, scraping, and a host of other new methods to try to glimpse from the outside what firm insiders could easily see on the inside.

As difficult as it was before 2016, the Cambridge Analytica scandal further chilled any platform efforts to make it easy for outside researchers to gain access to individual-level content. That scandal involved a university researcher operating in his individual capacity harvesting "friend" data on Facebook and turning it over to a political consulting firm during the 2016 U.S. election campaign. As a result of the implications of these actions for user privacy, Facebook eventually paid a $5 billion fine to the Federal Trade Commission and shut down some of the APIs (automated programmer interfaces, or tools for efficiently downloading data directly without having to first render webpages in order to so) that academics had used for research. For any subsequent data-sharing effort at Facebook or another platform, the Cambridge Analytica scandal looms large and has created a presumption against open researcher access that might be exploited by bad actors or lead to leakage of private user data.

If the 2016 election and the Cambridge Analytica scandal represented a turning point in *public* concern about social media, the 2021 Haugen revelations seem to have focused *legislative* attention in the United States and around the world. Haugen has testified before parliamentary committees in several jurisdictions. Legislators in the U.S. and around the world have followed with additional hearings and various proposals to regulate social media algorithms, to remove the platform's legal immunity for user generated content, to break up Facebook and Google, and to regulate advertising. Unfortunately, there remains a real risk that legislation, particularly as it relates to content moderation, will be based on the snippets of data and research found in the recent document disclosures. To fill the void, Congress should mandate an unprecedented corporate data-sharing program to enable outside, independent researchers to conduct the kinds of analysis on social media platforms that firm insiders routinely perform.

The current barriers to social media research

Given the tremendous public interest in understanding social media's impact on the quality of American democracy, it is important to note that unlike the administrative (e.g., election results, economic indicators) or self-created (e.g., surveys, lab experiments) data that social scientists mined to understand political phenomena in the pre-internet age, some of the most important data related to political behavior is now locked up in a few large internet companies. As a result, there may be more politically relevant data than ever before, but a smaller share of it is now accessible to outside researchers. Researchers have deployed creative methods from the outside, but nothing can substitute for access to the raw data held by the firms themselves.

Researchers have tried for the last decade to get access to data from Facebook, Google/YouTube, and Twitter. Twitter has been the most open of the big three, in part because analogous privacy concerns do not arise on a platform where most tweets are public.[3] Facebook has experimented with several notable data sharing programs. Google/YouTube and TikTok have tended to be the most closed off to independent research. All of these firms will often bring in researchers for one or another project, sometimes even resulting in publication.[4] But when it comes to truly independent research, academics tend to be forced to come up with efforts from the outside.

The system of APIs set up by Twitter over the past decade—most of which were set up for business purposes as opposed to research—led to a flowering of academic research using Twitter data.[5] And to be very clear, Twitter deserves kudos for making so much data available, including specialized collections around Russian Internet Research Agency (IRA) trolls and COVID-19. However, even the data that Twitter makes accessible to outside researchers has left out information that is crucial for academic research, such as data about which users have seen which tweets, known as "exposure data," or even by how many users have seen each tweet. Therefore, understanding which tweets and associated news items reach which classes of people remains an area of inquiry that outsiders to Twitter cannot investigate. Moreover, some data, such as friends and follower networks, have become harder to collect at scale over time as new API access rules are rolled out with new rate limits governing how often researchers can access the API. Nevertheless, because more research has been performed on Twitter than any other platform, our understanding of the relationship of social media to online harms is highly biased toward what is occurring on that particular platform.

Facebook has had grand ambitions around sharing data. Both authors were involved with Social Science One, an effort to share Facebook data with academics in a secure, privacy-protected way. Privacy concerns derailed that effort from its inception, as Facebook later decided that it could only release a dataset at the "URL level," meaning no individual-level data would be accessible to academics, only information about exposure and engagement with URLs. Even at high levels of aggregation, Facebook still added noise to the data through methods of differential privacy that made the dataset difficult to use, and serious omissions were later found in that dataset. The dataset still exists as one of the largest ever made accessible to social scientists, but researchers have been slow to use it.

In the wake of the Social Science One difficulties, Facebook has embarked on a different model for a specific research project. To study the platform's impact on the 2020 election, a team of academics (co-led by one of the authors—Tucker) has worked with Facebook researchers to analyze data related to the 2020 U.S. election.[6] The partnership promises to make possible for the 2020 election the kind of research that was never done with respect to the 2016 election. Facebook has invested a significant amount of money and time from internal research teams in the project. Regardless of the importance and value

of this particular effort, the model—while replicable for other focused studies—is by design not something that is scalable to meet the ongoing needs of the larger research community.[7]

In contrast to these efforts, Google/YouTube and TikTok have not embarked on any dedicated data sharing program with academics. This has not stopped academics from attempting to study the workings of these platforms by, for example, seeing how and when links to YouTube videos spread on other platforms or by running experimental studies of what videos are recommended by the YouTube algorithm. YouTube does have an API that provides some information about videos that can be useful for research, although API limits can hinder the ability to conduct such research at scale. Google's trends feature also allows the public to get a sense of trends in searches, such as how often people search for a particular candidate's name, for example, but there is no analogous tool for YouTube.

In the end, independent academic researchers remain reliant on the kindness of platforms to make data available. Even access to supposedly public YouTube or Twitter data can be upended by changes in rules regarding what data is available through APIs, how much data researchers can collect via APIs and how fast they can do so, and whether existing APIs will be shut down (or "deprecated") . The problem, of course, is that platforms may feel that they have little incentive to share data with academic researchers. Merely making the data available could expose them to liability for violating applicable privacy rules. Moreover, independent publications based off of such research, in some instances, will put the platforms in a bad light. Therefore, the public good of greater access is usually not seen as outweighing the real legal risk of another Cambridge Analytica or the reputational risk of embarrassing findings.

The path forward for researcher access to platform data

To break through the logjam, we need federal legislation. That legislation, such as the Platform Transparency and Accountability Act proposed by one of the authors (Persily), could come in many forms but it should have three essential characteristics. First, a federal agency must be empowered to force the large internet platforms to share data akin to what firm insiders are able to access with outsiders not selected by the firm. Second,

that agency (perhaps working with a nongovernmental organization or another arm of the federal government, such as the National Science Foundation) should vet researchers and research projects that will be given access to platform data.[8] Third, data should reside at the firm, and regulations should specify in detail the process for accessing data and publishing results in a way that would not endanger user privacy.

Apart from these three critical features, there are several different paths that legislation and regulation could take. The relevant enforcement body could be the Federal Trade Commission, given that it has been out front in dealing both with fraud and user privacy, or a wholly new government agency. The researchers could be limited to academics,[9] or they could be expanded to other groups such as journalists or think tanks if those groups could be adequately defined by law. The universe of firms could be limited to the largest social media companies, such as Facebook, Google/YouTube, Twitter, and TikTok. Perhaps it could also be extended to other large technology companies, such as Amazon or Apple, or other critical companies in the internet stack. The type of research enabled by this law could be defined by its purposes (such as politics or mental health) or it could be expanded to all possible scientific questions that could be answerable with access to firm data. Finally, penalties both for non-compliant firms and researchers engaging in malfeasance could be significant. A platform's immunity under Section 230 of the Communications Decency Act could depend on providing researcher access, for example. Researchers, too, and the universities with which they are affiliated could be subject to extensive fines or other criminal punishment if they attempt to repeat another Cambridge Analytica.

If outsiders get access to firm data, it will have immediate effects on platform behavior and long-term effects on informing governmental policy. The mere fact that outsiders will have data access will lead the platforms to know that they are being watched. Like any person or institution that knows they cannot operate in secret, the platforms will know that their algorithms and content moderation policies will be perpetually under scrutiny. The resulting research will not only keep them in check but also help inform their interventions and policies going forward. Ideally, the emergence of a more open research ecosystem around platform data will also encourage platforms to share more of their internal research (e.g., the materials described in the Facebook Papers) publicly, as the idea that internal research must be kept private would become less appealing. At the same

time, research conducted by employees of the platforms might come to be seen as more credible by journalists, the scientific community, and policy makers in an environment where replication of such research is automatically possible by researchers who are not employees of the platform.

Independent research on platform data is a prerequisite to sound government policy. On the table right now are any number of legislative proposals dealing with privacy, antitrust, child welfare, and amendments to section 230 of the Communications Decency Act to make the platforms liable for user-generated content. For the most part, legislators are legislating in the dark—with faint light being cast by whistleblowers or well-spun public reports from the firms. Whether the issue is the harm of Instagram use on teen girls' health or the ubiquity of hate speech and disinformation, the conventional wisdom that has led to promotion of different policy interventions can only be evaluated with access to internal data. Likewise, outside research might also improve platform policies and even the products themselves. Outside researchers will ask questions that those tied to the profit-making mission of the company may not want to ask. But the results could help platforms better understand online harms and develop more targeted policies to address them.

Finally, even apart from knowledge gained about online harms—such as misinformation, hate speech, and online harassment—and platform policies, analysis of platform data is critical to understanding larger social and policy questions, such as the nature of the impact of social media on the quality of democracy or the impact of the platforms on mental health and wellbeing.[10] More and more of the human experience is taking place online. To understand fundamental aspects of the economy, politics, and society requires a better understanding of online behavior. Whether the topic is the effectiveness of COVID-19 interventions or the racial and gender biases of online marketplaces or the changing nature of the news media, platform data represents an ever-growing share of the data necessary to understand social phenomena and craft appropriate public policy responses.

Conclusion: Research access and the transparency agenda

Researcher access is only one component of a larger transparency agenda, and transparency is only one aspect of tech regulation. In addition to researcher access along a privacy-protected pathway described above, the platforms should make more information available to the public. We can envision a tiered system of transparency and data access in which the most sensitive data requires the kind of vetting and security measures described above. But other privacy-protected datasets could be made more widely accessible to outside researchers, including journalists and civil society groups. Finally, the platforms should be pushed to make publicly available tools and APIs, such as Google Trends or Crowdtangle, that will allow anyone to gain insights as to the magnitude of certain online phenomena.

Policy should also facilitate outside research efforts that do not depend on platform compliance. Sometimes the only way to check on the accuracy of platform-provided data is to deeply analyze what is publicly available. Researchers must therefore be protected when they develop "adversarial" methods to analyze platform data. In particular, we need to shield researchers from criminal and civil liability when they scrape publicly available websites.

We have reached a critical moment in the attention paid to digital communication and online harms and in the widespread recognition that answers to the relevant policy questions cannot be assessed without access to platform-controlled data. The current equilibrium is unsustainable. We cannot live in a world where the platforms know everything about us and we know next to nothing about them. We should not need to wait for whistleblowers to whistle before we can begin to understand all that is happening online.

Joshua A. Tucker is the co-Chair of the Independent Academic Research Team for the 2020 US Elections Facebook Research Project, a research collaboration with Facebook for which he is not receiving any monetary compensation.

The authors did not receive financial support in the last year from any firm or person for this article or from any firm or person with a financial or political interest in this article. The authors are not currently an officer, director, or board member of any organization with a financial or political interest in this article.

Footnotes

1. 1

   Persily is the James B. McClatchy Professor of Law at Stanford Law School and co-Director of the Stanford Cyber Policy Center. Tucker is Professor of Politics at New York University and the co-Director of NYU's Center for Social Media and Politics. They are the co-editors of Social Media and Democracy: The State of the Field and Prospects for Reform (Cambridge University Press, 2020).

2. 2 Tucker, Joshua A., Yannis Theocharis, Margaret E. Roberts, and Pablo Barberá. "From liberation to turmoil: Social media and democracy." *Journal of Democracy* 28, no. 4 (2017): 46-59.
3. 3

   Reddit, another platform on which posts are almost entirely public, has also been largely open to academic analysis thanks to the work of Jason Baumgartner and the pushshift.io website he set up to share Reddit data.

4. 4 The "sometimes" nature of publications highlights another problem revealed by the Facebook Papers, which is that most research conducted internally by the platforms will only make its way into the public domain if the platforms choose to release the research publicly. In academia, this is known as the "file drawer" problem, where less interesting (and often null) results fail to be published, and as a consequence the overall accumulation of knowledge is biased (see Franco, Annie, Neil Malhotra, and Gabor Simonovits. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345, no. 6203 (2014): 1502-1505). When we consider this from the perspective of for-profit corporations, the net result can be even more pernicious, which is that the overall accumulation of knowledge would likely be biased in the direction of research that puts the platforms in a better light. However, knowing the potential for such biases to exist should lead outside observers to discount such research accordingly, making knowledge accumulation that much more difficult. The exception here – and a possible path forward – are mechanisms by which the platforms can bind themselves *a priori* to share research *ex post*; we discuss one such mechanism below.
5. 5

   Tucker, Joshua Aaron, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan, "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature" (March 19, 2018). Available at SSRN: https://ssrn.com/abstract=3144139.

6. 6 https://medium.com/@2020_election_research_project/a-proposal-for-understanding-social-medias-impact-on-elections-4ca5b7aae10
7. 7

As of the time of this writing, Facebook has begun testing a new Research API with select groups of researchers that promises to make data available from "four buckets of real-time Facebook data: pages, groups, events and posts" although as of now the data is limited to public posts and to posts from the U.S. and EU (https://techcrunch.com/2021/11/15/facebooks-researcher-api-meta-academic-research/). As the API has not officially been launched yet, we hold off on any further commentary for now.

8. 8

The idea of the government vetting researchers for access to sensitive *administrative* data is of course not new; one example is the process known as "Special Sword Status" by which researchers can be certified to work with certain data at the U.S. Census Bureau that is not available to the general public.

9. 9 One of the reasons the Platform Transparency and Accountability Act is limited to academic researchers is that a university is an easily defined entity. They also have Institutional Review Boards (IRBs) that evaluate the impact of proposed research on human subjects. IRB approval is a necessary predicate for applying to get access to firm data under that proposal.
10. 10 Braghieri, Luca and Levy, Roee and Makarin, Alexey, Social Media and Mental Health (August 12, 2021). Available at SSRN: https://ssrn.com/abstract=3919760

Dr. Mary Anne Franks
Response to Post-Hearing Questions for the Record
From Senator Kyrsten Sinema
Dec. 17, 2021

"Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"
October 28, 2021

1. According to internal Facebook documents recently brought to light, we know that, after the election, Facebook ended a number of practices it had put in place to protect against election-related disinformation. We quickly saw the Stop the Steal hashtag take root. After January 6th, many of those protective policies were reinstated. It seems this is one example where more transparency and giving academics access to company data could help inform these companies and the public as to responsible practices that should be standard for social media companies. Can you expand on how granting access to researchers could improve company policies and practice?
    a. In the case of the example above, what specific information would you want to see to better understand how these policies impacted the spread of election-related disinformation?

Answer: I would defer to my colleague Professor Nate Persily (who also testified at this hearing) on this issue, as he is far more well-versed on it than I am, but I would emphasize that transparency efforts are (1. meaningless without an enforcement mechanism and (2. run the risk of legitimizing unethical and harmful practices. Regarding the first, while it would be immensely valuable to learn more about how platforms operate behind the scenes, how they make decisions that impact the public welfare, what data their policies are based on, what tools they have at their disposal to minimize harm, and so on, it is difficult to imagine why companies would feel compelled to turn this information over as long as they retain near-total immunity from liability under Section 230. The most direct and effective way to get transparency, in my view, is to limit Section 230 immunity so that individuals who have suffered foreseeable, avoidable, technology-facilitated harms can sue, which would force companies in some cases to turn over valuable information they would otherwise not be inclined to disclose and encouraging all companies to engage in safer practices to avoid litigation. As to the second, I want to sound a note of caution about academic access to internal research: academics are not invulnerable to regulatory capture, and Silicon Valley already funds a great deal of academic research into itself. There is a real danger that the tech industry will actually gain more legitimacy and social control by granting academic access to its restricted data.

2. When I think of positive examples of granting researchers access to private company databases, I immediately think of like vulnerability detection or bug bounty programs for companies concerned about their cybersecurity. These are an industry best practice and broadly used. What models can we look to as we discuss granting researchers access to company data?

Answer: I'm afraid this is also not in my area of expertise, but I would emphasize my concern above and recommend that any model used be sensitive to and provide countermeasures against the potential for industry capture and control.

3. In your testimony, you mentioned that the use of artificial intelligence (AI) will not be as effective as some of these companies have stated in identifying content that violates company policies or spreads disinformation and harmful content. What are the limitations of AI for these purposes?

Answer: I believe testimony about this specific subject was offered by a different expert on the panel with deep technical expertise in AI, which I am afraid I do not have. Based on the work I have done that touches on ethics in AI issues, however, I would offer the observation that AI is for many purposes just human beings multiplied. Machine learning cannot be used to escape the hard work of human judgment of complex issues such as what constitutes "disinformation" or "harm." And models based on biased data will produce biased results, whether the operation is being performed by a human or a machine.

4. Certain platforms have indicated an interest in creating a platform for kids. The Facebook Files highlighted internal studies about the negative impact of Instagram on kids, particularly teen girls. Given that we don't have visibility to the algorithms guiding the current platforms, I am concerned about the impact new products targeted at kids could have. What actions should Congress and other leaders take now to protect kids from the potential negative impact of these platforms?

Answer: The single most important step Congress can take to address this and other technology-facilitated harms is to end the special treatment afforded to the tech industry by Section 230. Right now, there is simply no incentive for tech companies to take care—indeed, there is no standard of care for the tech industry. If a toy company wants to make and market a product to children, it has to conform to a slew of product safety regulations before the product ever gets to the public. And if that product gets through those regulations and still causes injury or death, that company can be sued. But a tech company that wants to make and market an online product to minors can simply roll it out to them, untested, and if it contributes to injury or death, the company can invoke Section 230 and avoid suit. That has to change. My specific proposal to reform Section 230 can be found here.

5. A recent study showed that online misinformation is six times more likely to see user engagement than other content. What steps should users be taking now to not contribute to misinformation sharing and to protect themselves against falling prey to it?
    a. Certain groups, such as veterans, servicemembers and law enforcement, can be especially targeted for recruitment by extremist organizations. Are there additional steps that need to be taken for specific populations who might be more often be targeted to help them identify and not fall prey to disinformation and extremist or harmful content?

Answer: Media literacy programs, which teach people how to identify and resist false information, can be extremely valuable on this front. Finland has made impressive strides on this issue and could serve as a potential model for the United States. Tech companies themselves could do far more than they are currently doing to flag and filter false or misleading content, especially when it is targeted at vulnerable communities. But again, the question is how to incentivize the tech industry to invest in such practices, which comes back to the issue of reforming Section 230.

○