NEW AMERICA

March 17, 2020

It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge

Nathalie Maréchal & Ellery Roberts Biddle

Last edited on March 16, 2020 at 7:16 p.m. EDT

Acknowledgments

RDR Director Rebecca MacKinnon, Deputy Director Jessica Dheere, Research Director Amy Brouillette, Communications Manager Kate Krauss, Company Engagement Lead Jan Rydzak, and Communications Associate Hailey Choi also contributed to this report.

We wish to thank all the reviewers who provided their feedback on the report, including:

- Jen Daskal, Professor and Faculty Director, Tech, Law & Security Program, American University Washington College of Law
- Sharon Bradford Franklin, Policy Director, Open Technology Institute
- Brandi Geurkink, European Campaigner, Mozilla Foundation
- Spandana Singh, Policy Analyst, Open Technology Institute
- Joe Westby, Researcher, Technology and Human Rights, Amnesty International

Original art by Paweł Kuczyński.

We would also like to thank Craig Newmark Philanthropies for making this report possible.

For a full list of RDR funders and partners, please see: www.rankingdigitalrights.org/who/partners.

For more about RDR's vision, impact, and strategy see: www.rankingdigitalrights.org/about.

For more about New America, please visit www.newamerica.org.

For more about the Open Technology Institute, please visit www.newamerica.org/oti.

About the Author(s)

Nathalie Maréchal is a senior policy analyst at Ranking Digital Rights.

Ellery Roberts Biddle is a journalist and digital rights advocate.

About New America

We are dedicated to renewing America by continuing the quest to realize our nation's highest ideals, honestly confronting the challenges caused by rapid technological and social change, and seizing the opportunities those changes create.

About Open Technology Institute

OTI works at the intersection of technology and policy to ensure that every community has equitable access to digital technology and its benefits. We promote universal access to communications technologies that are both open and secure, using a multidisciplinary approach that brings together advocates, researchers, organizers, and innovators.

About Ranking Digital Rights

Ranking Digital Rights (RDR) works to promote freedom of expression and privacy on the internet by creating global standards and incentives for companies to respect and protect users' rights.

Contents

Executive Summary

Democracies are struggling to address disinformation, hate speech, extremism, and other problematic online speech without requiring levels of corporate censorship and surveillance that will violate users' free expression and privacy rights. This report argues that instead of seeking to hold digital platforms liable for content posted by their users, regulators and advocates should instead focus on holding companies accountable for how content is amplified and targeted. It's not just the content, but tech companies' surveillance-based business models that are distorting the public sphere and threatening democracy.

The first in a two-part series aimed at U.S. policymakers and anybody concerned with the question of how internet platforms should be regulated, this report by Ranking Digital Rights (RDR) draws on new findings from a study examining company policies and disclosures about targeted advertising and algorithmic systems to propose more effective approaches to curtailing the most problematic online speech. Armed with five years of research on corporate policies that affect online speech and privacy, we make a case for a set of policy measures that will protect free expression while holding digital platforms much more accountable for the effects of their business models on democratic discourse.

We begin by describing two types of algorithms—content shaping and content moderation algorithms—and how each can violate users' freedom of expression and information as well as their privacy. While these algorithms can improve user experience, their ultimate purpose is to generate profits for the companies by keeping users engaged, showing them more ads, and collecting more data about them—data that is then used to further refine targeting in an endless iterative process.

Next, we give examples of the types of content that have sparked calls for content moderation and outline the risks posed to free expression by deploying algorithms to filter, remove, or restrict content. We emphasize that because these tools are unable to understand context, intent, news value, and other factors that are important to consider when deciding whether a post or an advertisement should be taken down, regulating speech can only result in a perpetual game of whack-a-mole and, while damaging democracy, will not fix the internet.

We then describe the pitfalls of the targeted advertising business model, which relies on invasive data collection practices and black-box algorithmic systems to create detailed digital profiles. Such profiling not only results in unfair (and sometimes even illegal) discrimination, but enables any organization or person who is allowed to buy ads on the platform to target specific groups of people who share certain characteristics with manipulative and often misleading messages. The implications for society are compounded by companies' failure to conduct due diligence on the social impact of their targeted advertising systems, failure to

impose and enforce rules that prevent malicious targeted manipulation, and failure to disclose enough information to users so that they can understand who is influencing what content they see online.

Nowhere in the U.S. have the effects of targeted advertising and misinformation been felt more strongly than in recent election cycles. Our next section describes how we are only beginning to understand how powerful these systems can be in shaping our information environment—and in turn our politics. Despite efforts to shut down foreign-funded troll farms and external interference in online political discourse, we note that companies remain unacceptably opaque about how we can otherwise be influenced and by whom. This opacity makes it impossible to have an informed discussion about solutions, and how best to regulate the industry.

We conclude the report with a warning against recent proposals to revoke or dramatically revise Section 230 of the 1996 Communications Decency Act (CDA), which protects companies from liability for content posted by users on their platforms. We believe that such a move would undermine American democracy and global human rights for two reasons. First, algorithmic and human content moderation systems are prone to error and abuse. Second, removing content prior to posting requires pervasive corporate surveillance that could contribute to even more thorough profiling and targeting—and would be a potential goldmine for government surveillance. That said, we believe that in addition to strong privacy regulation, companies must take immediate steps to maximize transparency and accountability about their algorithmic systems and targeted advertising business models.

The steps listed below should be legally mandated if companies do not implement them voluntarily. Once regulators and the American public have a better understanding of what happens under the hood, we can have an informed debate about whether to regulate the algorithms themselves, and if so, how.

Key recommendations for corporate transparency

- Companies' rules governing content shared by users must be clearly explained and consistent with established human rights standards for freedom of expression.

- Companies should explain the purpose of their content-shaping algorithms and the variables that influence them so that users and the public can understand the forces that cause certain kinds of content to proliferate, and other kinds to disappear.

- Companies should enable users to decide whether and how algorithms can shape their online experience.

- Companies should be transparent about the ad targeting systems that determine who can pay to influence users.

- Companies should publish their rules governing advertising content (what can be advertised, how it can be displayed, what language and images are not permitted).

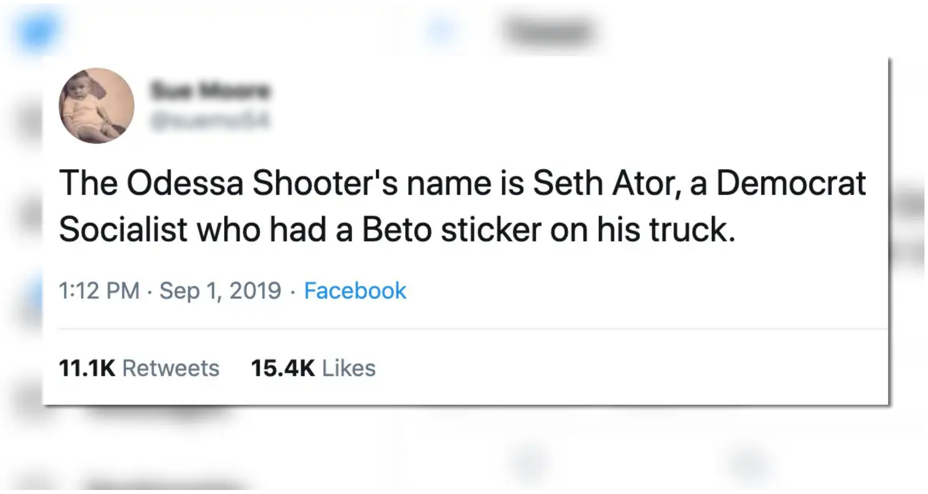
- All company rules governing paid and organic user-generated content must be enforced fairly according to a transparent process.

- People whose speech is restricted must have an opportunity to appeal.

- Companies must regularly publish transparency reports with detailed information about the outcomes of the steps taken to enforce the rules.

Introduction

Within minutes of the August 2019 mass shooting in Odessa, Texas, social media was ablaze with safety warnings and messages of panic about the attack. The shooter, later identified as Seth Ator, had opened fire from inside his car, and then continued his spree from a stolen postal service van, ultimately killing seven people and injuring 25. Not 30 minutes after the shooting, a tweet alleging that Ator was "a Democrat Socialist who had a Beto sticker on his truck," suggesting a connection between the shooter and former Rep. Beto O'Rourke (D-Texas), went viral. The post quickly spread to other social media sites, accompanied by a rear angle photo of a white minivan adorned with a Beto 2020 sticker. A cascade of speculative messages followed, alleging deeper ties between the shooter and O'Rourke, who was running for the Democratic presidential nomination at the time.



The Odessa Shooter's name is Seth Ator, a Democrat Socialist who had a Beto sticker on his truck.

1:12 PM · Sep 1, 2019 · Facebook

11.1K Retweets **15.4K** Likes

A screenshot of a tweet falsely alleging a connection between mass shooter Seth Ator and former U.S. presidential candidate Beto O'Rourke, which received more than 11,100 retweets (Twitter, Sept. 1, 2019).

But the post was a fabrication. The O'Rourke campaign swiftly spoke out, denying any link between the candidate and the shooter and condemning social media platforms for hosting messages promoting the false allegation. Local law enforcement confirmed that neither of the vehicles driven by Ator had Beto stickers on them, and that Ator was registered as "unaffiliated," undercutting follow-on social media messages that called Ator a "registered Democrat." Twitter temporarily suspended the account, but much of the damage had already been done.

Within just two days, the post became "Google's second-highest trending search query related to O'Rourke," according to the Associated Press.[1] Within a week, it had garnered 11,000 retweets and 15,000 favorites.[2]

What made this post go viral? We wish we knew.

One might ask why Twitter or other platforms didn't do more to stop these messages from spreading, as the O'Rourke campaign did. The incident raises questions about how these companies set and enforce their rules, what messages should or shouldn't be allowed, and the extent to which internet platforms should be held liable for users' behavior.

But we also must ask what is happening under the hood of these incredibly powerful platforms. Did this message circulate so far and fast thanks only to regular users sharing and re-sharing it? Or did social media platforms' algorithms —technical systems that can gauge a message's likely popularity, and sometimes cause it to go viral—exert some influence? The technology underlying social media activity is largely opaque to the public and policymakers, so we cannot know for sure (this is a major problem—which is discussed more later in the report). But research indicates that in a scenario like this one, algorithmic systems can drive the reach of a message by targeting it to people who are most likely to share it, and thus influence the viewpoints of thousands or even millions of people.[3]

As a society, we are facing a problem stemming not just from the existence of disinformation and violent or hateful speech on social media, but from the systems that make such speech spread to so many people. We know that when a piece of content goes viral, it may not be propelled by genuine user interest alone. Virality is often driven by corporate algorithms designed to prioritize views or clicks, in order to raise the visibility of content that appears to inspire user interest. Similarly, when a targeted ad reaches a certain voter, and influences how or whether they vote, it is rarely accidental. It is the result of sophisticated systems that can target very specific demographics of voters in a particular place.[4]

Why do companies manipulate what we see? Part of their goal is to improve user experience—if users had to navigate all the information that comes across our social media feeds with no curation, we would indeed be overwhelmed. But companies build algorithms not simply to respond to our interests and desires but to generate profit. The more we engage, the more data we give up. The more data they have, the more decisions they can make about what to show us, and the more money they make through targeted advertising, including political ads.[5]

Surveillance-based business models have driven the distortion of our information environment in ways that are bad for us individually and potentially catastrophic for democracy.

The predominant business models of the most powerful American internet platforms are surveillance-based. Built on a foundation of mass user-data collection and analysis, they are part of a market ecosystem that Harvard Professor Shoshana Zuboff has labeled surveillance capitalism.[6] Evidence suggests that surveillance-based business models have driven the distortion of our information environment in ways that are bad for us individually and potentially catastrophic for democracy.[7] Between algorithmically-induced filter bubbles, disinformation campaigns foreign and domestic, and political polarization exacerbated by targeted advertising, our digital quasi-public sphere has become harder to navigate and more harmful to the democratic process each year. Yet policymakers in the United States and abroad have been primarily focused on purging harmful content from social media platforms, rather than addressing the underlying technical infrastructure and business models that have created an online media ecosystem that is optimized for the convenience of advertisers rather than the needs of democracy.

Some foreign governments have responded to these kinds of content by directly regulating online speech and imposing censorship requirements on companies. In contrast, the First Amendment of the U.S. Constitution forbids the government from passing laws that ban most types of speech directly, and companies are largely free to set and enforce their own rules about what types of speech or behaviors are permitted on their platforms.[8] But as private rules and enforcement mechanisms have failed to curb the types of online extremism, hate speech, and disinformation that many Americans believe are threatening democracy, pressure is mounting on Congress to change the law and hold internet platforms directly liable for the consequences of speech appearing on their platforms.

Policymakers, civil society advocates, and tech executives alike have spent countless hours and resources developing and debating ways that companies could or should remove violent extremist content, reduce viral disinformation (often described as "coordinated inauthentic behavior") online, and minimize the effects of advertising practices that permit the intentional spread of falsehoods. Here too, technical solutions like algorithmic systems are often deployed. But efforts to identify and remove content that is either illegal or that violates a company's rules routinely lead to censorship that violates users' civil and political rights.

Today's technology is not capable of eliminating extremism and falsehood from the internet without stifling free expression to an unacceptable degree.

There will never be a perfect solution to these challenges, especially not at the scale at which the major tech platforms operate.[9] But we suspect that if they changed the systems that decide so much of what actually happens to our speech (paid and unpaid alike) once we post it online, companies could significantly reduce the problems that disinformation and hateful content often create.

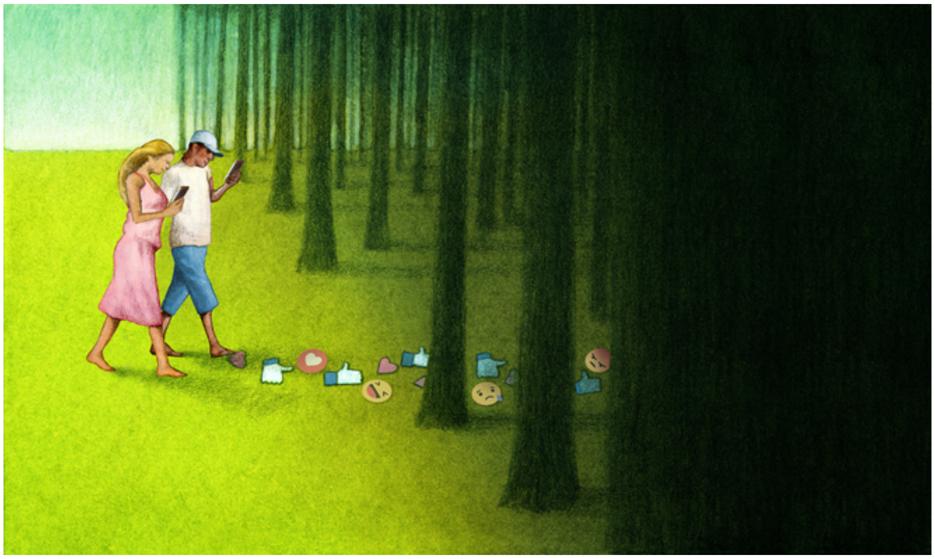
At the moment, determining exactly how to change these systems requires insight that only the platforms possess. Very little is publicly known about how these algorithmic systems work, yet the platforms know more about us each day, as they track our every move online and off. This information asymmetry impedes corporate accountability and effective governance.[10]

This report, the first in a two-part series, articulates the connection between surveillance-based business models and the health of democracy. Drawing from Ranking Digital Rights's extensive research on corporate policies and digital rights, we examine two overarching types of algorithms, give examples of how these technologies are used both to propagate and prohibit different forms of online speech (including targeted ads), and show how they can cause or catalyze social harm, particularly in the context of the 2020 U.S. election. We also highlight what we *don't* know about these systems, and call on companies to be much more transparent about how they work.

Since 2015, the Ranking Digital Rights (RDR) Corporate Accountability Index has measured how transparent companies are about their policies and practices that affect online freedom of expression and privacy. Despite measurable progress, most companies still fail to disclose basic information about how they decide what content should appear on their platforms, how they collect and monetize user information, and what corporate governance processes guide these decisions. This report offers recommendations for companies and regulators that would make our online ecosystem work for democracy rather than undermine it.

The second installment in this series will dive deeper into our solutions, which focus on corporate governance. We will look at the ways tech companies should anticipate and address the societal harms that their products cause or contribute to, and how society can hold the private sector accountable through elected government and democratic institutions. Strong privacy legislation will be a key first step towards curbing the spread of extremism and disinformation, by placing limits on whether and how vast amounts of personal data can be collected and used to target people. This must come hand in hand with greater transparency about how companies rule themselves, and how they make decisions that affect us all. Only then will citizens and their elected representatives have the tools and information they need to hold tech companies accountable (including with regulation) in a way that advances the public interest.

A Tale of Two Algorithms



Source: Original art by Paweł Kuczyński

In recent years, policymakers and tech executives alike have begun to invoke the almighty algorithm as a solution for controlling the harmful effects of online speech. Acknowledging the pitfalls of relying primarily on human content moderators, technology companies promote the idea that a technological breakthrough, something that will automatically eliminate the worst kinds of speech, is just around the corner.[11]

But the public debate often conflates two very different types of algorithmic systems that play completely different roles in policing, shaping, and amplifying online speech.[12]

First, there are content-shaping algorithms. These systems determine the content that each individual user sees online, including user-generated or organic posts and paid advertisements. Some of the most visible examples of content-shaping algorithms include Facebook's News Feed, Twitter's Timeline, and YouTube's recommendation engine.[13] Algorithms also determine which users should be shown a given ad. The advertiser usually sets the targeting parameters (such as demographics and presumed interests), but the platform's algorithmic systems pick the specific individuals who will see the ad and determine the ad's placement within the platform. Both kinds of personalization are only possible because of the vast troves of detailed information that the companies have accumulated about their users and their online behavior, often without the knowledge or consent of the people being targeted.[14]

While companies describe such algorithms as matching users with the content that is most relevant to them, this relevance is measured by predicted engagement: how likely users are to click, comment on, or share a piece of content. Companies make these guesses based on factors like users' previous interaction with similar content and the interactions of other users who are similar to them. The more accurate these guesses are, the more valuable the company becomes for advertisers, leading to ever-increasing profits for internet platforms. This is why mass data collection is so central to Big Tech's business models: companies need to surveil internet users in order to make predictions about their future behavior.[15]

Companies can and do change their algorithms anytime they want, without any legal obligation to notify the public.

Second, we have content moderation algorithms, built to detect content that breaks the company's rules and remove it from the platform. Companies have made tremendous investments in these technologies. They are increasingly able to identify and remove some kinds of content without human involvement, but this approach has limitations.[16]

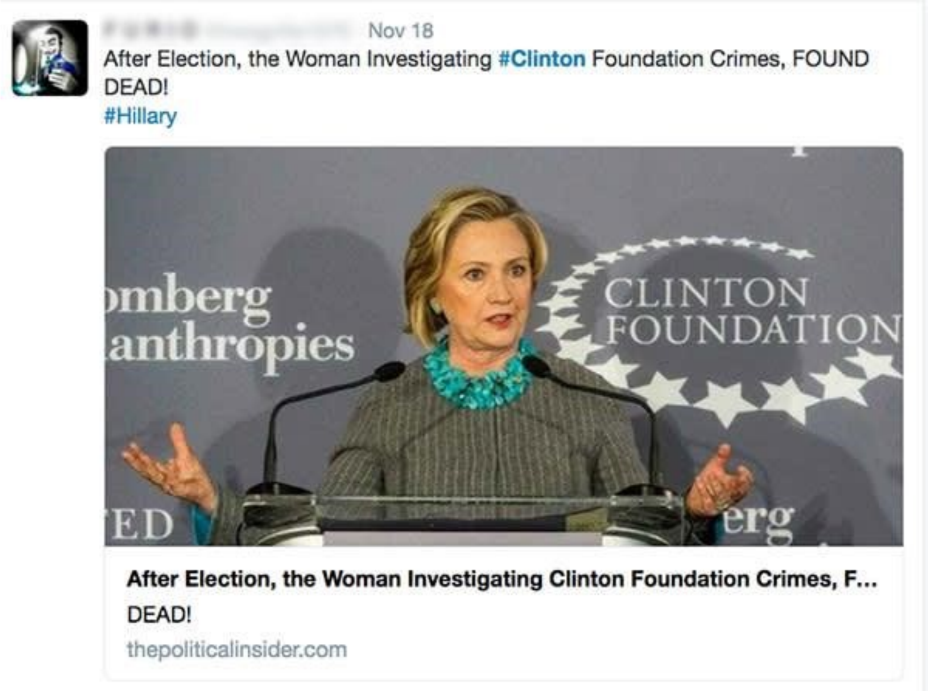
Content moderation algorithms work best when they have a hard and fast rule to follow. This works well when seeking to eliminate images of a distinct symbol, like a swastika. But machine-driven moderation becomes more difficult, if not impossible, when content is violent, hateful, or misleading and yet has some public interest value. Companies are, in effect, adjudicating such content, but this requires the ability to reason—to employ careful consideration of context and nuance. Only humans with the appropriate training can make these kinds of judgments—this is beyond the capability of automated decision-making systems.[17] Thus, in many cases, human reviewers remain involved in the moderation process. The consequences that this type of work has had for human reviewers has become an important area of study unto itself, but lies beyond the scope of this report.[18]

Both types of systems are extraordinarily opaque, and thus unaccountable. Companies can and do change their algorithms anytime they want, without any legal obligation to notify the public. While content-shaping and ad-targeting algorithms work to show you posts and ads that they think are most relevant to your interests, content moderation processes (including algorithms) work alongside this stream of content, doing their best to identify and remove those posts that might cause harm. Let's look at some examples of these dynamics in real life.

Russian Interference, Radicalization, and Dishonest Ads: What Makes Them So Powerful?

Russian interference in recent U.S. elections and online radicalization by proponents of violent extremism are just two recent, large-scale examples of the content problems that we reference above.

Following the 2016 election, it was revealed that Russian government actors had attempted to influence U.S. election results by promoting false online content (posts and ads alike) and using online messaging to stir tensions between different voter factions. These influence efforts, alongside robust disinformation campaigns run by domestic actors,[19] were able to flourish and reach millions of voters (and perhaps influence their choices) in an online environment where content-shaping and ad-targeting algorithms play the role of human editors. Indeed, it was not the mere existence of misleading content that interfered with people's understanding of what was true about each candidate and their positions—it was the reach of these messages, enabled by algorithms that selectively targeted the voters whom they were most likely to influence, in the platforms' estimation.[20]



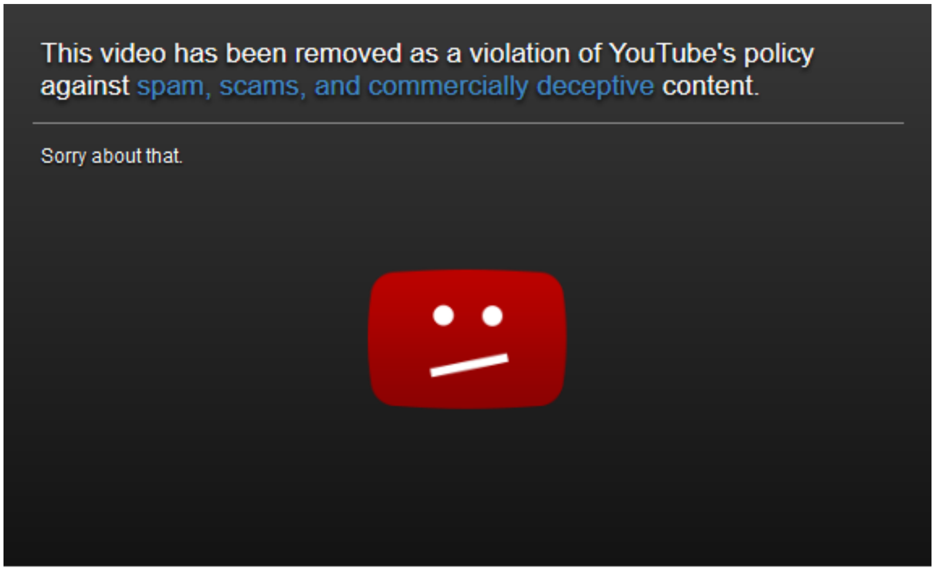
[redacted] Nov 18
After Election, the Woman Investigating #Clinton Foundation Crimes, FOUND DEAD!
#Hillary

After Election, the Woman Investigating Clinton Foundation Crimes, F...
DEAD!
thepoliticalinsider.com

During the 2016 U.S. presidential election, disinformation circulated widely online. The screenshot above is from a fake story shared on Twitter about then-candidate Hillary Clinton (Twitter, Nov. 21, 2016).

The revelations set in motion a frenzy of content restriction efforts by major tech companies and fact-checking initiatives by companies and independent groups alike,[21] alongside Congressional investigations into the issue. Policymakers soon demanded that companies rein in online messages from abroad meant to skew a voter's perspective on a candidate or issue.

How would companies achieve this? With more algorithms, they said.[22] Alongside changes in policies regarding disinformation, companies soon adjusted their content moderation algorithms to better identify and weed out harmful election-related posts and ads. But the scale of the problem all but forced them to stick to technical solutions, with the same limitations as those that caused these messages to flourish in the first place.



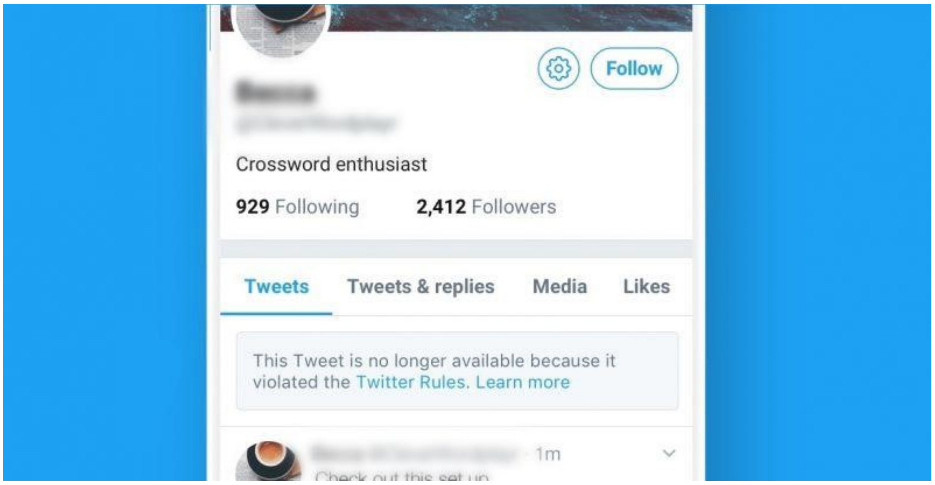
This video has been removed as a violation of YouTube's policy against spam, scams, and commercially deceptive content.

Sorry about that.

A screenshot of what YouTube users see when they try to access a video that has been removed by the company for violating its content policy (YouTube, Feb. 26, 2020).

Content moderation algorithms are notoriously difficult to implement effectively, and often create new problems. Academic studies published in 2019 found that algorithms trained to identify hate speech for removal were more likely to flag social media content created by African Americans, including posts using slang to discuss contentious events and personal experiences related to racism in America.[23] While companies are free to set their own rules and take down any content that breaks those rules, these kinds of removals are in tension with U.S. free speech values, and have elicited the public blowback to match.

Content restriction measures have also inflicted collateral damage on unsuspecting users outside the United States with no discernible connection to disinformation campaigns. As companies raced to reduce foreign interference on their platforms, social network analyst Lawrence Alexander identified several users on Twitter and Reddit whose accounts were suspended simply because they happened to share some of the key characteristics of disinformation purveyors.



Crossword enthusiast

929 Following 2,412 Followers

This Tweet is no longer available because it violated the Twitter Rules. Learn more

Check out this set up

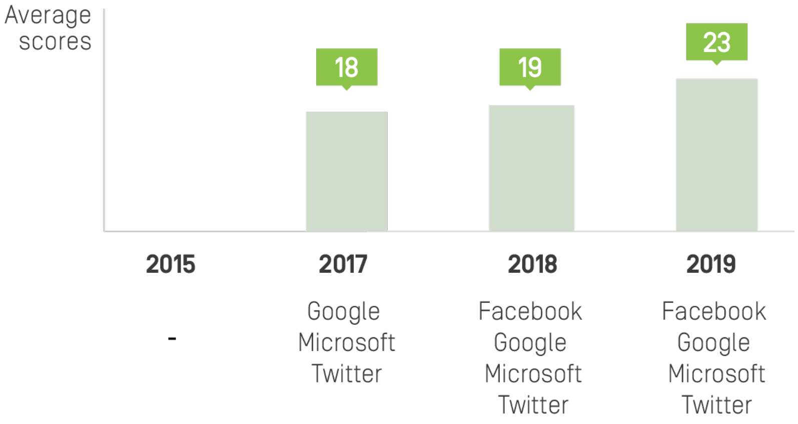
A screenshot of the message users see when a tweet has been removed by Twitter for a violation of the "Twitter Rules" (Twitter, Feb. 28, 2020).

One user had even tried to notify Twitter of a pro-Kremlin campaign, but ended up being banned himself. "[This] quick-fix approach to bot-hunting seemed to have dragged a number of innocent victims into its nets," wrote Alexander, in a research piece for *Global Voices*. For one user who describes himself in his Twitter profile as an artist and creator of online comic books, "it appears that the key 'suspicious' thing about their account was their location—Russia."[24]

The lack of corporate transparency regarding the full scope of disinformation and malicious behaviors on social media platforms makes it difficult to assess how effective these efforts actually are.

For four years, RDR has tracked whether companies publish key information about how they enforce their content rules. In 2015, none of the companies we evaluated published any data about the content they removed for violating their platform rules.[25] Four years later, we found that Facebook, Google, Microsoft, and Twitter published at least some information about their rules enforcement,[26] including in transparency reports.[27] But this information still doesn't demonstrate how effective their content moderation mechanisms have actually been in enforcing their rules, or how often acceptable content gets caught up in their nets.

How transparent are companies about terms of service enforcement (2015-2019)?



Companies have also deployed automated systems to review election-related ads, in an effort to better enforce their policies. But these efforts too have proven problematic. Various entities, ranging from media outlets[28] to LGBTQ-rights groups[29] to Bush's Baked Beans,[30] have reported having their ads rejected for violating election-ad policies, despite the fact that their ads had nothing to do with elections.[31] Yet companies' disclosures about how they police ads are even less transparent than those pertaining to user-generated content, and there's no way to know how effective these policing mechanisms have been in enforcing the actual rules as intended. [32]

The issue of online radicalization is another area of concern for U.S. policymakers, in which both types of algorithms have been in play. Videos and social media channels enticing young people to join violent extremist groups or to commit violent acts became remarkably easy to stumble upon online, due in part to what seems to be these groups' savvy understanding of how to make algorithms that amplify and spread content work to their advantage.[33] Online extremism and radicalization are very real problems that the internet has exacerbated. But efforts to address this problem have led to unjustified censorship.

Widespread concern that international terrorist organizations were recruiting new members online has led to the creation of various voluntary initiatives, including the Global Internet Forum to Counter Terrorism (GIFCT), which helps companies jointly assess content that has been identified as promoting or celebrating terrorism.

Scale matters—the societal impact of a single message or video rises exponentially when a powerful algorithm is driving its distribution.

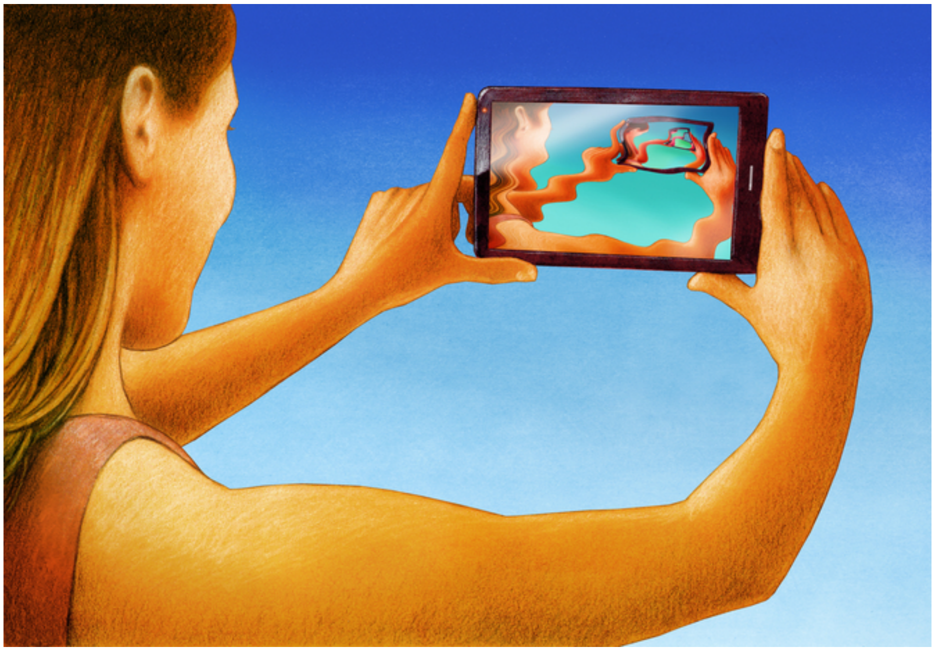
The GIFCT has built an industry-wide database of digital fingerprints or “hashes” for such content. Companies use these hashes to filter offending content, often prior to upload. As a result, any errors made in labeling content as terrorist in this database[34] are replicated on all participating platforms, leading to the censorship of photos and videos containing speech that should be protected under international human rights law.

Thousands of videos and photos from the Syrian civil war have disappeared in the course of these efforts—videos that one day could be used as evidence against perpetrators of violence. No one knows for sure whether these videos were removed because they matched a hash in the GIFCT database, because they were flagged by a content moderation algorithm or human reviewer, or some other reason. The point is that this evidence is often impossible to replace. But little has been done to change the way this type of content is handled, despite its enormous potential evidentiary value.[35]

In an ideal world, violent extremist messages would not reach anyone. But the public safety risks that these carry rise dramatically when such messages reach tens of thousands, or even millions, of people.

The same logic applies to disinformation targeted at voters. Scale matters—the societal impact of a single message or video rises exponentially when a powerful algorithm is driving its distribution. Yet the solutions for these problems that we have seen companies, governments, and other stakeholders put forth focus primarily on eliminating content itself, rather than altering the algorithmic engines that drive its distribution.

Algorithmic Transparency: Peeking Into the Black Box



Source: Original art by Paweł Kuczyński

While we know that algorithms are often the underlying cause of virality, we don't know much more—corporate norms fiercely protect these technologies as private property, leaving them insulated from public scrutiny, despite their immeasurable impact on public life.

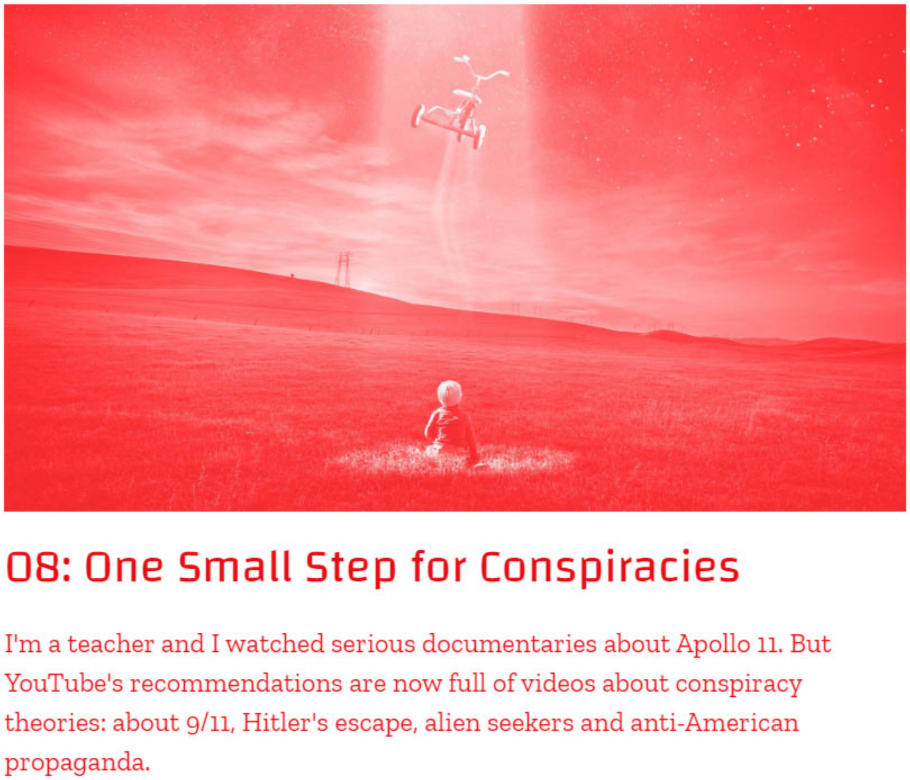
Since early 2019, RDR has been researching the impact of internet platforms' use of algorithmic systems, including those used for targeted advertising, and how companies should govern these systems.[36] With some exceptions, we found that companies largely failed to disclose sufficient information about these processes, leaving us all in the dark about the forces that shape our information environments. Facebook, Google, and Twitter do not hide the fact that they use algorithms to shape content, but they are less forthcoming about how the algorithms actually work, what factors influence them, and how users can customize their own experiences. [37]

Companies largely failed to disclose sufficient information about their algorithmic systems, leaving us all in the dark about the forces that shape our information environments.

U.S. tech companies have also avoided publicly committing to upholding international human rights standards for how they develop and use algorithmic systems. Google, Facebook, and other tech companies are instead leading the push for ethical or responsible artificial intelligence as a way of steering the discussion away from regulation.[38] These initiatives lack established, agreed-upon principles, and are neither legally binding nor enforceable—in contrast to international human rights doctrine, which offers a robust legal framework to guide the development and use of these technologies.

Grounding the development and use of algorithmic systems in human rights norms is especially important because tech platforms have a long record of launching new products without considering the impact on human rights.[39] Neither Facebook, Google, nor Twitter disclose any evidence that they conduct human rights due diligence on their use of algorithmic systems or on their use of personal information to develop and train them. Yet researchers and journalists have found strong evidence that algorithmic content-shaping systems that are optimized for user engagement prioritize controversial and inflammatory content.[40] This can help foster online communities centered around specific ideologies and conspiracy theories, whose members further radicalize each other and may even collaborate on online harassment campaigns and real-world attacks.[41]

One example is YouTube's video recommendation system, which has come under heavy scrutiny for its tendency to promote content that is extreme, shocking, or otherwise hard to look away from, even when a user starts out by searching for information on relatively innocuous topics. When she looked into the issue ahead of the 2018 midterm elections, University of North Carolina sociologist Zeynep Tufekci wrote, "What we are witnessing is the computational exploitation of a natural human desire: to look 'behind the curtain,' to dig deeper into something that engages us. ... YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales."[42]



08: One Small Step for Conspiracies

I'm a teacher and I watched serious documentaries about Apollo 11. But YouTube's recommendations are now full of videos about conspiracy theories: about 9/11, Hitler's escape, alien seekers and anti-American propaganda.

This image appeared in the Mozilla Foundation's 2019 "YouTube Regrets" campaign highlighting how the YouTube recommendation algorithm suggests videos with ever-more extreme viewpoints (The Mozilla Foundation, Sept. 10, 2019).

Tufekci's observations have been backed up by the work of researchers like Becca Lewis, Jonas Kaiser, and Yasodara Cordova. The latter two, both affiliates of Harvard's Berkman Klein Center for Internet and Society, found that when users searched videos of children's athletic events, YouTube often served them recommendations of videos of sexually themed content featuring young-looking (possibly underage) people with comments to match.[43]

After the *New York Times* reported on the problem last year, the company removed several of the videos, disabled comments on many videos of children, and made some tweaks to its recommendation system. But it has stopped short of turning off recommendations on videos of children altogether. Why did the company stop there? The *New York Times's* Max Fisher and Amanda Taub summarized YouTube's response: "The company said that because recommendations are the biggest traffic driver, removing them would hurt creators who rely on those clicks."[44] We can surmise that eliminating the recommendation feature would also compromise YouTube's ability to keep users hooked on its platform, and thus capture and monetize their data.

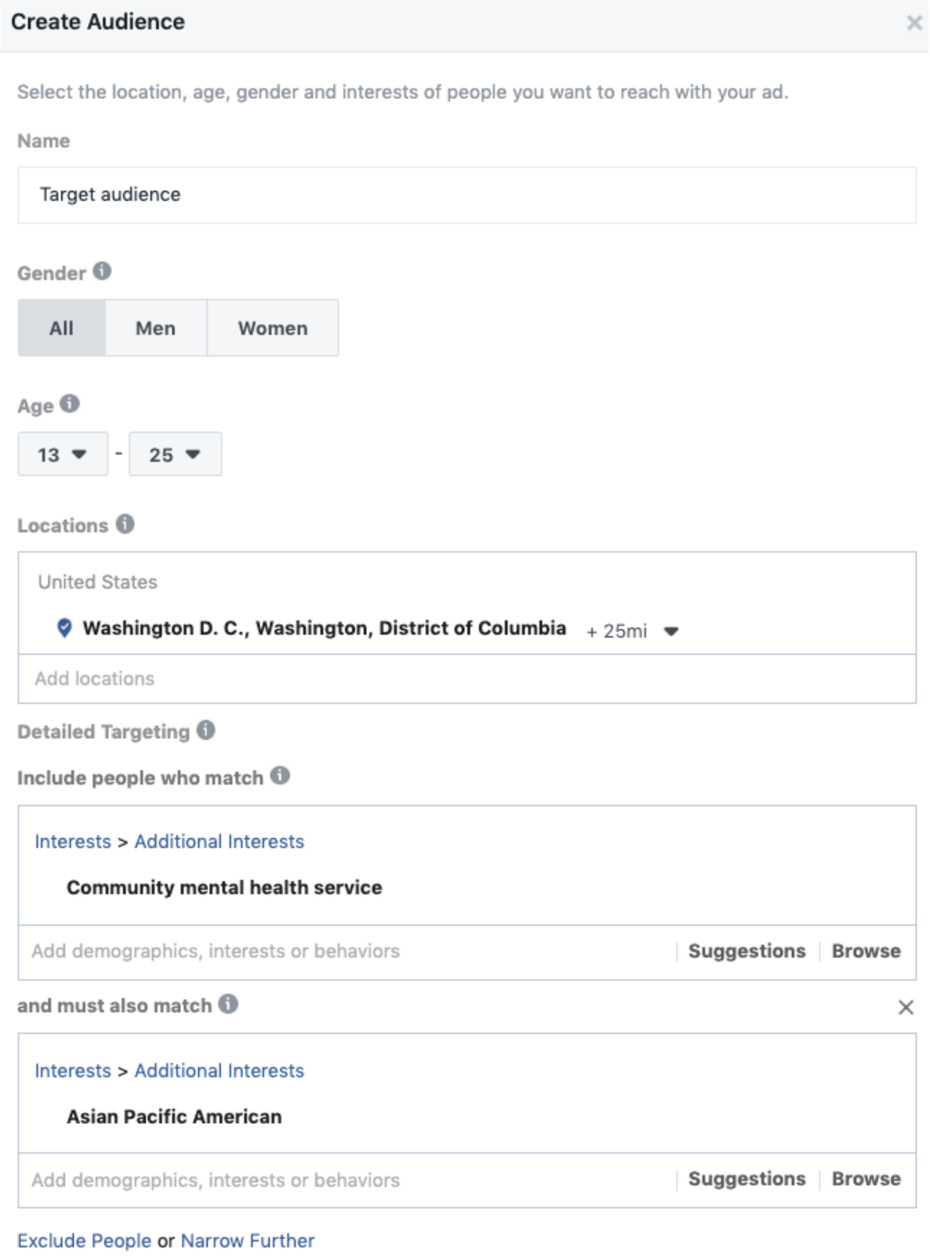
If companies were required by law to meet baseline standards for transparency in algorithms like this one, policymakers and civil society advocates alike would be better positioned to push for changes that benefit the public interest. And if they were compelled to conduct due diligence on the negative effects of their products, platforms would likely make very different choices about how they use and develop algorithmic systems.

Who Gets Targeted—Or Excluded—By Ad Systems?

Source: Original art by Paweł Kuczyński

Separate from the algorithms that shape the spread of user-generated content, advertising is the other arena in which policymakers must dig deeper and examine the engines that determine and drive distribution. Traditional advertising places ads based on context: everyone who walks by a billboard or flips through a magazine sees the same ad. Online targeted advertising, on the other hand, is personalized based on what advertisers and ad networks know (or think they know) about each person, based on the thick digital dossiers they compile about each of us.

In theory, advertisers on social media are responsible for determining which audience segments (as defined by their demographics or past behavior) will see a given ad, but in practice, platforms further optimize the audience beyond what the advertiser specified.[45] Due to companies' limited transparency efforts, we know very little about how this stage of further optimization works.



Create Audience

Select the location, age, gender and interests of people you want to reach with your ad.

Name: Target audience

Gender: All

Age: 13 - 25

Locations: United States
Washington D. C., Washington, District of Columbia + 25mi

Detailed Targeting

Include people who match: Interests > Additional Interests
Community mental health service

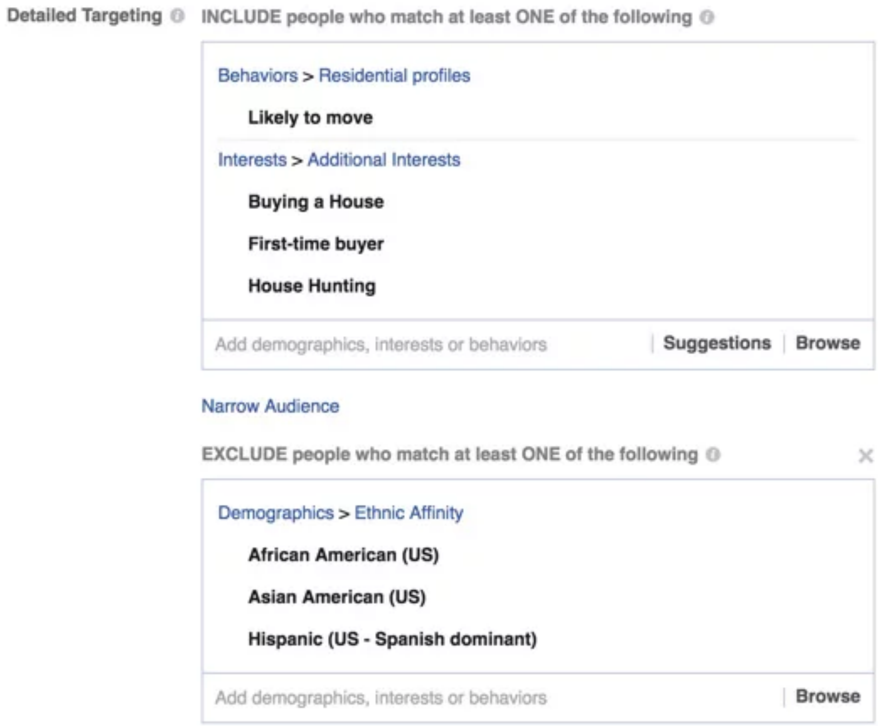
and must also match: Interests > Additional Interests
Asian Pacific American

All
Men
Women
Add locations
Add demographics, interests or behaviors
Suggestions
Browse
Exclude People or Narrow Further

A screenshot of the settings to create a Custom Audience for an ad on Facebook, with targeting by gender, age, location, interests, and other attributes. Narrowing the audience this way could enable discriminatory practices (Facebook, Feb. 28, 2020).

Nevertheless, from what we do know, these systems are already designed to discriminate—when you place an ad on a given platform, you are encouraged to select (from a list of options) different types of people you'd like the ad to reach. Such differentiation amounts to unfair and even illegal discrimination in many instances.[46] One powerful example of these dynamics (and of companies' reticence to make changes at the systemic level) emerged from an investigation conducted by *ProPublica*, in which the media outlet attempted to test Facebook's ad-targeting tools by purchasing a few ads, selecting targeted recipients for the ads, and observing the process by which they were vetted and approved.[47]

"The ad we purchased was targeted to Facebook members who were house hunting and excluded anyone with an 'affinity' for African-American, Asian-American or Hispanic people," the reporters wrote. They explained that Facebook's ad-sales interface allowed them to tick different boxes, selecting who would—and would not—see their ads.[48]



Detailed Targeting

INCLUDE people who match at least ONE of the following

Behaviors > Residential profiles

Likely to move

Interests > Additional Interests

Buying a House

First-time buyer

House Hunting

Narrow Audience

EXCLUDE people who match at least ONE of the following

Demographics > Ethnic Affinity

African American (US)

Asian American (US)

Hispanic (US - Spanish dominant)

Add demographics, interests or behaviors
Suggestions
Browse
Add demographics, interests or behaviors
Browse

A screenshot from a 2016 ProPublica article. The upper panel allows advertisers to select specific audiences that they want to reach; the lower panel allows them to select those audiences that they want to exclude (ProPublica, Oct. 28, 2016).

The ads were approved 15 minutes after they were submitted for review. Civil liberties experts consulted by *ProPublica* confirmed that the option of excluding people with an "affinity" for African-American, Asian-American, or Hispanic people was a clear violation of the U.S. Fair Housing Act, which prohibits real estate entities from discriminating against prospective renters or buyers on the basis of their race, ethnicity, or other identity traits.

After *ProPublica* wrote about and confronted Facebook on the issue, the company added an "advertiser education" section to its ad portal, letting advertisers know that such discrimination was illegal under U.S. law. It also began testing machine learning that would identify discriminatory ads for review. But the company preserved the key tool that allowed these users to be excluded in the first place: the detailed targeting criteria, pictured above, which allowed advertisers to target or exclude African-Americans and Hispanics.

Neither Facebook, Google, nor Twitter show evidence that they conduct due diligence on their targeted advertising practices.

Rather than addressing systemic questions around the social and legal consequences of this type of targeting system, Facebook focused on superficial remedies that left its business model untouched. In another investigation, *ProPublica* learned that these criteria are in fact generated not by Facebook employees, but by technical processes that comb the data of Facebook's billions of users and then establish targeting categories based on users' stated interests. In short, by an algorithm.[49]

In spite of the company's apparent recognition of the problem, Facebook did not take away or even modify these capabilities until many months later, after multiple groups took up the issue and filed a class action lawsuit claiming that the company had violated the Fair Housing Act.[50]

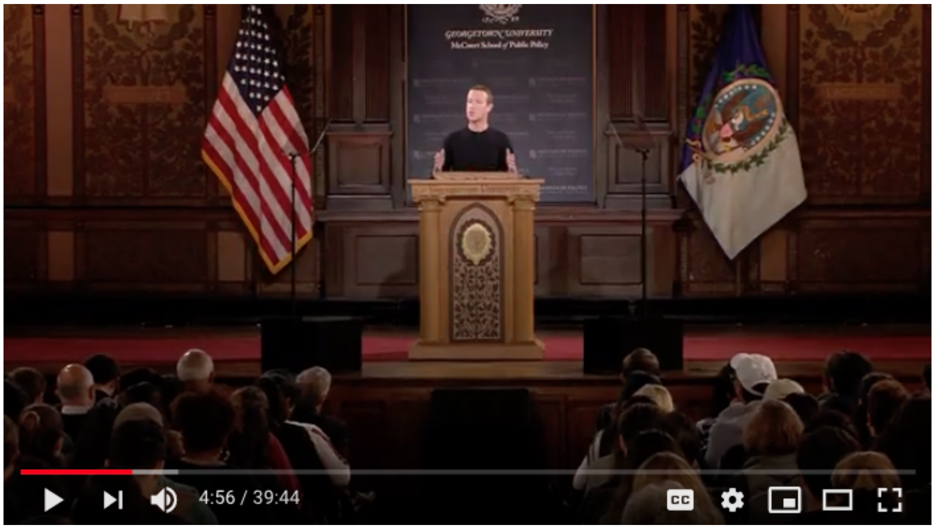
Here again, a company-wide policy for using and developing algorithms combined with human rights due diligence would most likely have identified these risks ahead of time and helped Facebook develop its ad-targeting systems in a way that respects free expression, privacy, and civil rights laws like the Fair Housing Act. But this is the norm rather than the exception. Neither Facebook, Google, nor Twitter show evidence that they conduct due diligence on their targeted advertising practices.[51]

When Ad Targeting Meets the 2020 Election

Discriminating in housing ads is against the law.[52] Discriminating in campaign ads may not be against the law, but many Americans strongly believe that it is bad for democracy.[53]

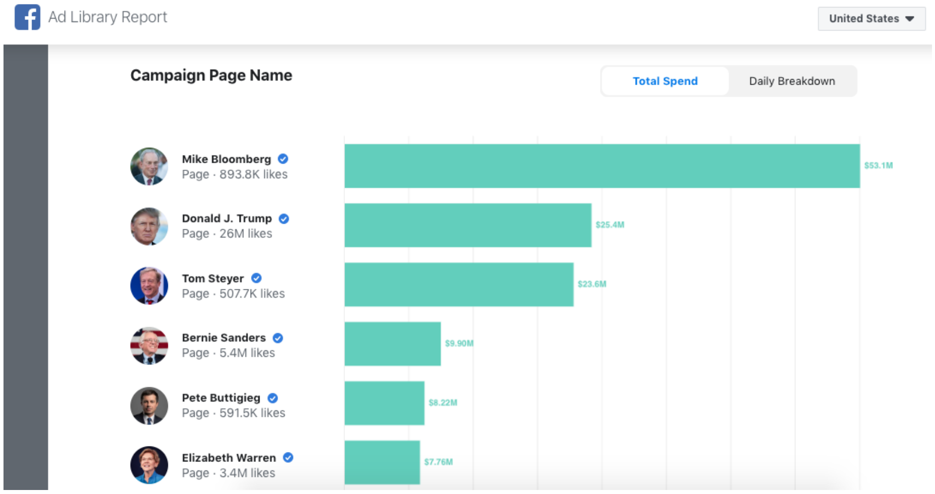
The same targeted advertising systems that are used to target people based on their interests and affinities were used to manipulate voters in the 2016 and 2018 elections. Most egregious, voters who were thought to lean toward Democratic candidates were targeted with ads containing incorrect information about how, when, and where to vote.[54] Facebook, Google, and Twitter now prohibit this kind of disinformation, but we don't know how effectively the rule is enforced: none of these companies publish information about their processes for enforcing advertising rules or about the outcomes of those processes.[55]

At the scale that these platforms operate, fact-checking ad content is hard enough when the facts in question are indisputable, such as the date of an upcoming election. It's even thornier when subjective claims about an opponent's character or details of their policy proposals are in play. Empowering private companies to evaluate truth would be dangerously undemocratic, as Facebook CEO Mark Zuckerberg has himself argued.[56] In the absence of laws restricting the content or the targeting of campaign ads, campaigns can easily inundate voters with ads peddling misleading claims on issues they care about, in an effort to sway their votes.



A screenshot of Mark Zuckerberg speaking about Facebook's stance on free expression and political advertising at Georgetown University in Washington, D.C. (YouTube, Oct. 17, 2019).

Barack Obama and Donald Trump both owe their presidencies to this type of targeting to an extent, though exactly to what extent is impossible to quantify. The 2008 Obama campaign pioneered the use of voters' personal information, and his reelection team refined the practice in 2012. In 2016, thanks to invaluable guidance from Facebook itself, Trump ran "the single best digital ad campaign [...] ever seen from any advertiser," as Facebook executive Andrew "Boz" Bosworth put it in an internal memo.[57] The president's 2020 reelection campaign is on the same track.[58] More than ever, elections have turned into marketing contests. This shift long predates social media companies, but targeted advertising and content-shaping algorithms have amplified the harms and made it harder to address them.



Campaign Page Name	
Mike Bloomberg ✓ Page · 893.8K likes	$53.1M
Donald J. Trump ✓ Page · 26M likes	$25.4M
Tom Steyer ✓ Page · 507.7K likes	$23.6M
Bernie Sanders ✓ Page · 5.4M likes	$9.90M
Pete Buttigieg ✓ Page · 591.5K likes	$8.22M
Elizabeth Warren ✓ Page · 3.4M likes	$7.76M

A screenshot of political campaigns' ad spending on Facebook, as shown on the Facebook Ad Library Report website, illustrating the extent to which elections may have become marketing contests (Facebook, Feb. 26, 2020).

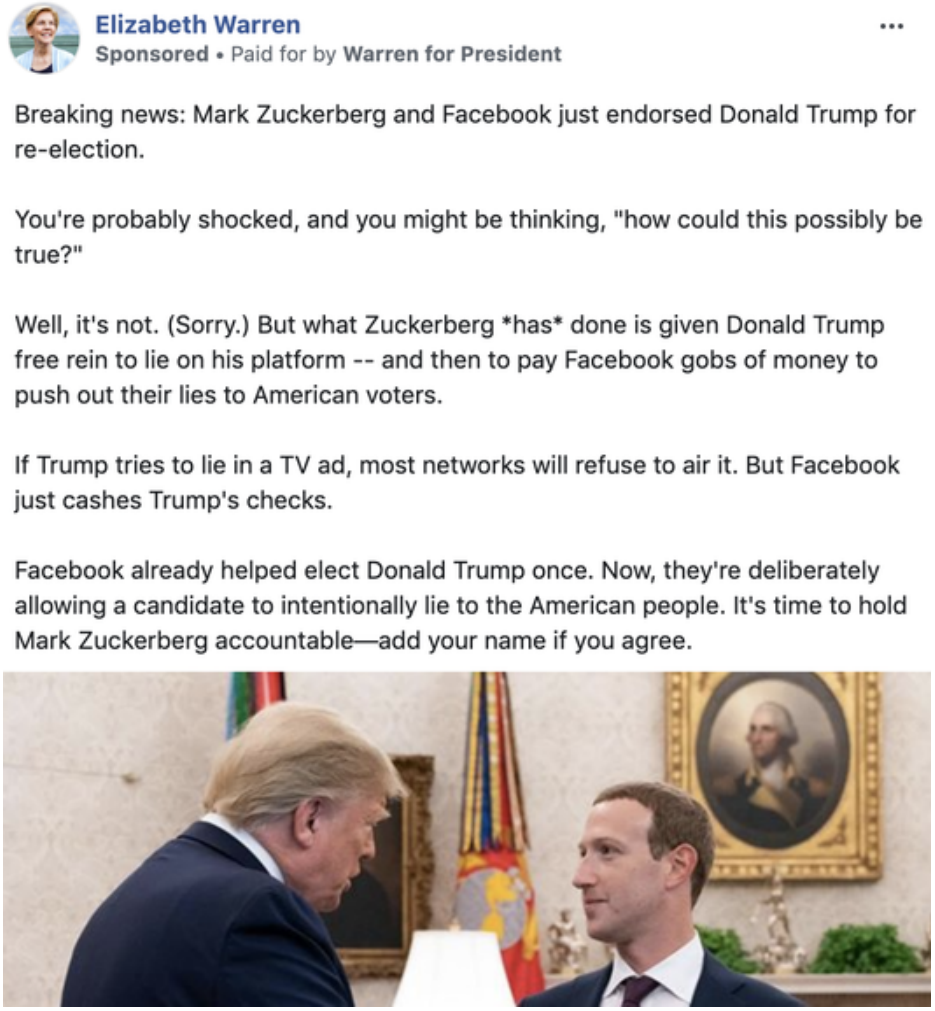
In the 2020 election cycle, we find ourselves in an online environment dominated by algorithms that appear ever-more powerful and effective at spreading content to precisely the people who will be most affected by it, thanks to continued advances in data tracking and analysis. Some campaigns are now using cell phone location data to identify churchgoers, Planned Parenthood patients, and similarly sensitive groups.[59] Many of the risks we've articulated in unique examples thus far will be in play, and algorithms likely will multiply their effects for everyone who relies on social media for news and information.

We are entering a digital perfect storm fueled by deep political cleavages, opaque technological systems, and billions of dollars in campaign ad money that may prove disastrous for our democracy.

We need look no further than the bitter debates that played out around political advertising in the final months of 2019 to see just how high the stakes have become.

In October 2019, when Donald Trump's reelection campaign purchased a Facebook ad claiming that "Joe Biden promised Ukraine $1 billion if they fired the prosecutor investigating his son's company," Facebook accepted it, enabling thousands of users to see (and perhaps believe) it. It didn't matter that the claim was unfounded, and had been debunked by two of Facebook's better-known fact-checking partners, PolitiFact and Factcheck.org.

When Facebook decided to stand by this decision, and to let the ad stay up, Sen. Elizabeth Warren (D-Mass.)—who was running for the Democratic nomination at the time, alongside Biden—ran a Facebook ad of her own, which made the intentionally false claim that "Mark Zuckerberg and Facebook just endorsed Donald Trump for re-election."[60]



Elizabeth Warren
Sponsored • Paid for by **Warren for President**

Breaking news: Mark Zuckerberg and Facebook just endorsed Donald Trump for re-election.

You're probably shocked, and you might be thinking, "how could this possibly be true?"

Well, it's not. (Sorry.) But what Zuckerberg *has* done is given Donald Trump free rein to lie on his platform -- and then to pay Facebook gobs of money to push out their lies to American voters.

If Trump tries to lie in a TV ad, most networks will refuse to air it. But Facebook just cashes Trump's checks.

Facebook already helped elect Donald Trump once. Now, they're deliberately allowing a candidate to intentionally lie to the American people. It's time to hold Mark Zuckerberg accountable—add your name if you agree.

A screenshot of a deliberately misleading Facebook ad (cropped) run by presidential candidate Senator Elizabeth Warren meant to call attention to how easily misinformation can spread on the platform (Facebook, Oct. 10, 2019).

The ad was intended to draw attention to how easily politicians can spread misinformation on Facebook. Indeed, unlike print and broadcast ads, online political ads are completely unregulated in the United States: parties, campaigns, and outside groups are free to run any ads they want, if the platform or advertising network lets them. This gives companies like Google, Facebook, and Twitter tremendous power to set the rules by which political campaigns operate.

Soon after Warren's attention-grabbing ad, the *New York Times* published a letter that had circulated internally at Facebook, and was signed by 250 staff members. The letter's authors criticized Facebook's refusal to fact-check political ads and tied the issue to ad targeting, arguing that it "allows politicians to weaponize our platform by targeting people who believe that content posted by political figures is trustworthy" and could "cause harm in coming elections around the world."[61]

Among other demands, the authors urged Facebook to restrict targeting for political ads. But the company did not relent, reportedly at the insistence of board member Peter Thiel.[62] The only subsequent change that Facebook has made to this system is to allow users to opt out of custom audiences, a tool that allows advertisers to upload lists of specific individuals to target.[63]

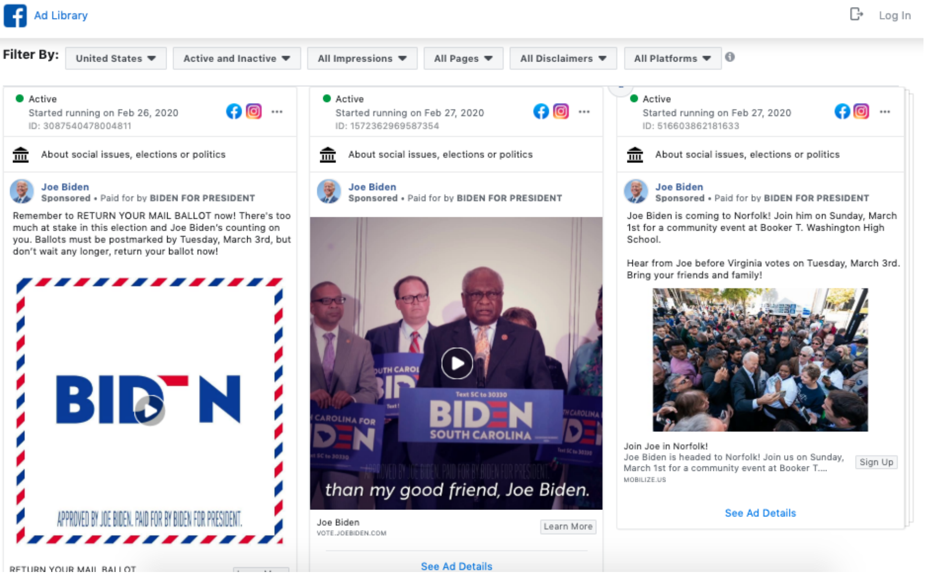
In response to increased public scrutiny around political advertising, the other major U.S. platforms also moved to tweak their own approaches to political ads. Twitter, which only earned $3 million from political ads during the 2018 U.S. midterms,[64] announced in October that it would no longer accept political ads,[65] and restrict how "issue-based" ads (which argue for or against a policy position without naming specific candidates) can be targeted.[66] Google elected to limit audience targeting for election ads to age, gender, and zip code, though it remains unclear precisely what kind of algorithm will be able to correctly identify (and then disable targeting for) election ads. None of the companies have given any indication that they conducted a human rights impact assessment or other due diligence prior to announcing these changes.[67]

The companies' insistence on drawing unenforceable lines around "political ads," "issue ads," and "election ads" highlights how central targeting is to their business models.

We might read Twitter and Google's decisions as an acknowledgment that the algorithms underlying the distribution of targeted ads are in fact a major driver of the kinds of disinformation campaigns and platform weaponization that can so powerfully affect our democracy. However, the companies' insistence on drawing unenforceable lines around "political ads," "issue ads," and "election ads" highlights how central targeting is to their business models. Facebook's decision regarding custom audiences signals the same thing: as long as users are included in custom audiences by default, the change will have limited effects.

Ad targeting is just the beginning of such influence campaigns. As a Democratic political operative told the *New York Times*, "the real goal of paid advertising is for the content to become organic social media."[68] Once a user boosts an ad by sharing it as their own post, the platform's content-shaping algorithms treat it as normal user content and highlight it to people in the user's network who are more likely to click, like, and otherwise engage with it, allowing campaigns to reach audiences well beyond the targeted segments.

The major platforms' newly introduced ad libraries typically allow the public to find out who paid for an ad, how much they spent, and some targeting parameters. They shed some light into targeted campaigns themselves. But it is impossible to know how far messages travel without meaningful algorithmic transparency.[69] All we know is that between ad-targeting and content-shaping algorithms, political campaigns are dedicating more resources to mastering the dark arts of data science.



Filter By: United States ▼ Active and Inactive ▼ All Impressions ▼ All Pages ▼ All Disclaimers ▼ All Platforms ▼

● Active
Started running on Feb 26, 2020
ID: 3087540478004811

About social issues, elections or politics

Joe Biden
Sponsored • Paid for by BIDEN FOR PRESIDENT

Remember to RETURN YOUR MAIL BALLOT now! There's too much at stake in this election and Joe Biden's counting on you. Ballots must be postmarked by Tuesday, March 3rd, but don't wait any longer, return your ballot now!

RETURN YOUR MAIL BALLOT

● Active
Started running on Feb 27, 2020
ID: 1572362969587354

About social issues, elections or politics

Joe Biden
Sponsored • Paid for by BIDEN FOR PRESIDENT

Joe Biden
VOTE.JOEBIDEN.COM

● Active
Started running on Feb 27, 2020
ID: 516603862181633

About social issues, elections or politics

Joe Biden
Sponsored • Paid for by BIDEN FOR PRESIDENT

Joe Biden is coming to Norfolk! Join him on Sunday, March 1st for a community event at Booker T. Washington High School.

Hear from Joe before Virginia votes on Tuesday, March 3rd. Bring your friends and family!

Join Joe in Norfolk!
Joe Biden is headed to Norfolk! Join us on Sunday, March 1st for a community event at Booker T....
MOBILIZE.US

A screenshot of political U.S. advertisements in the Facebook Ad Library, capturing three of former Vice President Joe Biden's campaign ads (Facebook, Feb. 27, 2020).

We have described the nature and knock-on effects of two general types of algorithms. The first drives the distribution of content across a company's platform. The second seeks to identify and eliminate specific types of content that have been deemed harmful, either by the law, or by the company itself. But we have only been able to scratch the surface of how these systems really operate, precisely because we cannot see them. To date, we only see evidence of their effects when we look at patterns of how certain kinds of content circulate online.

Although the 2020 election cycle is already in full swing, we are only beginning to understand just how powerful these systems can be in shaping our information environments, and in turn, our political reality.

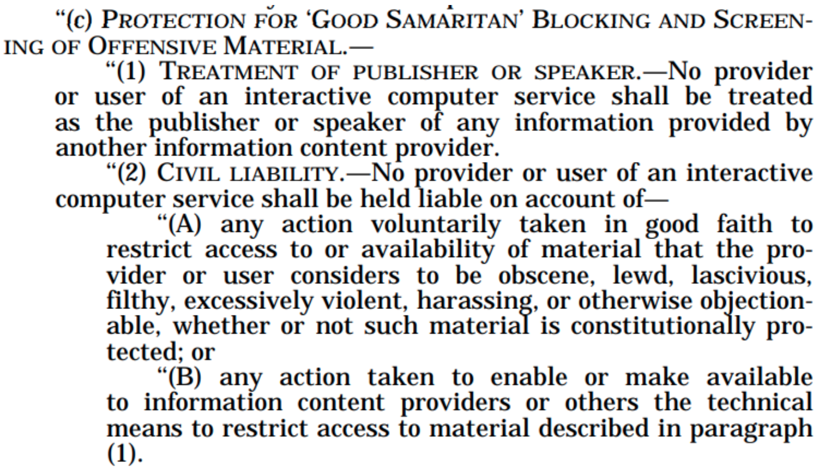
Regulatory Challenges: A Free Speech Problem—and a Tech Problem

Until now, Congress has largely put its faith in companies' abilities to self-regulate. But this has clearly not worked. We have reached a tipping point—a moment in which protecting tech companies' abilities to build and innovate unfettered might actually be putting our democracy at grave risk. We have come to realize that we have a serious problem on our hands, and that the government must step in with regulation. But what should this regulation look like?

There is no clear or comprehensive solution to these problems right now. But we know that we need more information—likely through government-mandated transparency and impact-assessment requirements—in order to assess the potential for damage and propose viable solutions. It is also clear that we need a strong federal privacy law. These recommendations will be explored in greater depth in the second part of this report series.

We have reached a tipping point—a moment in which protecting tech companies' abilities to build and innovate unfettered might actually be putting our democracy at grave risk.

Members of Congress are understandably eager to hold tech platforms accountable for the harms they enable, but should resist the temptation of quick-fix solutions: not all types of regulation will actually solve the problems of disinformation and violent extremism online without also seriously corroding democracy. We urge policymakers to refrain from instituting a broad intermediary liability regime, such as by revoking or dramatically revising Section 230 of the 1996 Communications Decency Act (CDA). Section 230 provides that companies cannot be held liable for the content that their users post on their platforms, within the bounds of U.S. law. It also protects companies' ability to develop their own methods for identifying and removing content. Without this protection, companies that moderate their users' content would be held liable for damages caused by any content that they failed to remove, creating strong incentives for companies to censor users' posts. Instead, Section 230 allows companies to do their best to govern the content on their platforms through their terms of service and community standards.



"(c) Protection for 'Good Samaritan' Blocking and Screening of Offensive Material.—

"(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 provided by another information content provider.

"(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).

Source: 47. Communications Decency Act, U.S.C. § 230(c).

Experts in media and technology policy are all but unanimous that eliminating CDA 230 would be disastrous for free speech — domestically and globally.[70] Perfect enforcement is impossible, and holding companies liable for failing to do the impossible will only lead to over-censorship.

Foreign governments are not constrained by the U.S. First Amendment and have the power to regulate speech on internet platforms more directly. This is already happening in various jurisdictions, including Germany, where the 2018 NetzDG law requires social media companies to swiftly remove illegal speech, with a specific focus on hate speech and hate crimes, or pay a fine.[71] While this may reduce illegal online speech, it bypasses important measures of due process, delegating judicial authority (normally reserved for judges) to private companies. It also incentivizes them to err on the side of censorship rather than risk paying fines.[72]

Another example comes with the anti-terrorist content regulation currently pending before the European Commission[73] that would, among other things, require companies to institute upload filters. These still-hypothetical algorithmic systems would theoretically be able to evaluate not only the content of an image or video but also its context, the user's intent in posting it, and the competing arguments for removing the content versus taking it down. Automated tools may be able to detect an image depicting a terrorist atrocity, but they cannot recognize or judge the context or deeper significance of a piece of content. For that, human expertise and judgment are needed.[74]

A desire to see rapid and dramatic reduction in disinformation, hate speech, and violent extremism leads to a natural impulse to mandate outcomes. But technology simply cannot achieve these results without inflicting unacceptable levels of collateral damage to human rights and civil liberties.

So What Should Companies Do?

What would a better framework, one that puts democracy, civil liberties, and human rights above corporate profits, look like?

It is worth noting that Section 230 stipulates that companies are protected from liability in taking "any action voluntarily taken in good faith" to enforce its rules. Section 230 doesn't provide any guidance as to what "good faith" actually means for how companies should govern their platforms.[75] Some legal scholars have proposed reforms that would keep Section 230 in place but would clarify steps that companies need to take in order to demonstrate good-faith efforts to mitigate harm, in order to remain exempt from liability.[76]

Companies hoping to convince lawmakers not to abolish or drastically change Section 230 would be well advised to proactively and voluntarily implement a number of policies and practices to increase transparency and accountability. This would help to mitigate real harms that users or communities can experience when social media is used by malicious or powerful actors to violate their rights.

First, companies' speech rules must be clearly explained and consistent with established human rights standards for freedom of expression. Second, these rules must be enforced fairly according to a transparent process. Third, people whose speech is restricted must have an opportunity to appeal. And finally, the company must regularly publish transparency reports with detailed information about the steps that the company takes to enforce its rules.[77]

Since 2015, RDR has encouraged internet and telecommunications companies to publish basic disclosures about their policies and practices that affect their users' rights. Our annual **RDR Corporate Accountability Index** benchmarks major global companies against each other and against standards grounded in international human rights law.

Much of our work measures companies' transparency about the policies and processes that shape users' experiences on their platforms. We have found that—absent a regulatory agency empowered to verify that companies are conducting due diligence and acting on it—transparency is the best accountability tool at our disposal. Once companies are on the record describing their policies and practices, journalists and researchers can investigate whether they are actually telling the truth.

Transparency allows journalists, researchers, and the public and their elected representatives to make better informed decisions about the content they receive and to hold companies accountable.

We believe that platforms have the responsibility to set and enforce the ground rules for user-generated and ad content on their services. These rules should be grounded in international human rights law, which provides a framework for balancing the competing rights and interests of the various parties involved.[78] Operating in a manner consistent with international human rights law will also strengthen the U.S.' long-standing bipartisan policy of promoting a free and open global internet.

But again, content is only part of the equation. Companies must take steps to publicly disclose the different technological systems at play: the content-shaping algorithms that determine what user-generated content users see, and the ad-targeting systems that determine who can pay to influence them. Specifically, companies should explain the purpose of their content-shaping algorithms and the variables that influence them so that users can understand the forces that cause certain kinds of content to proliferate, and other kinds to disappear.[79] Currently, companies are not transparent or accountable for how their targeted-advertising policies and practices and their use of automation shape the online public sphere by determining the content and information that internet users receive.[80]

Companies also need to publish their rules for ad targeting, and be held accountable for enforcing those rules. Our research shows that while Facebook, Google, and Twitter all publish ad-targeting rules that list broad audience categories that advertisers are prohibited from using, the categories themselves can be excessively vague and unclear—Twitter for instance bans advertisers from using audience categories "that we consider sensitive or are prohibited by law, such as race, religion, politics, sex life, or health."[81] Nor do these platforms disclose any data about the number of ads they removed for violating their ad-targeting rules (or other actions they took).[82]

Facebook says that advertisers can target ads to custom audiences, but prohibits them from using targeting options "to discriminate against, harass, provoke, or disparage users or to engage in predatory advertising practices." However, not everyone can see what these custom audience options are, since these are only available to Facebook users. And Facebook publishes no data about the number of ads removed for breaching ad-targeting rules.

Platforms should set and publish rules for targeting parameters, which should apply equally to all ads—a practice like this would make it much more difficult for companies to violate anti-discrimination laws like the Fair Housing Act. Moreover, once an advertiser has chosen their targeting parameters, companies should refrain from further optimizing ads for distribution, as this may lead to further discrimination.[83]

Platforms should not differentiate between commercial, political, and issue ads, for the simple reason that drawing such lines fairly, consistently, and at a global scale is impossible and complicates the issue of targeting.

Eliminating targeting practices that exploit individual internet users' characteristics (real or assumed) would protect privacy, reduce filter bubbles, and make it harder for political advertisers to send different messages to different constituent groups.

Limiting targeting, as Federal Elections Commissioner Ellen Weintraub has argued,[84] is a much better approach, though here again, the same rules should apply for all types of ads. Eliminating targeting practices that exploit individual internet users' characteristics (real or assumed) would protect privacy, reduce filter bubbles, and make it harder for political advertisers to send different messages to different constituent groups. This is the kind of reform that will be addressed in the second part of this report series.

In addition, companies should conduct due diligence through human rights impact assessments on all aspects of what their rules are, how they are enforced and what steps the company takes to prevent violations of users' rights. This process forces companies to anticipate worst case scenarios, and change their plans accordingly, rather than simply rolling out new products or entering new markets and hoping for the best.[85] A robust practice like this could reduce or eliminate some of the phenomena described above, ranging from the proliferation of election-related disinformation to YouTube's tendency to recommend extreme content to unsuspecting users.

All systems are prone to error, and content moderation processes are no exception. Platform users should have access to timely and fair appeals processes to contest a platform's decision to remove or restrict their content. While the details of individual enforcement actions should be kept private, transparency reporting provides essential insight into how the company is addressing the challenges of the day. Facebook, Google, Microsoft, and Twitter have finally started to do so,[86] though their disclosures could be much more specific and comprehensive.[87] Notably, they should include data about the enforcement of ad content and targeting rules.

Our complete transparency and accountability standards can be found on our website. Key transparency recommendations for content shaping and content moderation are presented in the next section.

Key Transparency Recommendations for Content Shaping and Moderation

The recommendations below are drawn from the **RDR Corporate Accountability Index**, and reflect over a decade of civil society and academic research into platform accountability. Many of the companies that RDR ranks (including Facebook, Google, and Twitter) already meet some of these standards, but their disclosures are by no means comprehensive. As we have argued throughout this report, we don't know nearly enough about how companies use algorithmic systems to determine what we see online—and what we don't see. Congress should consider requiring internet platforms to disclose key information about their content shaping and content moderation practices as a first step toward potentially regulating the practices themselves.

Access to Key Policy Documents

- Companies should publish the rules (otherwise known as terms of service or community guidelines) for what user-generated content and behavior are or aren't permitted.

- Companies should publish the content rules for advertising (e.g., what kinds of products and services can or cannot be advertised, how ads should be formatted, and what kind of language may be prohibited in ads, such as curse words or vulgarity).

- Companies should publish the targeting rules for advertising (e.g., who users are, where they live, and what their interests are can be used to target ads).

Notification of Changes

- Companies should notify users when the rules for user-generated content, for advertising content, or for ad targeting change so that users can make an informed decision about whether to continue using the platform.

Rules and Processes for Enforcement

- Companies should disclose the processes and technologies (including content moderation algorithms) used to identify content or accounts that violate the rules for user-generated content, advertising content, and ad targeting.

- Companies should notify users when they make significant changes to these processes and technologies.

Transparency Reporting

- Companies should regularly publish transparency reports with data about the volume and nature of actions taken to restrict content that violates the rules for user-generated content, for advertising content, and for ad targeting.

- Transparency reports should be published at least once a year, preferably once a quarter.

Content-shaping Algorithms

- Companies should disclose whether they use algorithmic systems to curate, recommend, and/or rank the content that users can access through their platforms.

- Companies should explain how such algorithmic systems work, including what they optimize for and the variables they take into account.

- Companies should enable users to decide whether to allow these algorithms to shape their online experience, and to change the variables that influence them.

Conclusion

Companies have long accepted the need to moderate content, and to interface with policymakers and civil society about their content moderation practices, as a cost of doing business. It may be in their short-term commercial interest to keep the public debate at the level of content, without questioning the core assumptions of the surveillance-based, targeted advertising business model: scale, collection and monetization of user information, and the use of opaque algorithmic systems to shape users' experiences. But this focus may backfire if Section 230 is abolished or drastically changed. Furthermore, U.S. business leaders, investors, and consumers have been voicing growing expectations that the American economy should serve all parts of society, not just Big Tech's shareholders.[88] Companies that want to be considered responsible actors in society must make credible efforts to understand and mitigate the harms caused by their business models, particularly companies whose platforms have the power to shape public discourse and thereby our democracy.

Reliance on revenue from targeted advertising incentivizes companies to design platforms that are addictive, that manufacture virality, and that maximize the information that the company can collect about its users.[89] Policymakers and the American public are starting to understand this, but have not taken this insight to its logical conclusion: the business model needs to be regulated.

Reliance on revenue from targeted advertising incentivizes companies to design platforms that are addictive, that manufacture virality, and that maximize the information that the company can collect about its users.

Instead, as privacy bills languish in Congress, calls to reform Section 230 put the focus on interventions at the content level. Such reforms risk endangering free speech by incentivizing companies to remove much more user content than they currently do. They may not even address lawmakers' concerns, as much of the speech in question is protected by the First Amendment (like hate speech).

We have to pursue a different path, one that allows us to preserve freedom of expression and hold internet platforms accountable. Policymakers and activists alike must shift their focus to the power that troubling content can attain when it is plugged into the algorithmic and ad-targeting systems of companies like Google, Facebook, and Twitter. This is where regulatory efforts could truly shift our trajectory.

We will never be able to eliminate all violent extremism or disinformation online any more than we can eliminate all crime or national security threats in a city or country—at least not without sacrificing core American values like free speech, due process, and the rule of law. But we can drastically reduce the power of such content—its capacity to throw an election or bring about other kinds of real-life harm—if we focus on regulating companies' underlying data-driven (and money-making) technological systems and on good corporate governance. Our next report will do just that.

Notes

[1] Weissert, William, and Amanda Seitz. 2019. “False Claims Blur Line between Mass Shootings, 2020 Politics.” *AP NEWS*. https://apnews.com/bd653f4eb5ed4f34b6c936221c35a3e5

[2] Kelly, Makena. 2019. “Facebook, Twitter, and Google Must Remove Disinformation, Beto O’Rourke Demands.” *The Verge*. https://www.theverge.com/2019/9/6/20853447/facebook-twitter-google-beto-orourke-odessa-midland-texas-shooting-disinformation

[3] Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. New York, NY: Oxford University Press.

[4] Amnesty International. 2019. *Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights*. London: Amnesty International. https://amnestyusa.org/wp-content/uploads/2019/11/Surveillance-Giants-Embargo-21-Nov-0001-GMT-FINAL-report.pdf

[5] Gary, Jeff, and Ashkan Soltani. 2019. “First Things First: Online Advertising Practices and Their Effects on Platform Speech.” *Knight First Amendment Institute*. https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech

[6] Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition. New York: PublicAffairs.

[7] O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

[8] Companies are required to police and remove content that is not protected by the First Amendment, including child sexual abuse materials and content that violates copyright laws.

[9] Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.

[10] Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition. New York: PublicAffairs.

[11] Harwell, Drew. 2018. “AI Will Solve Facebook’s Most Vexing Problems, Mark Zuckerberg Says. Just Don’t Ask When or How.” *Washington Post*. https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

[12] It is worth nothing that terms like algorithms, machine learning and artificial intelligence have very specific meanings in computer science, but are used more or less interchangeably in policy discussions to refer to computer systems that use Big Data analytics to perform tasks that humans would otherwise do. Perhaps the most useful non-technical definition of an algorithm is Cathy O’Neil’s: “Algorithms are opinions embedded in code” (see O’Neil, Cathy. 2017. *The Era of Blind Faith in Big Data Must End*. https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end).

[13] Singh, Spandana. 2019. *Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds*. Washington D.C.: New America’s Open Technology Institute. https://www.newamerica.org/oti/reports/rising-through-ranks/

[14] Singh, Spandana. 2020. *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*. Washington D.C.: New America’s Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

[15] Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition. New York: PublicAffairs.

[16] Singh, Spandana. 2019. *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*. Washington D.C.: New America's Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

[17] Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt; Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7(1): 205395171989794.

[18] Newton, Casey. 2020. "YouTube Moderators Are Being Forced to Sign a Statement Acknowledging the Job Can Give Them PTSD." *The Verge*. https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health; Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

[19] Benkler, Yochai, Rob Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York, NY: Oxford University Press.

[20] U.S. Senate Select Committee on Intelligence. 2019a. *Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election Volume 1: Russian Efforts against Election Infrastructure*. Washington, D.C.: U.S. Senate. https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures; U.S. Senate Select Committee on Intelligence. 2019b. *Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election Volume 2: Russia 's Use of Social Media*. Washington, D.C.: U.S. Senate. https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures

[21] Many independent fact-checking efforts stemmed from this moment, but it is difficult to assess their effectiveness.

[22] Harwell, Drew. 2018. "AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How." *Washington Post*. https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

[23] Ghaffary, Shirin. 2019. "The Algorithms That Detect Hate Speech Online Are Biased against Black People." *Vox*. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter; Sap, Maarten et al. 2019. "The Risk of Racial Bias in Hate Speech Detection." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 1668–78. https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf; Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 25–35. https://www.aclweb.org/anthology/W19-3504.pdf

[24] Lawrence Alexander. "In the Fight against Pro-Kremlin Bots, Tech Companies Are Suspending Regular Users." 2018. *Global Voices*. https://globalvoices.org/2018/05/18/in-the-fight-against-pro-kremlin-bots-tech-companies-are-suspending-regular-users/

[25] Several companies (including Facebook, Google, and Twitter) did publish some information about content they removed as a result of a government request or a copyright claim. Ranking Digital Rights. 2015. *Corporate Accountability Index*. Washington, DC: New America. https://rankingdigitalrights.org/index2015/

[26] Ranking Digital Rights. 2019. *Corporate Accountability Index*. Washington, DC: New America. https://rankingdigitalrights.org/index2018/

[27] Transparency reports contain aggregate data about the actions that companies take to enforce their rules. See Budish, Ryan, Liz Woolery, and Kevin Bankston. 2016. The Transparency Reporting Toolkit. Washington D.C.: New America's Open Technology Institute. https://www.newamerica.org/oti/policy-papers/the-transparency-reporting-toolkit/

[28] Jeremy B. Merrill, Ariana Tobin. 2018. "Facebook's Screening for Political Ads Nabs News Sites Instead of Politicians." *ProPublica*. https://www.propublica.org/article/facebook-new-screening-system-flags-the-wrong-ads-as-political

[29] Eli Rosenberg. "Facebook Blocked Many Gay-Themed Ads as Part of Its New Advertising Policy, Angering LGBT Groups." *Washington Post*. https://www.washingtonpost.com/technology/2018/10/03/facebook-blocked-many-gay-themed-ads-part-its-new-advertising-policy-angering-lgbt-groups/

[30] Mak, Aaron. 2018. "Facebook Thought an Ad From Bush's Baked Beans Was 'Political' and Removed It." *Slate Magazine*. https://slate.com/technology/2018/05/bushs-baked-beans-fell-victim-to-facebooks-political-ads-system.html

[31] In a 2018 experiment, Cornell University professor Nathan Matias and a team of colleagues tested the two systems by submitting several hundred software-generated ads that were not election-related, but contained some information that the researchers suspected might get caught in the companies' filters. The ads were primarily related to national parks and Veterans' Day events. While all of the ads were approved by Google, 11 percent of the ads submitted to Facebook were prohibited, with a notice citing the company's election-related ads policy. See Matias, J. Nathan, Austin Hounsel, and Melissa Hopkins. 2018. "We Tested Facebook's Ad Screeners and Some Were Too Strict." *The Atlantic*. https://www.theatlantic.com/technology/archive/2018/11/do-big-social-media-platforms-have-effective-ad-policies/574609/

[32] Our research found that while Facebook, Google and Twitter did provide some information about their processes for reviewing ads for rules violations, they provided no data about what actions they actually take to remove or restrict ads that violate ad content or targeting rules once violations are identified. And because none of these companies include information about ads that were rejected or taken down in their transparency reports, it's impossible to evaluate how well their internal processes are working. Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising and Algorithmic Systems – Pilot Study and Lessons Learned*. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

[33] Popken, Ben. 2018. "As Algorithms Take over, YouTube's Recommendations Highlight a Human Problem." NBC News. https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596; Marwick, Alice, and Rebecca Lewis. 2017. Media Manipulation and Disinformation Online. New York, NY: Data & Society Research Institute. https://datasociety.net/output/media-manipulation-and-disinfo-online/; Caplan, Robin, and danah boyd. 2016. Who Controls the Public Sphere in an Era of Algorithms? New York, NY: Data & Society Research Institute. https://www.datasociety.net/pubs/ap/MediationAutomationPower_2016.pdf.

[34] Companies take different approaches to defining what constitutes “terrorist” content. This is a highly subjective exercise.

[35] Biddle, Ellery Roberts. “‘Envision a New War’: The Syrian Archive, Corporate Censorship and the Struggle to Preserve Public History Online.” *Global Voices*. https://globalvoices.org/2019/05/01/envision-a-new-war-the-syrian-archive-corporate-censorship-and-the-struggle-to-preserve-public-history-online/; MacDonald, Alex. “YouTube Admits ‘wrong Call’ over Deletion of Syrian War Crime Videos.” *Middle East Eye*. http://www.middleeasteye.net/news/youtube-admits-wrong-call-over-deletion-syrian-war-crime-videos

[36] Maréchal, Nathalie. 2019. “RDR Seeks Input on New Standards for Targeted Advertising and Human Rights.” Ranking Digital Rights. https://rankingdigitalrights.org/2019/02/20/rdr-seeks-input-on-new-standards-for-targeted-advertising-and-human-rights; Brouillette, Amy. 2019. “RDR Seeks Feedback on Standards for Algorithms and Machine Learning, Adding New Companies.” Ranking Digital Rights. https://rankingdigitalrights.org/2019/07/25/rdr-seeks-feedback-on-standards-for-algorithms-and-machine-learning-adding-new-companies We identified three areas where companies should be much more transparent about their use of algorithmic systems: advertising policies and their enforcement, content-shaping algorithms, and automated enforcement of content rules for users’ organic content (also known as content moderation). Ranking Digital Rights. *RDR Corporate Accountability Index: Draft Indicators. Transparency and Accountability Standards for Targeted Advertising and Algorithmic Decision-Making Systems.* Washington D.C.: New America.

[37] Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising and Algorithmic Systems – Pilot Study and Lessons Learned*. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

[38] Ochigame, Rodrigo. 2019. “The Invention of ‘Ethical AI’: How Big Tech Manipulates Academia to Avoid Regulation.” *The Intercept*. **https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/**; Vincent, James. 2019. “The Problem with AI Ethics.” *The Verge*. **https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech**.

[39] Facebook’s record in Myanmar, where it was found to have contributed to the genocide of the Rohingya minority, is a particularly egregious example. See Warofka, Alex. 2018. “An Independent Assessment of the Human Rights Impact of Facebook in Myanmar.” Facebook Newsroom. https://about.fb.com/news/2018/11/myanmar-hria/

[40] Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. New York, NY: Oxford University Press; Gary, Jeff, and Ashkan Soltani. 2019. “First Things First: Online Advertising Practices and Their Effects on Platform Speech.” *Knight First Amendment Institute*. https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech

[41] Lewis, Becca. 2020. “All of YouTube, Not Just the Algorithm, Is a Far-Right Propaganda Machine.” *Medium*. https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430

[42] Tufekci, Zeynep. 2018. “YouTube, the Great Radicalizer.” *The New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

[43] Unpublished study by Kaiser, Cordova & Rauchfleisch, 2019. The authors have not published the study out of concern that the algorithmic

manipulation techniques they describe could be replicated in order to exploit children.

[44] Fisher, Max, and Amanda Taub. 2019. “On YouTube’s Digital Playground, an Open Gate for Pedophiles.” *The New York Times*. https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

[45] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-30.

[46] Zuiderveen Borgesius, Frederik. 2018. *Discrimina tion, Artificial Intelligence, and Algorithmic Decision-Making*. Strasbourg: Council of Europe. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73; Wachter, Sandra. 2019. *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=3388639 Forthcoming in Berkeley Technology Law Journal, Vol. 35, No. 2, 2020.

[47] Angwin, Julia, and Terry Parris Jr. 2016. “Facebook Lets Advertisers Exclude Users by Race.” *ProPublica*. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[48] Julia Angwin, Terry Parris Jr. 2016. “Facebook Lets Advertisers Exclude Users by Race.” *ProPublica*. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[49] Julia Angwin, Madeleine Varner. 2017. “Facebook Enabled Advertisers to Reach ‘Jew Haters.’” *ProPublic a*. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

[50] In a March 2019 settlement, Facebook agreed to create a distinct advertising portal for housing, employment, and credit ads, as civil rights law prohibits discriminatory advertising in these areas. The company also committed to create a new “Housing Search Portal” allowing users to view all housing ads on the platform, regardless of whether the users are in the target audience selected by the advertiser.

[51] Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising and Algorithmic Systems – Pilot Study and Lessons Learned*. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

[52] *Fair Housing Act*. 42 U.S.C. § 3604(c). https://www.justice.gov/crt/fair-housing-act-2

[53] Anderson, Janna, and Lee Rainie. 2020. “Many Tech Experts Say Digital Disruption Will Hurt Democracy.” *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2020/02/21/many-tech-experts-say-digital-disruption-will-hurt-democracy/; McCarthy, Justin. 2020. *In U.S., Most Oppose Micro-Targeting in Online Political Ads*. Knight Foundation. https://knightfoundation.org/articles/in-us-most-oppose-micro-targeting-in-online-political-ads/

[54] This capability was used for voter suppression in both 2016 and 2018. The major platforms now prohibit ads that “are designed to deter or prevent people from voting,” but it is not at all clear how they will detect violations. See Hsu, Tiffany. 2018. “Voter Suppression and Racial Targeting: In Facebook’s and Twitter’s Words.” *The New York Times*. https://www.nytimes.com/2018/12/17/business/russia-voter-suppression-facebook-twitter.html; Leinwand, Jessica. 2018. “Expanding Our Policies on Voter Suppression.” *Facebook Newsroom*. https://about.fb.com/news/2018/10/voter-suppression-policies/

[55] Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising*

and Algorithmic Systems – Pilot Study and Lessons Learned. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

56 Zuckerberg, Mark. 2019. "Zuckerberg: Standing For Voice and Free Expression." Speech at Georgetown University, Washington D.C. https://www.washingtonpost.com/technology/2019/10/17/zuckerberg-standing-voice-free-expression/

57 Roose, Kevin, Sheera Frenkel, and Mike Isaac. 2020. "Don't Tilt Scales Against Trump, Facebook Executive Warns." The New York Times. https://www.nytimes.com/2020/01/07/technology/facebook-trump-2020.html

58 Wong, Julia Carrie. 2020. "One Year inside Trump's Monumental Facebook Campaign." The Guardian. https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election

59 Edsall, Thomas B. 2020. "Trump's Digital Advantage Is Freaking Out Democratic Strategists." *T he New York Times*. https://www.nytimes.com/2020/01/29/opinion/trump-digital-campaign-2020.html

60 The ad continued: "if Trump tries to lie in a TV ad, most networks will refuse to air it. But Facebook just cashes Trump's checks. Facebook already helped elect Donald Trump once. Now, they're deliberately allowing a candidate to intentionally lie to the American people. It's time to hold Mark Zuckerberg accountable – add your name if you agree." Epstein, Kayla. 2019. "Elizabeth Warren's Facebook Ad Proves the Social Media Giant Still Has a Politics Problem." *Washington Post*. https://www.washingtonpost.com/politics/2019/10/12/elizabeth-warrens-facebook-ad-proves-social-media-giant-still-has-politics-problem/

61 Isaac, Mike. 2019. "Dissent Erupts at Facebook Over Hands-Off Stance on Political Ads." *The New York Times.* https://www.nytimes.com/2019/10/28/technology/facebook-mark-zuckerberg-political-ads.html; The New York Times. 2019. "Read the Letter Facebook Employees Sent to Mark Zuckerberg About Political Ads." *The New York Times*. https://www.nytimes.com/2019/10/28/technology/facebook-mark-zuckerberg-letter.html

62 Glazer, Emily, Deepa Seetharaman, and Jeff Horwitz. 2019. "Peter Thiel at Center of Facebook's Internal Divisions on Politics." Wall Street Journal. https://www.wsj.com/articles/peter-thiel-at-center-of-facebooks-internal-divisions-on-politics-11576578601

63 Barrett, Bridget, Daniel Kreiss, Ashley Fox, and Tori Ekstrand. 2019. *Political Advertising on Platforms on the United States: A Brief Primer*. Chapel Hill: University of North Carolina. https://citapdigitalpolitics.com/wp-content/uploads/2020/01/PlatformAdvertisingPrimer_CITAP.pdf

64 Kovach, Steve. 2019. "Mark Zuckerberg vs. Jack Dorsey Is the Most Interesting Battle in Silicon Valley." *CNBC*. https://www.cnbc.com/2019/10/31/twitters-dorsey-calls-out-facebook-ceo-zuckerberg-on-political-ads.htm

65 Defined as "content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome." In the U.S., this applies to independent expenditure groups like PACs, Super PACs, and 501(c)(4) organizations.

66 Its policies prohibit advertisers from using zip codes or keywords and audience categories related to politics (like "conservative" or "liberal," or presumed interest in a specific candidate). Instead, issue ads can be targeted at the state, province, or regional level, or using keywords and audience categories that are unrelated to politics. See Barrett, Bridget, Daniel Kreiss, Ashley Fox, and Tori Ekstrand. 2019. *Political Advertising on Platforms in the United States: A Brief Primer*. Chapel Hill: University of North Carolina. https://citapdigitalpolitics.com/wp-

content/uploads/2020/01/PlatformAdvertisingPrimer_CITAP.pdf

67 Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising and Algorithmic Systems – Pilot Study and Lessons Learned*. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

68 Edsall, Thomas B. 2020. "Trump's Digital Advantage Is Freaking Out Democratic Strategists." *T he New York Times.* https://www.nytimes.com/2020/01/29/opinion/trump-digital-campaign-2020.html.

69 Rosenberg, Matthew. 2019. "Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised." *The New York Times*. https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html

70 Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.

71 Germany's Netzwerkdurchsetzungsgesetz ("Network Enforcemencement Act") went into effect in 2018.

72 Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet.* New York: Columbia Global Reports.

73 European Parliament. 2019. "Legislative Train Schedule." https://www.europarl.europa.eu/legislative-train

74 Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.

75 Kosseff, Jeff. 2019. *The Twenty-Six Words That Created the Internet*. Ithaca: Cornell University Press.

76 Citron, Danielle Keats, and Benjamin Wittes. 2017. "The Internet Will Not Break: Denying Bad Samaritans §230 Immunity." *Fordham Law Review* 86(2): 401–23.

77 See also Pírková, Eliška, and Javier Pallero. 2020. *26 Recommendations on Content Governance: A Guide for Lawmakers, Regulators, and Company Policy Makers*. Access Now. https://www.accessnow.org/recommendations-content-governance

78 Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.

79 Singh, Spandana. 2019. *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*. Washington D.C.: New America's Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/; Singh, Spandana. 2019. *Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds*. Washington D.C.: New America's Open Technology Institute. https://www.newamerica.org/oti/reports/rising-through-ranks/

80 Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising and Algorithmic Systems – Pilot Study and Lessons Learned*. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

81 Twitter. 2020. "Privacy Policy." https://twitter.com/content/twitter-com/legal/en/privacy.html (Accessed on February 20, 2020).

82 Ranking Digital Rights. 2020. *The RDR Corporate Accountability Index: Transparency and Accountability Standards for Targeted Advertising*

and Algorithmic Systems – Pilot Study and Lessons Learned. Washington D.C.: New America. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

[83] Ali, Muhammad et al. 2019. “Discrimination through Optimization: How Facebook’s Ad Delivery Can Lead to Biased Outcomes.” Proceedings of the ACM on Human-Computer Interaction 3(CSCW): 1–30.

[84] Weintraub, Ellen L. 2019. “Don’t Abolish Political Ads on Social Media. Stop Microtargeting.” *Washingt on Post*. https://www.washingtonpost.com/opinions/2019/11/01/dont-abolish-political-ads-social-media-stop-microtargeting/

[85] Allison-Hope, Dunstan. 2020. “Human Rights Assessments in the Decisive Decade: Applying UNGPs in the Technology Sector.” *Business for Social Responsibility*. https://www.bsr.org/en/our-insights/blog-view/human-rights-assessments-in-the-decisive-decade-ungp-challenges-technology

[86] Ranking Digital Rights. 2019. Corporate Accountability Index. Washington, DC: New America. https://rankingdigitalrights.org/index2019/

[87] In particular, Microsoft only reports requests from individuals to remove nonconsensual pornography, also referred to as “revenge porn,” which is the sharing of nude or sexually explicit photos or videos online without an individual’s consent. See “Content Removal Requests Report – Microsoft Corporate Social Responsibility.” Microsoft. https://www.microsoft.com/en-us/corporate-responsibility/crrr

[88] Business Roundtable. 2019. “Business Roundtable Redefines the Purpose of a Corporation to Promote ‘An Economy That Serves All Americans.’” **https://www.businessroundtable.org/business-roundtable-redefines-the-purpose-of-a-corporation-to-promote-an-economy-that-serves-all-americans** ; Behar, Andrew. 2019. “CEOs of World’s Largest Corporations: Shareholder Profit No Longer Sole Objective.” *As You Sow*. **https://www.asyousow.org/blog/business-roundtable-ceos-corporations-shareholder-value**

[89] Gary, Jeff, and Ashkan Soltani. 2019. “First Things First: Online Advertising Practices and Their Effects on Platform Speech.” *Knight First Amendment Institute*. https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech