Book Event
The McGannon Center presents

AntiSocial Media
Facebook and the Future of Democracy

Thursday, October 4, 2018
Costantino Room (Second Floor)
5 - 5:30 p.m. | check-in
5:30 - 6:45 p.m. | program
Reception to Follow

Featuring:

Siva Vaidhyanathan
Author of Antisocial Media: How Facebook Disconnects Us and Undermines Democracy; Professor of Media Studies and Director of the Center for Media and Citizenship, University of Virginia

Crystal Patterson
Global Civic Partnerships, Facebook

Julia Angwin
Editor-in-Chief and Co-Founder, The Markup; Award-winning Investigative Journalist, formerly of ProPublica and The Wall Street Journal

Danielle Keats Citron ’94
Visiting Professor of Law, Fordham Law School; Morton & Sophia Macht Professor of Law, Maryland University

Olivier Sylvain
Director, The McGannon Center; Professor of Law, Fordham Law School

CLE
Course Materials
Table of Contents

1. CLE Materials

Cohen, D. Facebook Will No Longer Allow Exclusion of Certain Groups Via It’s Ad Targeting Tools. (July 2018)

(see in document)

ProPublica. Facebook Lets Advertisers Exclude Users by Race

(see in document)

ProPublica. Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads

(see in document)

Sylvain, O. Discriminatory Designs on User Data. April 2018

(see in document)

Citron, D & Jurecic, Q. Platform Justice.

(see in document)

Facebook Community Standards

(see in document)
Facebook Will No Longer Allow Exclusion of Certain Groups Via Its Ad Targeting Tools

How will it affect legitimate, nondiscriminatory uses of those options?

By David Cohen

July 27, 2018
Facebook had earlier signed a legally binding agreement regarding targeting with Washington state.

Getty Images, Facebook

Facebook signed a legally binding agreement earlier this week with the office of Washington Attorney General Bob Ferguson in which it agreed to ensure that third-party advertisers cannot exclude ethnic and religious minorities, immigrants, LGBTQ individuals and other protected groups from seeing their ads, and to implement those policy changes within 90 days.

However, not all usage of those targeting capabilities is discriminatory in nature. Periscope vice president of media Jen Brady said, “Targeting by race is helpful when reaching dedicated ethnicities, such as with a Hispanic-focused campaign versus a general marketing campaign, or by gender. Being able to reach those audiences effectively is key. There are some advertisers who want to focus on religion.”

One agency executive, who wished to remain anonymous, said, “You can get around any rule. Facebook has a really interesting creative strategy
that can work well with advertising only if it is done in a creative way that doesn’t think it will limit them. People

However, another agency executive had a different experience, saying that a Hispanic-focused campaign for a paper towel manufacturer experienced delays due to having to go through an unanticipated review process, and adding, “It would be nice if our Facebook reps would say, ‘Hey, by the way, we know you guys typically do this, so you might want to build this into your timeline.’”

The 20-month investigation by Ferguson’s office (https://www.atg.wa.gov/news/news-releases/ag-ferguson-investigation-leads-facebook-making-nationwide-changes-prohibit) was prompted by an October 2016 story in ProPublica (https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race), by Julia Angwin and Terry Parris Jr., which found that advertisers could use the social network’s extensive ad targeting capabilities to ensure that people belonging to groups such as those mentioned above would not see their ads, potentially leading to discriminatory situations in sectors such as housing.

Facebook responded in November 2016 (https://www.adweek.com/digital/facebook-removes-ethnic-ad-targeting-option-housing-employment-and-credit-ads-174585/) with steps including:
3rd party ads directly into its ad-buying software and fully remove its affinity-marketing tools when marketers buy certain types of ads that have historically faced discrimination,” such as housing, employment and credit.

Moving the ethnic-affinity section in its ad-buying software to its behaviors section from its demographics section.

Updating its advertising policies website to make affinity-marketing issues “front and center for all businesses” and making it a bigger part of its education process for businesses and ad partners.

However, a follow-up last November by Angwin, Ariana Tobin and Madeleine Varner of ProPublica (https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin) revealed that the three reporters were able to buy “dozens of rental housing ads” on Facebook and exclude people such as African Americans, Jews and people interested in wheelchair ramps from seeing those ads, with ProPublica saying, “Every single ad was approved within minutes.”

Facebook then said it would temporarily block advertisers from excluding racial and ethnic groups from seeing ads and conduct an audit of how advertisers used those capabilities.

However, the ongoing investigation by Ferguson’s office found that advertisers were still able to exclude people from seeing their ads based on “several other protected classes,” such as sexual
In addition to housing, credit and employment ads, Facebook will no longer provide advertisers with options to exclude ethnic groups from advertisements for insurance and public accommodations. Public accommodations include all businesses open to the public. Places of public accommodation range from auto dealers, beauty salons and restaurants to colleges, hospitals and professional sports stadiums.

Facebook will no longer provide advertisers with tools to discriminate based on race, creed, color, national origin, veteran or military status, sexual orientation and disability status. These exclusion options will not be present on any advertisement for employment, housing, credit, insurance and/or places of public accommodation. These changes will be permanent and legally binding. The assurance of discontinuance requires Facebook to remove these exclusion options for any advertiser whose ad Washingtonians may have the ability to see. According to Facebook, these changes to its platform will be nationwide. According to the assurance of discontinuance, Facebook will fix its advertising platform to remove the unlawful targeting options within 90 days. The social network service also will pay the Washington State Attorney General’s Office $90,000 in costs and fees.
The first anonymous agency executive said brands can turn to lookalike audiences (https://www.adweek.com/digital/facebook-debuted-3-new-ad-products-for-retailers-at-shoptalk/), which are “much more sophisticated and work better in general” than the targeting capabilities affected by this week’s ruling, adding of the latter, “You spend so much more to do that—get super neat with the way you target.”

And Brady said, “Even if you take away those parameters of self-identification, you can still scrub Facebook to find those who might affiliate with one religion over another.”

Ferguson said in a statement, “Facebook’s advertising platform allowed unlawful discrimination on the basis of race, sexual orientation, disability and religion. That’s wrong, illegal, and unfair.”

David Cohen is editor of Adweek’s Social Pro Daily.
Video: What We Can Learn From the Impact of Streetwear on Retail

by Nick Gardner
Marvel Stars Join Parkland Activists in Sharing Stories of Their 'First Time' (Voting, That Is)

by David Griner

Spotify Celebrates Its 10th Anniversary and Sets Sights on Solidifying Role as a Leader in Music Streaming

by Adweek Staff
IKEA'S 'PEE AD' WINS GRAND Clio, ALONGSIDE PALAU PLEDGE, WENDY'S MIXTAPE AND MORE

by David Griner
Without a Brand Owning the Toy Category, Is It Possible to Successfully Revive Toys R Us?

by Kristina Monllos

4 Lessons From Challenger Brands Bumble, Flipboard and More on How They Usurped a Market

by Ann-Marie Alcántara
CNN Launches New Website and San Francisco Office

by Sara Jerde
MACHINE BIAS

Facebook Lets Advertisers Exclude Users by Race

Facebook’s system allows advertisers to exclude black, Hispanic, and other “ethnic affinities” from seeing ads.

by Julia Angwin and Terry Parris Jr., Oct. 28, 2016, 1 p.m. EDT

Imagine if, during the Jim Crow era, a newspaper offered advertisers the option of placing ads only in copies that went to white readers.

That’s basically what Facebook is doing nowadays.

The ubiquitous social network not only allows advertisers to target users by their interests or background, it also gives advertisers the ability to exclude specific groups it calls “Ethnic Affinities.” Ads that exclude people based on race, gender and other sensitive factors are prohibited by federal law in housing and employment.

Here is a screenshot of an ad we purchased in Facebook’s housing categories via the company’s advertising portal <https://www.facebook.com/business>.
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. (Here's the ad itself <https://www.propublica.org/documents/item/3191165-Facebook-Propublica-Ad.html>.)

When we showed Facebook’s racial exclusion options to a prominent civil rights lawyer John Relman <http://www.relmanlaw.com/attorneys/relman.php>, he gasped and said, “This is horrifying. This is massively illegal. This is about as blatant a violation of the federal Fair Housing Act as one can find.”

The Fair Housing Act of 1968 makes it illegal <https://www.law.cornell.edu/uscode/text/42/3604> "to make, print, or publish, or cause to be made, printed, or published any notice, statement, or advertisement, with respect to the sale or rental of a dwelling that indicates any preference, limitation, or discrimination based on race, color, religion, sex, handicap, familial status, or national origin.” Violators can face tens of thousands of dollars in fines <http://www.fairhousingnc.org/2014/justice-department-increases-civil-monetary-penalties-for-fair-housing-violations/>.

The Civil Rights Act of 1964 <https://www.eeoc.gov/laws/statutes/titlevii.cfm> also prohibits the “printing or publication of notices or advertisements indicating prohibited preference, limitation, specification or discrimination” in employment recruitment.

Facebook’s business model is based on allowing advertisers to target specific groups — or, apparently to exclude specific groups — using huge
reams of personal data the company has collected about its users. Facebook’s microtargeting is particularly helpful for advertisers looking to reach niche audiences, such as swing-state voters concerned about climate change. ProPublica recently offered a tool allowing users to see how Facebook is categorizing them. We found nearly 50,000 unique categories in which Facebook places its users.

Facebook says its policies prohibit advertisers from using the targeting options for discrimination, harassment, disparagement or predatory advertising practices.

“We take a strong stand against advertisers misusing our platform: Our policies prohibit using our targeting options to discriminate, and they require compliance with the law,” said Steve Satterfield, privacy and public policy manager at Facebook. “We take prompt enforcement action when we determine that ads violate our policies.”

Satterfield said it’s important for advertisers to have the ability to both include and exclude groups as they test how their marketing performs. For instance, he said, an advertiser “might run one campaign in English that excludes the Hispanic affinity group to see how well the campaign performs against running that ad campaign in Spanish. This is a common practice in the industry.”

He said Facebook began offering the “Ethnic Affinity” categories within the past two years as part of a “multicultural advertising” effort.

Satterfield added that the “Ethnic Affinity” is not the same as race — which Facebook does not ask its members about. Facebook assigns members an “Ethnic Affinity” based on pages and posts they have liked or engaged with on Facebook.

When we asked why “Ethnic Affinity” was included in the “Demographics” category of its ad-targeting tool if it’s not a representation of demographics, Facebook responded that it plans to move “Ethnic Affinity” to another section.

Facebook declined to answer questions about why our housing-categories ad excluding minority groups was approved 15 minutes after we placed the order.

---

**Breaking the Black Box**

We live in an era of increasing automation. But as machines make more decisions for us, it is increasingly important to understand the algorithms that produce their judgments.

See the series.


**Machine Bias**
By comparison, consider the advertising controls that the New York Times has put in place to prevent discriminatory housing ads. After the newspaper was successfully sued under the Fair Housing Act in 1989, it agreed to review ads for potentially discriminatory content before accepting them for publication.

Steph Jespersen, the Times’ director of advertising acceptability, said that the company’s staff runs automated programs to make sure that ads that contain discriminatory phrases such as “whites only” and “no kids” are rejected.

The Times’ automated program also highlights ads that contain potentially discriminatory code words such as “near churches” or “close to a country club.” Humans then review those ads before they can be approved.

Jespersen said the Times also rejects housing ads that contain photographs of too many white people. The people in the ads must represent the diversity of the population of New York, and if they don’t, he says he will call up the advertiser and ask them to submit an ad with a more diverse lineup of models.

But, Jespersen said, these days most advertisers know not to submit discriminatory ads: “I haven’t seen an ad with ‘whites only’ for a long time.”

**Clarification, Oct. 28, 2016:** We’ve updated the story to explain more clearly that the ad we bought was not for housing itself — it was placed in Facebook’s housing categories.

---

**Julia Angwin**

Julia Angwin is a senior reporter at ProPublica. From 2000 to 2013, she was a reporter at The Wall Street Journal, where she led a privacy investigative team that was a finalist for a Pulitzer Prize in Explanatory Reporting in 2011 and won a Gerald Loeb Award in 2010.

Email: Julia.Angwin@propublica.org  
Twitter: @JuliaAngwin

---

**Terry Parris Jr.**

Terry is ProPublica’s deputy editor, engagement. Prior to joining ProPublica, he led digital production and engagement at WDET 101.9 FM, NPR’s affiliate in Detroit.
MACHINE BIAS

Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads

Among the companies we found doing it: Amazon, Verizon, UPS and Facebook itself. “It’s blatantly unlawful,” said one employment law expert.


Mark Edelstein, a social media marketing strategist who is also legally blind, says he never had serious trouble finding a job until he turned 50. (Whitney Curtis for The New York Times)

This story was co-published with The New York Times.

A few weeks ago, Verizon placed an ad on Facebook to recruit applicants for a unit focused on financial planning and analysis. The ad showed a smiling, millennial-aged woman seated at a computer and promised that new hires could look forward to a rewarding career in which they would be “more than just a number.”

Some relevant numbers were not immediately evident. The promotion was set to run on the Facebook feeds of users 25 to 36 years old who lived in the nation’s capital, or had recently visited there, and had demonstrated an
interest in finance. For a vast majority of the hundreds of millions of people who check Facebook every day, the ad did not exist.

Verizon

This Verizon recruiting ad appeared on Facebook last month, seeking financial analysts aged 25 to 36. Verizon didn’t respond to repeated requests for comment about this ad.

Verizon is among dozens of the nation’s leading employers — including Amazon, Goldman Sachs, Target and Facebook itself — that placed recruitment ads limited to particular age groups, an investigation by ProPublica and The New York Times has found.

The ability of advertisers to deliver their message to the precise audience most likely to respond is the cornerstone of Facebook’s business model. But using the system to expose job opportunities only to certain age groups has raised concerns about fairness to older workers.

Several experts questioned whether the practice is in keeping with the federal Age Discrimination in Employment Act of 1967, which prohibits bias against people 40 or older in hiring or employment. Many jurisdictions make it a crime to “aid” or “abet” age discrimination, a provision that could apply to companies like Facebook that distribute job ads.

“It’s blatantly unlawful,” said Debra Katz, a Washington employment lawyer who represents victims of discrimination.
Facebook defended the practice <https://newsroom.fb.com/news/h/addressing-targeting-in-recruitment-ads/>. “Used responsibly, age-based targeting for employment purposes is an accepted industry practice and for good reason: it helps employers recruit and people of all ages find work,” said Rob Goldman, a Facebook vice president.

The revelations come at a time when the unregulated power of the tech companies is under increased scrutiny, and Congress is weighing whether to limit the immunity that it granted to tech companies in 1996 for third-party content on their platforms.

Facebook has argued in court filings that the law, the Communications Decency Act, makes it immune from liability for discriminatory ads.

Although Facebook is a relatively new entrant into the recruiting arena, it is rapidly gaining popularity with employers. Earlier this year, the social network launched a section of its site devoted to job ads <https://www.facebook.com/business/news/take-the-work-out-of-hiring>. Facebook allows advertisers to select their audience, and then Facebook finds the chosen users with the extensive data it collects about its members.

The use of age targets emerged in a review of data originally compiled by ProPublica readers for a project about political ad placement <https://www.propublica.org/article/help-us-monitor-political-ads-online> on Facebook. Many of the ads include a disclosure by Facebook about why the user is seeing the ad, which can be anything from their age to their affinity for folk music.

The precision of Facebook’s ad delivery has helped it dominate an industry once in the hands of print and broadcast outlets. The system, called microtargeting, allows advertisers to reach essentially whomever they prefer, including the people their analysis suggests are the most plausible hires or consumers, lowering the costs and vastly increasing efficiency.

Targeted Facebook ads were an important tool in Russia’s efforts to influence the 2016 election. The social media giant has acknowledged that 126 million people <https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html> saw Russia-linked content, some of which was aimed at particular demographic groups and regions. Facebook has also come under criticism for the disclosure that it accepted ads aimed at “Jew-haters” <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters> as well as housing ads <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race> that discriminated by race, gender, disability and other factors.

Other tech companies also offer employers opportunities to discriminate by age. ProPublica bought job ads on Google and LinkedIn that excluded audiences older than 40 — and the ads were instantly approved. Google said it does not prevent advertisers from displaying ads based on the user’s
age. After being contacted by ProPublica, LinkedIn changed its system to prevent such targeting in employment ads.

The practice has begun to attract legal challenges. On Wednesday, a class-action complaint alleging age discrimination was filed in federal court <https://www.documentcloud.org/documents/4334241-Outten-Golden-Complaint-FILED.html> in San Francisco on behalf of the Communications Workers of America and its members — as well as all Facebook users 40 or older who may have been denied the chance to learn about job openings. The plaintiffs’ lawyers said the complaint was based on ads for dozens of companies that they had discovered on Facebook.

The database of Facebook ads collected by ProPublica shows how often and precisely employers recruit by age. In a search for “part-time package handlers,” United Parcel Service ran an ad aimed at people 18 to 24. State Farm <https://projects.propublica.org/graphics/facebook-job-ads#StateFarm> pitched its hiring promotion to those 19 to 35.

Some companies, including Target, State Farm and UPS, defended their targeting as a part of a broader recruitment strategy that reached candidates of all ages. The group of companies making this case included Facebook itself, which ran career ads on its own platform, many aimed at people 25 to 60. “We completely reject the allegation that these advertisements are discriminatory,” said Goldman of Facebook.

After being contacted by ProPublica and the Times, other employers, including Amazon, Northwestern Mutual <https://projects.propublica.org/graphics/facebook-job-ads#NorthwesternMutual> and the New York City Department of Education <https://projects.propublica.org/graphics/facebook-job-ads#NYCteachingFellows>, said they had changed or were changing their recruiting strategies.

“We recently audited our recruiting ads on Facebook and discovered some had targeting that was inconsistent with our approach of searching for any candidate over the age of 18,” said Nina Lindsey, a spokeswoman for Amazon, which targeted some ads for workers at its distribution centers between the ages of 18 and 50. “We have corrected those ads.”

Verizon did not respond to requests for comment.

Several companies argued that targeted recruiting on Facebook was comparable to advertising opportunities in publications like the AARP magazine or Teen Vogue, which are aimed at particular age groups. But
this obscures an important distinction. Anyone can buy Teen Vogue and see an ad. Online, however, people outside the targeted age groups can be excluded in ways they will never learn about.

“What happens with Facebook is you don’t know what you don’t know,” said David Lopez, a former general counsel for the Equal Employment Opportunity Commission who is one of the lawyers at the firm Outten & Golden bringing the age-discrimination case on behalf of the communication workers union.

‘They Know I’m Dead’

Age discrimination on digital platforms is something that many workers suspect is happening to them, but that is often difficult to prove.

Mark Edelstein at his home in St. Louis, Missouri (Whitney Curtis for The New York Times)

Mark Edelstein, a fitfully employed social-media marketing strategist who is 58 and legally blind, doesn’t pretend to know what he doesn’t know, but he has his suspicions.

Edelstein, who lives in St. Louis, says he never had serious trouble finding a job until he turned 50. “Once you reach your 50s, you may as well be dead,” he said. “I’ve gone into interviews, with my head of gray hair and my receding hairline, and they know I’m dead.”

Edelstein spends most of his days scouring sites like LinkedIn and Indeed and pitching hiring managers with personalized appeals. When he scrolled through his Facebook ads on a Wednesday in December, he saw a variety of ads reflecting his interest in social media marketing: ads for the marketing software HubSpot (https://www.hubspot.com/) (“15 free infographic templates!”) and TripIt, which he used to book a trip to visit his mother in
Florida.

What he didn’t see was a single ad for a job in his profession, including one identified by ProPublica that was being shown to younger users: a posting for a social media director job at HubSpot.

This HubSpot ad aimed at people aged 27 to 40 appeared on Facebook in November. The company said that this ad was a mistake and that the company would not use the age-targeting feature again.

The company asked that the ad be shown to people aged 27 to 40 who live or were recently living in the United States.

“Hypothetically, had I seen a job for a social media director at HubSpot, even if it involved relocation, I ABSOLUTELY would have applied for it,” Edelstein said by email when told about the ad.

A HubSpot spokeswoman, Ellie Botelho, said that the job was posted on many sites, including LinkedIn, The Ladders and Built in Boston, and was open to anyone meeting the qualifications regardless of age or any other demographic characteristic.

She added that “the use of the targeted age-range selection on the Facebook ad was frankly a mistake on our part given our lack of experience using that platform for job postings and not a feature we will use again.”

For his part, Edelstein says he understands why marketers wouldn’t want to target ads at him: “It doesn’t surprise me a bit. Why would they want a 58-year-old white guy who’s disabled?”
Looking for 'Younger Blood'

Although LinkedIn is the leading online recruitment platform, according to an annual survey by SourceCon, an industry website. Facebook is rapidly increasing in popularity for employers.

One reason is that Facebook’s sheer size — two billion monthly active users, versus LinkedIn’s 530 million total members — gives recruiters access to types of workers they can’t find elsewhere.

Consider nurses, whom hospitals are desperate to hire. “They’re less likely to use LinkedIn,” said Josh Rock, a recruiter at a large hospital system in Minnesota who has expertise in digital media. “Nurses are predominantly female, there’s a larger volume of Facebook users. That’s what they use.”

There are also millions of hourly workers who have never visited LinkedIn, and may not even have a résumé, but who check Facebook obsessively.

Deb Andrychuk, chief executive of the Arland Group, which helps employers place recruitment ads, said clients sometimes asked her firm to target ads by age, saying they needed “to start bringing younger blood” into their organizations. “It’s not necessarily that we wouldn’t take someone older,” these clients say, according to Andrychuk, “but if you could bring in a younger set of applicants, it would definitely work out better.”

Andrychuk said that “we coach clients to be open and not discriminate” and that after being contacted by The Times, her team updated all their ads to ensure they didn’t exclude any age groups.

But some companies contend that there are permissible reasons to filter audiences by age, as with an ad for entry-level analyst positions at Goldman Sachs that was distributed to people 18 to 64. A Goldman Sachs spokesman, Andrew Williams, said showing it to people above that age range would have wasted money: roughly 25 percent of those who typically click on the firm’s untargeted ads are 65 or older, but people that age almost never apply for the analyst job.
“We welcome and actively recruit applicants of all ages,” Williams said. “For some of our social-media ads, we look to get the content to the people most likely to be interested, but do not exclude anyone from our recruiting activity.”

Pauline Kim, a professor of employment law at Washington University in St. Louis, said the Age Discrimination in Employment Act, unlike the federal anti-discrimination statute that covers race and gender, allows an employer to take into account “reasonable factors” that may be highly correlated with the protected characteristic, such as cost, as long as they don’t rely on the characteristic explicitly.

**The Question of Liability**

In various ways, Facebook and LinkedIn have acknowledged at least a modest obligation to police their ad platforms against abuse.

Earlier this year, Facebook said it would require advertisers to “self-certify” that their housing, employment and credit ads were compliant with anti-discrimination laws, but that it would not block marketers from purchasing age-restricted ads.

Still, Facebook didn’t promise to monitor those certifications for accuracy. And Facebook said the self-certification system, announced in February, was still being rolled out to all advertisers.

LinkedIn, in response to inquiries by ProPublica, added a self-certification step that prevents employers from using age ranges when placing an employment ad, unless they affirm the ad is not discriminatory.

With these efforts evolving, legal experts say it is unclear how much liability the tech platforms could have. Some civil rights laws, like the Fair Housing Act, explicitly require publishers to assume liability for discriminatory ads.

But the Age Discrimination in Employment Act assigns liability only to employers or employment agencies, like recruiters and advertising firms.

The lawsuit filed against Facebook on behalf of the communications workers argues that the company essentially plays the role of an employment agency — collecting and providing data that helps employers locate candidates, effectively coordinating with the employer to develop
the advertising strategies, informing employers about the performance of
the ads, and so forth.

Regardless of whether courts accept that argument, the tech companies
could also face liability under certain state or local anti-discrimination
statutes. For example, California’s Fair Employment and Housing Act
makes it unlawful to “aid, abet, incite, compel or coerce the doing” of
discriminatory acts proscribed by the statute.

“They may have an obligation there not to aid and abet an ad that enables
discrimination,” said Cliff Palefsky, an employment lawyer based in San
Francisco.

The question may hinge on Section 230 of the federal Communications
Decency Act, which protects internet companies from liability for third-
party content.

Tech companies have successfully invoked this law to avoid liability for
offensive or criminal content — including sex trafficking
<http://caselaw.findlaw.com/us-1st-circuit/1728752.html>, revenge porn
<https://law.justia.com/cases/federal/district-courts/new-
york/nysdae/1:2017cv00932/468549/21/> and calls for violence against Jews
<http://caselaw.findlaw.com/us-dc-circuit/1669834.html>. Facebook is currently
arguing in federal court that Section 230 immunizes it against liability for ad placement that blocks members of
certain racial and ethnic groups from seeing the ads.

“Advertisers, not Facebook, are responsible for both the content of their
ads and what targeting criteria to use, if any,” Facebook argued in its
motion to dismiss <https://www.documentcloud.org/documents/4333515-Outten-FB-
FB-Motion-to-Discard-4-3-17.html> allegations that its ads violated a host of civil
rights laws. The case does not allege age discrimination.

Eric Goldman, professor and co-director of the High Tech Law Institute at
the Santa Clara University School of Law, who has written extensively
about Section 230, says it is hard to predict how courts would treat
Facebook’s age-targeting of employment ads.

Goldman said the law covered the content of ads, and that courts have
made clear that Facebook would not be liable for an advertisement in
which an employer wrote, say, “no one over 55 need apply.” But it is not
clear how the courts would treat Facebook’s offering of age-targeted
customization.

According to a federal appellate court decision in a fair-housing case
<http://caselaw.findlaw.com/us-9th-circuit/1493375.html>, a platform can be
considered to have helped “develop unlawful content” that users play a
role in generating, which would negate the immunity.
“Depending on how the targeting is happening, you can make potentially different sorts of arguments about whether or not Google or Facebook or LinkedIn is contributing to the development” of the ad, said Deirdre K. Mulligan, a faculty director of the Berkeley Center for Law and Technology.

*Jeff Larson and Madeleine Varner contributed research.*

Want to help monitor ads on Facebook? Download our tool for [Firefox](https://addons.mozilla.org/en-US/firefox/addon/facebook-ad-collector/) or [Chrome](https://chrome.google.com/webstore/detail/facebook-political-ad-col/eniecaalhkkhinhcmnbqfmnmjikjcin?hl=en) web browsers.

**Julia Angwin**
Julia Angwin is a senior reporter at ProPublica. From 2000 to 2013, she was a reporter at The Wall Street Journal, where she led a privacy investigative team that was a finalist for a Pulitzer Prize in Explanatory Reporting in 2011 and won a Gerald Loeb Award in 2010.

✉️ Julia.Angwin@propublica.org  ➤️ @JuliaAngwin

**Ariana Tobin**
Ariana is an engagement reporter at ProPublica, where she works to cultivate communities to inform our coverage.

✉️ ariana.tobin@propublica.org  ➤️ @Ariana_Tobin
Discriminatory Designs on User Data

Olivier Sylvain
April 2018

The stated aim of online intermediaries like Facebook, Twitter, and Airbnb is to provide the platforms through which users freely meet people, purchase products, and discover information.¹ As “conduits” for speech and commerce,² intermediaries such as these are helping to create a more vibrant and democratic marketplace for goods and ideas than any the world has seen before.³

That, at least, is the theory on which Congress enacted Section 230 of the Communications Decency Act (CDA) in 1996.⁴ One of the central objectives of Section 230’s drafters was to ensure that intermediaries are “unfettered” by the obligation to police third-party user content.⁵ They believed that conventional tort principles and regulatory rules were simply not workable in an environment in which so much user content flows,⁶ and they doubted that intermediaries would be able to create new value for users if they constantly had to monitor, block, or remove illicit content. In the words of free speech doctrine, members of Congress worried the intermediaries would be “chilled” by the fear that they could be held legally responsible for content posted by users.⁷

Section 230 of the CDA therefore protects intermediaries from liability for distributing third-party user content. Courts have read Section 230 broadly, creating an immunity for intermediaries who do all but “materially contribute” to the user content they distribute.⁸ That is, courts have read the statute’s protections to cover services that “augment[]” user content, but not services that demonstrably “help” to develop the alleged illegal expressive conduct.⁹ Many believe that the internet would not be as dynamic and beguiling today were it not for the protection that Section 230 has been construed to provide for online intermediaries.¹⁰

This may be true. But Section 230 doctrine has also had a perverse effect. By providing intermediaries with such broad legal protection, the courts’ construction of Section 230 effectively underwrites content that foreseeably targets the most vulnerable among us. In their ambition to encourage an “unfettered” market for online speech, the developers of Section 230 immunity have set up a regime that makes online engagement more difficult for children, women, racial minorities, and other predictable targets of harassment and discriminatory expressive conduct. Examples abound: the gossip site that enabled users to anonymously post salacious images of unsuspecting young women;¹¹ the social media site through which an adult male lured a young teenage girl into a sexual assault;¹² the classifieds site that has allegedly facilitated the sex trafficking of minors;¹³ the online advertising platform that allows companies to exclude Latinos from apartment rentals and older people from job postings;¹⁴ the unrelenting social media abuse of feminist media critics¹⁵ and a prominent black female comedian;¹⁶ the live video stream of a gang rape of a teenage girl.¹⁷

The standard answer to the charge that current immunity doctrine enables these acts is that the originators of the illicit content are to blame, not the “neutral” services that facilitate online interactions.¹⁸ Intermediaries, this position holds, merely pass along user speech; they do not encourage its production or dissemination, and, in any case, Section 230 immunity exists to protect against a different problem: the “collateral censorship” of lawful content.¹⁹
This answer, however, is either glib or too wedded to an obsolete conception of how online intermediaries operate. Intermediaries today do much more than passively distribute user content or facilitate user interactions. Many of them elicit and then algorithmically sort and repurpose the user content and data they collect. The most powerful services also leverage their market position to trade this information in ancillary or secondary markets.  

Intermediaries, moreover, design their platforms in ways that shape the form and substance of their users’ content. Intermediaries and their defenders characterize these designs as substantively neutral technical necessities, but as I explain below, recent developments involving two of the most prominent beneficiaries of Section 230 immunity, Airbnb and Facebook, suggest otherwise. Airbnb and Facebook have enabled a range of harmful expressive acts, including violations of housing and employment laws, through the ways in which they structure their users’ interactions. At a minimum, companies should not get a free pass for enabling unlawful discriminatory conduct, regardless of the social value their services may otherwise provide. But more than this, I argue here, Section 230 doctrine requires a substantial reworking if the internet is to be the great engine of democratic engagement and creativity that it should be. Section 230 is no longer serving all the purposes it was meant to serve. The statute was intended at least in part to ensure the vitality and diversity, as well as the volume, of speech on new communications platforms. By allowing intermediaries to design their platforms without internalizing the costs of the illegal speech and conduct they facilitate, however, the statute is having the opposite effect.

This paper has four parts. The first discusses the basic contours of the prevailing doctrine, including the legislative purposes behind Section 230 and the logic courts have relied on to support broad immunity for intermediaries. The second part identifies ways in which the doctrine, in assuming that intermediaries are passive disseminators of information, may accelerate the mass distribution of content that harms vulnerable people and members of historically subordinated groups. I focus in particular on the distribution of nonconsensual pornography as a species of content that not only exacts a discrete reputational or privacy toll on victims but also fuels the circulation of misogynist views that harm young women in particular.

The third part of the paper turns to the designs that intermediaries employ to structure and enhance their users’ experience, and how these designs themselves can further discrimination. While the implications of this analysis reach beyond injuries to historically marginalized groups, my goal is to explain how the designs employed by two of the most prominent intermediaries today, Airbnb and Facebook, have enabled unlawful discrimination. The fourth and final part of the paper proposes a reform to the doctrine: I argue that courts should account for the specific ways in which intermediaries’ designs do or do not enable or cause harm to the predictable targets of discrimination and harassment. As recent developments underscore, Section 230 immunity doctrine must be brought closer in line with longstanding equality and universality norms in communications law.

I. Section 230 Immunity: A Brief Overview

The immunity that intermediaries enjoy under Section 230 of the CDA has helped to bring about the teeming abundance of content in today’s online environment. The prevailing interpretation of Section 230, bars courts from imposing liability on intermediaries that are the “mere conduits” through which user-generated content passes. This doctrine protects services that host all kinds of content — everything from customer product reviews to fake news to dating profiles.

Congress invoked a very old concept when it drafted this law. The central provision of Section 230, titled “Protection for ‘Good Samaritan’ blocking and screening of offensive material,” resembles laws in all the states that in one way or another shield defendants from liability arising from their good-faith efforts to help those in distress. Good Samaritan laws are inspired by the Biblical parable that praises the do-gooder who risks ridicule and censure to help a stranger left for dead.
Section 230’s drafters applied this concept to online activity. They created an exception under tort law, which traditionally holds publishers liable for distributing material they know to be unlawful, but does not hold them liable if they lack notice about the illegality of the communicative act at issue. Proponents of Section 230 worried that, without this legislation, claims for secondary liability would either stifle expressive conduct in the then-nascent medium or discourage intermediaries from policing content altogether. They further insisted that government regulators such as the Federal Communications Commission should play no role in deciding what sorts of content prevailed online; viewers (and their parents) should make those decisions for themselves.

While an interest in both free speech and the Good Samaritan concept drove Congress to enact Section 230, courts interpreting the statute have been far more influenced by the free speech concerns. In contrast to the nuanced requirements of the Digital Millennium Copyright Act’s notice-and-takedown regime, online intermediaries have not been required under Section 230 to block or screen offensive material in any particular way. Today, Section 230 doctrine provides a near-blanket immunity to intermediaries for hosting tortious third-party content. Long-established internet companies like America Online and Craigslist that host massive amounts of user content have been clear beneficiaries. Relying on Section 230, courts have immunized them from liability for everything from defamatory posts on electronic bulletin boards to racially discriminatory solicitations in online housing advertisements. Leading opinions have reasoned that the scale at which third-party content passes through online services makes that content infeasible to moderate; requiring services to try would not only chill online speech but also stunt the internet’s development as a transformative medium of communication. This immunity now applies to a wide range of online services that host and distribute user content, including Twitter’s microblogging service, Facebook’s flagship social media platform, and Amazon’s online marketplace. Thanks to Section 230, these companies have no legal obligation to block or remove mendacious tweets, fraudulent advertisements, or anticompetitive customer reviews by rivals.

As a result, most targets of illicit online user content in the United States have little to no effective recourse under law to have that content blocked or removed. They can sue the original posters of the content. But such litigation often presents serious challenges, including the cost of bringing a lawsuit, the difficulty of discovering the identities of anonymous posters, and, even if the suit is successful on the merits, the difficulty of obtaining remedies that are commensurate with the harm. Targets can also enlist services like search engine optimizers that make it harder to find the offending material. They can complain to the intermediaries about offending posts. And they can press intermediaries to improve their policies generally. If none of these strategies succeeds, users can boycott the service, as many people did recently — for one day — to protest the failure of Twitter to protect women from “verbal harassment, death threats, and doxing.” Even if effective, however, this last option sometimes feels far from optimal, given that the promise of the internet is understood to lie in its unrivaled opportunities for commercial engagement and social integration. Exit would only exacerbate extant disparities.

The threat of losing consumers, it must be said, is potent enough to have moved many intermediaries to develop content-governance protocols and automated systems for content detection. Even though Section 230 doctrine has removed any legal duty to moderate third-party content, certain companies routinely block or remove content when its publication detracts from the character of the service they mean to provide. And so, for instance, Google demotes or delists search engine optimizers and sites that host “fake news” and offensive content. Facebook removes clickbait articles and has now partnered with fact-checking organizations like Snopes and PolitiFact to implement a notification process for removing “fake news.”

The reform that the news aggregation and discussion site Reddit undertook in 2015 is especially striking in this regard. Reddit, which had been evangelical about its laissez-faire approach to user-generated content, implemented rules that ban “illegal” content, “involuntary pornography,” material that “[e]ncourages or incites violence,” and content that “[t]hreatens, harasses, or bullies or encourages others to do so.” Many “redditors” rebelled, voting up user comments that addressed Reddit’s Asian American female CEO in racist and misogynist ways. These posts were popular enough among redditors to make it to the site’s front page, the prime position on the site that touts itself as “the first page of the Internet.” Reddit subsequently buttressed its restrictions on violent and harassing content.
Moreover, it recently banned a “subreddit” of self-identified misogynists.\textsuperscript{43} Reddit’s reforms have been met with fierce resistance from self-styled free speech enthusiasts.\textsuperscript{44} But the company does not appear to be backpedaling at this time.

As this example indicates, and as new scholarship illuminates,\textsuperscript{45} attention to consumer demand and a sense of corporate responsibility have motivated certain intermediaries to moderate certain user content. It may be tempting to conclude that reforms to Section 230 law are therefore unnecessary. Unregulated intermediaries might be the best gauges of authentic user sentiment about what is or is not objectionable. Section 230 doctrine, on this view, allows users to express and learn from each other in a dynamic fashion, without the distortions that may be caused by tort liability or government mandates. This is part of why free speech enthusiasts ascribe so much significance to the statute: Section 230 doctrine for them is premised on a noble faith in the moral and democratic power of unregulated information markets.\textsuperscript{46}

II. The Lived Human Costs of “Unfettered” Online Speech: The Example of Nonconsensual Pornography

These arguments for near-blanket immunity only go so far, though. As much as some intermediaries may try, the fact is that many others do not make any effort to block or remove harmful expressive conduct. According to their critics, sites like Backpage (a classified site through which users are known to engage in the sex trafficking of minors) or TheDirty (a gossip site known for soliciting derogatory content about unsuspecting young women) are unabashed solicitors and distributors of a species of content that attacks members of historically subordinated groups. Under current doctrine, they are immune for acting in this way. They are just as immune under Section 230 as are ostensibly content-consience intermediaries like Facebook and Twitter that purport to remove or block various categories of illicit user content but nevertheless sometimes distribute it.\textsuperscript{47} The prevailing justification for this approach is to protect against the “collateral censorship” of lawful content.\textsuperscript{48} This view holds that slippage in the direction of occasionally hosting hurtful material is the price of ensuring free speech online.

It may be correct that tolerating harmful content every now and again is the cost of promoting the statutory objective of an “unfettered” online speech environment. But just as a wide range of offline expressive acts like fraud, sexual harassment, and racially discriminatory advertisements for housing are not entitled to legal protection, we might wonder whether online services should be entirely immune for similar behaviors by their users.\textsuperscript{49} To be sure, there is a significant qualitative and quantitative difference between the reach of offline and online expressive acts: The latter travel further and faster than the former by a long shot. But this fact hardly removes the need to regulate harmful online behaviors. Quite the contrary. The human costs of “unfettered” online speech may be aggravated by the internet’s reach, and the costs themselves are disproportionately shouldered by those who are most likely to be the targets of attacks and abuse both online and off. That is to say, the victims of online abuse tend to be the same sorts of people who have always been subject to attack and harassment offline in the United States and elsewhere — in particular, young women, racial minorities, and sexual “deviants.”\textsuperscript{50}

The harm that these users experience is made worse by the way in which illicit or inflammatory content, once distributed, can spread across the internet at a speed and scale that is hard, if not impossible, to control. This unforgiving ecology raises the stakes of occasional slippage for the predictable targets and systemic victims of harmful content. The internet thus reinforces some of the classic arguments for the regulation of assaultive speech acts that target members of historically subordinated groups.\textsuperscript{51} The vitriolic content that flows through online intermediaries affects members of these groups distinctively, discouraging them from participating fully in public life online and making their social and commercial integration even more difficult than it might otherwise be.\textsuperscript{52}

Consider nonconsensual pornography, the distribution of nude images of a person who never authorized their distribution. On the internet, such images are generally shared in order to humiliate or harass the depicted person. In some instances, third parties then exploit the images to extort the victim, as in the case of sites that require a fee to take the images down.\textsuperscript{53} Other parties discover and distribute such images for free, without necessarily knowing anything about the depicted individual.
The injuries caused by nonconsensual pornography are clear and are felt most immediately and painfully by its victims. Section 230 jurisprudence is riddled with cases that illustrate these harms. In one of the more cited ones, *Barnes v. Yahoo!, Inc.*, a young woman sued Yahoo! for failing to remove a false dating site profile of her created by her ex-boyfriend. The profile contained her work phone number and address, as well as nude and suggestive photographs accompanied by promises of sex. Would-be suitors and predators soon came looking for her at work. The harm caused by this cruel hoax was plain.

Victims of nonconsensual pornography may experience many other indignities. Once posted, the offending image takes on a life of its own, exacting something that resembles an endlessly repeating privacy invasion. Danielle Citron and Mary Anne Franks, who have been thinking and writing compellingly about the issue for almost a decade now, explain the phenomenon:

Today, intimate photos are increasingly being distributed online, potentially reaching thousands, even millions of people, with a click of a mouse. A person’s nude photo can be uploaded to a website where thousands of people can view and repost it. In short order, the image can appear prominently in a search of the victim’s name. It can be e-mailed or otherwise exhibited to the victim’s family, employers, coworkers, and friends. The Internet provides a staggering means of amplification, extending the reach of content in unimaginable ways.

The scale of distribution magnifies the harm to depicted individuals far beyond what is possible through other communications technologies. In this environment, taking down nonconsensual pornography, once it has been posted on an online intermediary, often becomes a futile and agonizing game of whack-a-mole.

In addition to the direct harms to those whose images are being exploited, the distribution of nonconsensual pornography also exacts a more general harm that mirrors and reinforces the routine subjugation of young women. It is different in this regard from defamatory user posts, the prototypical subject of Section 230 jurisprudence, in which the injury caused by the defamatory posts are reputational in nature. Nonconsensual pornography sweeps its victims into a network of blogs, pornography sites, social media groups, Tumblrs, and Reddit discussion threads that enthusiastically traffic in the collective humiliation of young women.

And yet, Section 230 doctrine relieves online intermediaries of any legal obligation to block or remove nonconsensual pornography. When sued for distributing such images and videos, the intermediaries cite Section 230 to justify their passive role. Courts have generally sided with them, explaining that the immunity is not contingent on sites’ policing of illicit user content. The result is not only grief for the predictable victims of online abuse and harassment but also a regulatory regime that helps to reinforce systemic subordination.

### III. More than a Conduit: Online Intermediaries’ Designs on User Data

As pernicious as it is, cyberharassment does not reflect the full scope of the threat that such broad legal protection for online intermediaries poses to vulnerable persons. This is because, today, most if not all intermediaries affirmatively shape the form and substance of user content. Adding to the arguments that scholars like Citron and Franks have ably made, I want to call attention here to this crucial way in which Section 230 immunity entrenches extant barriers to social and commercial integration for historically subordinated groups. I want to suggest, furthermore, that over two decades into the development of the networked information economy, online intermediaries should not be able to claim blissful indifference when their designs predictably elicit or even encourage expressive conduct that perpetuates discrimination and subjugation.

I make these arguments in this part in several sections. In section A, I illustrate the ways in which intermediaries pervasively influence users’ online experiences. In section B, I explain how such designs can enable and exacerbate certain categories of harmful expressive acts. Section C looks at the courts’ responses.

### A. Intermediary Designs and User Experiences
Popular services like Facebook, Twitter, and Airbnb offer good examples of how intermediary designs interact with user experiences. Twitter immediately distributes its users’ posts (tweets) after the users type them. But its user interface affects the nature and content of those tweets. Twitter’s 280-character limitation, for example, has generated its own abbreviated syntax and conventions (for example, hashtags and retweets). The company also allows pseudonyms, effectively allowing users to be anonymous. This liberal approach to attribution invites creativity and useful provocation but also the harassment and targeted attacks mentioned above. Twitter knows this, and in many cases it will take down such attacks after the fact and remove users who routinely violate the company’s no-harassment policy.

These superficial interface design features are distinct from the designs on content that occur behind (so to speak) the user interface. Some companies are intentionally deceptive about how they acquire or employ content. Take, for example, the online marketing company that placed deceptive information about its clients’ products on affiliated “fake news” sites. Or consider the online sleuthing company that, in response to solicited user requests for information about people, routinely contracted with third-party researchers to retrieve information in ways it allegedly knew violated privacy law.

Without necessarily resorting to outright deception, many more intermediaries administer their platforms in obscure or undisclosed ways that are meant to influence how users behave on the site. Many intermediaries, for example, employ user interfaces designed to hold user attention by inducing something like addictive reliance. Facebook employs techniques to ensure that each user sees stories and updates in her “News Feeds” that she may not have seen on her last visit to site. And its engineers constantly tweak the algorithms that manage the user experience. In addition, many intermediaries analyze, sort, and repurpose the user content they elicit. Facebook and Twitter, for example, employ software to make meaning out of their users’ “reactions,” search terms, and browsing activity in order to curate the content of each user’s individual feed, personalized advertisements, and recommendations about “who to follow.” (A Wired magazine headline of three years ago comes to mind: “How Facebook Knows You Better than Your Friends Do.”) Intermediaries ostensibly do all of these things to improve user experiences, but their practices are often problematic and opaque to the outside world. As very recent revelations involving Cambridge Analytica underscore, Facebook for years shared its unrivaled trove of user data with third-party researchers, application developers, and data brokers in the interest of deepening user engagement. Facebook reportedly took 30 percent of developer profits in the process.

This is all to say that intermediaries now have near-total control of users’ online experience. They design and predict nearly everything that happens on their site, from the moment a user signs in to the moment she logs out. The lure of “big” consumer data pushes them to be ever more aggressive in their efforts to attract new users, retain existing users, and generate information about users that they can mine and market to others. It is neither surprising nor troubling that companies make handsome profits in this way. But these developments undermine any notion that online intermediaries deserve immunity because they are mere conduits for, or passive publishers of, their users’ expression. Online intermediaries pervasively shape, study, and exploit communicative acts on their services.

All of this, moreover, belies the old faith that such services operate at too massive a scale to be asked to police user content. Online intermediaries are already carefully curating and commoditizing this content through automated “black box” processes that would seem unworkable were they not working so well. The standard justifications for broad immunity under Section 230 — grounded in fears of imposing excessive burdens on intermediaries and chilling their distribution of lawful material—have become increasingly divorced from technological and economic realities. As intermediaries have figured out how to manage and distribute user data with ever greater precision, the traditional case for Section 230 immunity has become ever less compelling, if not altogether inapt.

B. Discriminatory Designs on User Content and Data: The Example of Online Housing Marketplaces
These developments in intermediary design have been underway for over a decade now and have become far-reaching and consequential enough in themselves to warrant rethinking of Section 230 doctrine. The problems with the doctrine, however, are made worse when intermediaries’ designs facilitate expressive conduct that harms vulnerable people and members of historically subordinated groups.\textsuperscript{71} We often hear about the dangerous content that intermediaries automatically distribute by algorithm, as in the notorious ways in which Facebook and Twitter facilitated the targeted dissemination of “fake news” in the months leading up to the 2016 presidential election,\textsuperscript{72} or the advertisement that Instagram made of a user’s personal photo of a violently misogynist threat she had received through her account.\textsuperscript{73} My point here, however, is that the stakes of automated intermediary designs are especially high for certain predictable communities. Unpoliced, putatively neutral online application and service designs can entrench longstanding racial and gender disparities.

Consider Airbnb’s popular home-sharing service. Quite unlike Twitter’s liberal approach to personal attribution, Airbnb’s main service requires each guest to create an online profile with certain information, including a genuine name and phone number. It also encourages inclusion of a real photograph.\textsuperscript{74} For Airbnb, the authenticity of this profile information is vital to the operation of the service, as it engenders a sense of trust and connection between hosts and guests. Guests’ physical characteristics may contain social cues that instill either familiarity and comfort, on the one hand, or suspicion and distrust, on the other. The sense of authentic connection that Airbnb is adamant about cultivating, however, has dangerous consequences in a market long plagued by discrimination against racial and ethnic minorities. In its more insidious manifestations, access to a guest’s name and profile picture affords hosts the ability to assess the trustworthiness of a guest based on illicit biases — against, say, Latinos or blacks — that do not accurately predict a prospective guest’s reliability as a tenant. In this way, Airbnb’s service directly reinforces discrimination when it requires users to share information that suggests their own race.

That race would matter so much to Airbnb hosts should not be a surprise. Race, after all, has long played an enormous—and pernicious—role in U.S. housing markets, online as well as offline. SketchFactor, the crowdsourced neighborhood safety rating application, for example, became little more than a platform for users to share racist stereotypes about “shady” parts of town.\textsuperscript{75} Match.com, the ostensibly race-neutral online dating application, facilitates users’ discrimination against blacks.\textsuperscript{76} Similarly, Airbnb hosts use the home-sharing service to discriminate against racial minorities whose identities as such are suggested in their profiles. Guests have complained publicly about this phenomenon, giving rise to the hashtag #AirbnbWhileBlack.\textsuperscript{77} One guest reported that a host abruptly cancelled her reservation after sending an unambiguously bigoted explanation: “I wouldn’t rent to u if u were the last person on earth. One word says it all. Asian.”\textsuperscript{78} Researchers at the Harvard Business School have substantiated individual claims like these, finding that Airbnb guests “with distinctively African-American names are 16 percent less likely to be accepted relative to identical guests with distinctively White names.”\textsuperscript{79} Airbnb felt compelled to commission a well-regarded civil rights attorney to conduct a study on the topic. Her review, too, found a distinct pattern of host discrimination against users whose profiles suggest they are a member of a racial minority group.\textsuperscript{80}

The difference between these racially discriminatory patterns as they appear on Airbnb versus dating or neighborhood rating apps is that the former are illegal because they violate fair housing laws. The 1968 Fair Housing Act (FHA), for example, specifically forbids home sellers or renters, as well as brokers, property managers, and agents, from distributing advertisements “that indicate[] any preference, limitation, or discrimination based on race, color, religion, sex, handicap, familial status, or national origin.”\textsuperscript{81} States have similar laws. In light of the mounting evidence that hosts use its service to discriminate unlawfully, Airbnb has augmented its efforts to police discriminatory behavior by hosts. In addition to requiring users to forswear that practice, the company now also requires new users to agree “to treat everyone in the Airbnb community—regardless of their race, religion, national origin, ethnicity, skin color, disability, sex, gender identity, sexual orientation or age—with respect, and without judgment or bias.”\textsuperscript{82} Airbnb has also promoted its “instant bookings” service as an alternative to its main service.\textsuperscript{83} “Instant bookings” does not require elaborate profiles (including racially suggestive names or pictures) to complete transactions.
However, Airbnb still facilitates discrimination through its main service to the extent that it continues to rely on names and pictures. The “instant bookings” feature, paired with the main service, creates a “two-tiered reservations system”: In one system (instant bookings), guests lose a sense of conviviality with hosts but obtain some peace of mind in knowing that they will not be discriminated against on the basis of race, while in the other system (the main service), discrimination is inevitable but also exploited to promote “authentic” connections.\textsuperscript{94}

Section 230 doctrine arguably insulates Airbnb’s design choices from antidiscrimination law’s scrutiny. The company and its defenders have routinely cited Section 230 as a protection against liability for a wide range of illicit host activities, including discrimination that violates fair housing laws.\textsuperscript{85} In their view, the statutory immunity is robust enough to protect Airbnb from liability for these expressive acts by third-party hosts because the company only facilitates transactions between users.\textsuperscript{86} It does not contribute anything material to the transactions themselves.\textsuperscript{87}

Airbnb is far from alone in deploying designs that routinely generate serious forms of discrimination. Late in 2016, ProPublica published the first in a series of illuminating reports on Facebook Ads, the social media company’s powerful microtargeted advertising platform. This service enables advertisers to customize campaigns to social media users based on the information that Facebook gathers about those users. Facebook Ads is a bargain (at a clip of $30 for each advertisement) compared to the going rate of top social media marketing and advertising firms. It can be a great help to entrepreneurs of all sizes because it identifies salient market segments in real time.

Facebook Ads is also distinctive because the company employs software, first, to analyze the unrivaled troves of user data that it collects and, second, to create dozens of categories from which advertisers may choose. These include targeted classifications within geographic locations, demographics, friendship networks, and online user behaviors.\textsuperscript{88} Among the more notorious categories in the recent past were ones that “enabled advertisers to direct their pitches to the news feeds of almost 2,300 people who expressed interest in the topics of ‘Jew hater,’ ‘How to burn jews,’ or ‘History of ‘why jews ruin the world.’”\textsuperscript{89} No human at Facebook created these specific anti-Semitic classifications. Facebook’s algorithms determined that they were salient based on user interest at the time.\textsuperscript{90}

Facebook’s algorithms likewise seem to have created various controversial demographic classifications for “ethnic” or “multicultural” affinities, a category that does not connote race as much as users’ cultural associations and inclinations.\textsuperscript{91} These classifications are predictive proxies, however, for race and ethnicity. Recent news reports have shown that, through these classifications, Facebook Ads has enabled building managers and employers to exclude racial minorities from advertisements about apartment rentals and to exclude older people from advertisements about jobs.\textsuperscript{92} When faced with stories of discrimination on the advertising platform in late 2016, Facebook immediately announced a plan to stamp out the practice.\textsuperscript{93} Among other things, Facebook now requires advertisers to certify that they do not discriminate in contravention of civil rights laws.\textsuperscript{94} But, as with Airbnb, reports of illicit use of the site continue to surface.\textsuperscript{95}

Critics and victims of these practices would greatly prefer to seek relief and reform from the intermediary itself—from Facebook—rather than from thousands of individual users. Aggrieved parties have thus filed federal class action lawsuits against Facebook alleging fair housing and employment discrimination violations.\textsuperscript{96} Predictably, Facebook has cited Section 230 to defend its advertising platform. It argues that the company does not control the reach or content of targeted ads; third-party advertisers do. According to Facebook, its platform is nothing more than a “neutral tool” to help these advertisers “target their ads to groups of users most likely to be interested in the goods or services being offered.”\textsuperscript{97} This activity, it asserts, falls squarely in the category of “publishing” for which companies like Facebook are granted immunity under the CDA.

\textbf{C. Doctrinal Responses — and Resources}

Section 230 doctrine could very well lead courts to side with Facebook on this matter. But it is hardly obvious that it should, given that the alleged discrimination would not be possible but for the way in which Facebook leverages its unrivaled access to social media user data to generate the illicit
categories. In Facebook’s favor, courts have read Section 230 to immunize intermediaries that host racially discriminatory advertisements or solicitations. In 2008, the U.S. Court of Appeals for the Seventh Circuit explained that the popular classifieds site Craigslist could not be held liable for hosting third-party housing advertisements that overtly expressed preferences for people on the basis of race, sex, religion, sexual orientation, and family status. The panel explained that Congress enacted the statute to protect services exactly like Craigslist. The company neither had a hand in the authorship of the discriminatory advertisements nor caused or induced advertisers to post such content. Craigslist, the panel reasoned, acts as nothing more than a publisher of (sometimes racist) user content and, as such, could not be liable under federal fair housing law. Had Congress meant to include an exception under Section 230 for such laws, it would have said so.

But the Section 230 case law also contains some resources and opportunities for plaintiffs like those in the current Facebook Ads case. In the same year that the Seventh Circuit ruled in favor of Craigslist, the Ninth Circuit sitting en banc held that an important design element of Roommates.com, a website that also brokers connections between people in the housing market, was not immune under Section 230. As a condition of participation on the site, Roommates.com required subscribers to express preferences that are strictly forbidden under fair housing law. Among other things, the site’s developers designed a dropdown menu that listed gender, sexual orientation, and family status as potential options. (Notably, the menu did not include race among the listed items.) A participant had to share such a preference to find a match. The Ninth Circuit held that this design feature “materially contributed” to a fair housing law violation every time a user expressed a preference for one of those prohibited classifications. This conclusion flowed from language in Section 230 that does not extend protection to intermediaries that help to “create or develop” illicit third-party content.

As important as the Roommates.com opinion has become in limiting the scope of immunity under Section 230, it is worth noting that the Ninth Circuit was very careful in how it discussed its holding. The court made a point of limiting its no-immunity conclusion to the dropdown menu. The plaintiffs had argued that a separate, blank dialogue box that Roommates.com makes available to subscribers also permits them to express bigoted preferences and share information in violation of fair housing law. For example, subscribers had posted comments that they “prefer white Male roommates,” that “the person applying for the room MUST be a BLACK GAY MALE,” or that they are “NOT looking for black muslims.” The court held that Section 230 immunizes Roommates.com from liability for statements like these. It is not enough, the court reasoned, that the site encourages subscribers to share preferences and information, as this is “precisely the kind of situation for which section 230 was designed to provide immunity.” Roommates.com only “passively displayed” the statements and had “no way to distinguish unlawful discriminatory preferences from perfectly legitimate statements.” This conclusion jibes with the Seventh Circuit’s approach to Craigslist. Indeed, these two opinions neatly mapped out the basic contours of Section 230 doctrine when they were decided in 2008. The Roommates.com opinion, in particular, is now routinely cited as authority for the “material contribution” standard.

The Ninth Circuit’s other notable conclusion in that case, decided a couple of years after a post-remand trial court finding for Roommates.com, was that the plaintiff civil rights organization, the Fair Housing Council of the San Fernando Valley (FHC), had standing to seek relief even if it was not itself the victim of a discrete discriminatory act. FHC had alleged that Roommates.com was strictly liable for designing its site in a way that discriminated against prospective renters. It claimed standing to sue, however, because its research into the company’s discriminatory designs was a drain on its resources and frustrated its mission. The Ninth Circuit agreed, holding that FHC had suffered an actual injury sufficient to have standing.

In essence, the court determined that the organization could stand in for a hypothetical Roommates.com subscriber who would be harmed by users’ discriminatory preferences and postings. This holding makes good sense, as discriminatory targeted advertisements and solicitations subjugate racial minorities even when their victims do not witness or otherwise experience the discriminatory act directly. Civil rights laws often reach beyond discrete acts of exclusion in order to redress systemic patterns of
subordination and exclusion. Roommates.com’s design choices, FHC had argued, facilitated communicative acts of discrimination in a market long plagued by that very problem. And if not for FHC’s intervention, the court reasoned, these patterns of bias would continue.

IV. Toward a More Nuanced Immunity Doctrine

The Roommates.com opinion, issued a decade ago, helps to show the way forward. The Ninth Circuit’s careful treatment of the two contested features of the website design of Roommates.com demonstrated an appreciation for the diversity of ways in which the company elicits content from users, and its standing ruling demonstrated an appreciation for the realities of civil rights harms.

However, the Ninth Circuit’s opinion did not go far enough; it did not address the increasingly subtle and tentacular kinds of control that online intermediaries exert over users’ experiences today. The system through which Facebook, for example, algorithmically sorts and repurposes user data to support microtargeted advertising is a far cry from the clumsy dropdown menu in the Roommates.com case. Two decades after the CDA’s enactment, it has become increasingly implausible to equate this powerful manipulation of users’ data and content with traditional publishing under Section 230.

Section 230 doctrine must be adapted to the political economy of contemporary online information flows. Judges and litigants already have a rich set of tools from antidiscrimination and consumer protection law for determining liability and providing remedies for harmful expressive conduct. But the current Section 230 doctrine cuts cyberspace off from these other bodies of law, foreclosing liability analysis for companies whose service designs routinely facilitate or even encourage illicit content.

It is important to emphasize, moreover, that holding intermediaries to account for such designs does not require anything like strict liability for the harms caused by nonconsensual pornography or any other user-generated content. Consistent with the neglected Good Samaritan goal of the statute, Section 230 can quite comfortably be interpreted to provide a safe harbor for intermediaries that try in good faith to block or take such content down. That is, after all, precisely what the text of Section 230(c)(2)(A) says, at least with regard to “objectionable” speech. At the same time, courts could allow plaintiffs to seek redress from intermediaries that knowingly or negligently facilitate the distribution of harmful content. As the Ninth Circuit’s ruling against Roommates.com shows, we do not need new statutory language to assess intermediary liability when the user interface at issue enables illegal online conduct.

But the experience of two decades of Section 230 litigation does suggest that new statutory language could help, particularly since the prevailing view prevents the plain meaning of the Good Samaritan title and Section 230(c)(2)(A) from doing any meaningful work. The statute itself, moreover, fails to give clear direction on the kinds of torts it covers. Nor, for that matter, does the statute address the extent to which a defendant must “create[] or develop[]” the offending material. This has been left to the courts to sort out. Distressed by the wide scope of the doctrine and some of these textual gaps, legislators and activists have been promoting amendments to Section 230 that would create exceptions for prostitution, nonconsensual pornography, and the sex trafficking of minors. There is no reason why Congress couldn’t also write in an explicit exception to Section 230 immunity for violations of civil rights laws.

Such proposals will face substantial pushback from intermediaries and others. A company like Facebook, for example, has a lot to lose from any change that would require it to be more careful about how it distributes user content or generates personal or targeted advertisements. Even a shift to what some are now calling “contextual advertising,” where an advertiser buys the context in which social media users engage with each other rather than individual users’ profiles, could cost a company like Facebook billions of dollars. And to be sure, apart from the commercial interests at stake, there are important free speech arguments for keeping Section 230 broad: The content and data flowing through the online speech environment may not be as abundant in a world in which intermediaries are held to account for their users’ content and their own designs on user data. But then again, it is difficult to weigh this “chilling” concern against the chilling of members of historically subordinated groups that is already happening under existing law.
Whether legal reform in this area takes place in the legislature or the judiciary or both, reform is necessary. Judges, lawyers, and legislators should stop shielding intermediaries from liability on the basis of implausible assumptions about their neutrality or passivity — and should instead start looking carefully at how intermediaries' designs on user content do or do not result in actionable injuries. This attention to design will further sensitize intermediaries to the ways in which their services perpetuate systemic harms. Equipped with a more nuanced approach to intermediary immunity, we might come to expect an online environment that is hospitable to all comers.

Olivier Sylvain is an Associate Professor of Law at Fordham University School of Law. He teaches legislation and regulation, administrative law, information law, and law related to communications technologies. He is also the Director of the McGannon Center for Communications Research. Olivier is also part of a team of research engineers and social entrepreneurs to whom the National Science Foundation in fall 2017 awarded a three-year one-million-dollar grant to prototype a secure, affordable, and energy-efficient computing network that is to be owned and operated as a "commons resource" for Harlem residents. Sylvain was a Karpatin Fellow in the National Legal Office of the American Civil Liberties Union in New York City and a litigation associate at Jenner & Block, LLC, in Washington, D.C. He is on the board of directors for the New York affiliate of the American Civil Liberties Union.


3 See Orly Lobel, The Law of the Platform, 101 Minn. L. Rev. 87, 89 (2016) (discussing "the digital platform revolution").


5 Id. § 230(b)(2).


8 See, e.g., Jones v. Dirty World Entm’t Recordings, 755 F.3d 398, 413 (6th Cir. 2014); Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1167–71 (9th Cir. 2008) (en banc).

9 Roommates.com, 521 F.3d at 1167–68.

10 See, e.g., id. at 1180 (McKeown, J., concurring in part and dissenting in part) (“We have underscored that this broad grant of webhost immunity gives effect to Congress’s stated goals ‘to promote the continued development of the Internet and other interactive computer services’ and ‘to preserve the
vibrant and competitive free market that presently exists for the Internet and other interactive computer services."

11 Jones v. Dirty World Entm’t Recordings, LLC, 755 F.3d 398 (6th Cir. 2014).


19 See Wu, supra note 7, at 315–18.

20 See generally Kenneth Bamberger & Orly Lobel, Platform Market Power, 32 Berk. Tech. L.J. 1, 37–39 (2018) (discussing ways in which intermediaries leverage their market position to exploit user data in different markets); Lina M. Khan, Note, Amazon’s Antitrust Paradox, 126 Yale L.J. 710 (2017) (discussing ways in which intermediaries may raise antitrust concerns to the extent they cultivate their position as “essential infrastructure” for commerce across industries).


22 On these norms, see generally Olivier Sylvain, Network Equality, 67 Hastings L.J. 443 (2016).

23 The pertinent language provides as follows:

(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).
47 U.S.C. § 230(c) (2012). The statute excludes from immunity intermediaries that are “responsible, in whole or in part, for the creation or development” of illicit user content. Id. § 230(f)(3). Applying this language, courts have subjected defendant intermediaries to liability when they “materially contribute” to the offending content. See, e.g., *Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157, 1167–71 (9th Cir. 2008) (en banc). In principle, the “material contribution” standard limits the scope of the protection to services that are only conduits of content. In practice, however, it is a very high bar for plaintiffs to clear.


27 *Luke* 10:23–37 (“[A] Samaritan, as he traveled, came where the man was; and when he saw him, he took pity on him. He went to him and bandaged his wounds, pouring on oil and wine. Then he put the man on his own donkey, brought him to an inn and took care of him.”).

28 *Zaran*, 129 F.3d at 330–32.


32 See, e.g., *Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157 (9th Cir. 2008) (en banc); *Zaran*, 129 F.3d 327.

33 See, e.g., *Zaran*, 129 F.3d at 333.


44 Id.


48 See sources cited supra note 7.


50 See Citron, supra note 35, at 13–16.


52 Cf. Richard Delgado & Jean Stefancic, *Understanding Words that Wound* 217–18 (2004) (advocating a “new approach” that “points out how speech and equality stand in reciprocal relation; neither can thrive without the other. Speech without equality is a lecture, a sermon, a rant. Speech, in other words, presumes equality, or something like it, among participants in a dialogue”).


54 570 F.3d 1096 (9th Cir. 2009).


58 See Citron, supra note 35, at 127 (“Cyber harassment reinforces gender stereotypes by casting women as sex objects that are unﬁt for life’s important opportunities.”).

59 See, e.g., Barnes v. Yahoo!, Inc., 570 F.3d 1096 (9th Cir. 2009); Jones v. Dirty World Entm’t Recordings, LLC, 755 F.3d 398 (6th Cir. 2014).

60 Twitter recognizes the signiﬁcance of its character limitation; it increased the limitation from 140 in November 2017 to improve the user experience. See Aatif Sulleyman, Twitter Introduces 280 Characters to All Users, Independent (Nov. 7, 2017), http://www.independent.co.uk/life-style/gadgets-and-tech/news/twitter-280-characters-tweets-start-when-get-latest-a8042716.html.

61 See FTC v. LeadClick Media, 838 F.3d 158 (2d Cir. 2016).

62 See FTC v. Accusearch, 570 F.3d 1187 (10th Cir. 2009).


70 Lewis, supra note 69.

71 Cf. Solon Barocas & Andrew D. Selbst, Big Data’s Disparate Impact, 104 Calif. L. Rev. 671 (2016) (discussing ways in which algorithmic analysis and machine learning may produce discriminatory impacts); Levy & Barocas, supra note 45 (discussing ways in which intermediary designs may have discriminatory impacts).


Airbnb also gives users the option of importing information from users’ Facebook accounts.


See Emanuella Grinberg, When It Comes to Dating Sites, Race Matters, CNN (Jan. 13, 2016), http://www.cnn.com/2016/01/13/living/where-white-people-meet-feat/index.html. This is to say nothing of sites like WhereWhitePeopleMeet that openly exploit this phenomenon. Id.


Javier, supra note 78.


87 On the other hand, Airbnb’s decision to settle in some of these cases may suggest that the company worries about its role in perpetuating discrimination, irrespective of whether Section 230 supplies immunity. Cf. Sam Levin, *Airbnb Gives in to Regulator’s Demand to Test for Racial Discrimination by Hosts*, Guardian (Apr. 27, 2017), http://www.theguardian.com/technology/2017/apr/27/airbnb-government-housing-test-black-discrimination.


91 ProPublica, which first broke the story about this practice, see Angwin & Parris, *supra* note 14, has not reported on whether Facebook generates these categories manually or by algorithm. I do not take up the question here, but the roles of automation and machine learning raise difficult questions about proof of intention under current nondiscrimination law.


95 Id.


97 Defendant’s Notice of Motion and Motion to Dismiss First Amendment Complaint; Memorandum of Points and Authorities in Support Thereof at 10, *Mobley v. Facebook, Inc.* (N.D. Cal.) (June 1, 2017) (No. 5:16-cv-06440-EJD), available at http://assets.documentcloud.org/documents/4333515/Outten-FB-FB-Motion-to-Dismiss-4-3-17.pdf. Facebook also asserts that the plaintiffs lack standing and that, in any event, it is not discriminating within the meaning of the pertinent civil rights laws. *Id.* at 14–25.

98 *Chicago Lawyers’ Comm. for Civil Rights v. Craigslist, Inc.*, 519 F.3d 666 (7th Cir. 2008).

99 *Id.* at 671.

100 *Id.*

101 *Id.*

102 *Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157 (9th Cir. 2008) (en banc).

103 *Id.* at 1165–72.
47 U.S.C. § 230(f)(3) (2012). After remand, a three-judge panel did nothing to alter this conclusion in its ruling four years later. *Fair Hous. Council of San Fernando Valley v. Roommate.com, LLC*, 666 F.3d 1216 (9th Cir. 2012). In this later opinion, the panel held that, while the immunity under Section 230 did not bar the suit against Roommates.com for its drop-down menu, Roommates.com’s specific conduct at issue did not violate the FHA because “the FHA doesn’t apply to the sharing of living units” as opposed to “the sale or rental of a dwelling.” *Id.* at 1222 (discussing the scope of 42 U.S.C. § 3604(c)).

*Roommates.com*, 521 F.3d at 1173.

*Id.* at 1174.

*Id.*

See *id.* at 1173 n.33 (explaining that the court’s holding is consistent with the Seventh Circuit’s Craigslist opinion).

See, e.g., *Jones v. Dirty World Entm’t Recordings*, 755 F.3d 398, 410–12 (6th Cir. 2014); *FTC v. Accusearch*, 570 F.3d 1187, 1200 (10th Cir. 2009).

*Fair Hous. Council of San Fernando Valley v. Roommate.com, LLC*, 666 F.3d 1216, 1219 (9th Cir. 2012).

*Id.*

*Id.*


See 47 U.S.C. § 230(c)(2)(A) (2012) (“No provider or user of an interactive computer service shall be held liable on account of . . . 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.”).

*Id.* § 230(f)(3). See generally *Sylvain, Design Duties, supra* note 21, at 239–42.


*Id.*

It is also difficult to disentangle this free speech argument from the intermediaries’ commercial interests. European regulators, for instance, fined Google almost two and a half billion Euros last summer for abusing its market dominance in search to give “an illegal advantage to another Google product.” European Commission, Press Release, *Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service* (June 27, 2017), http://europa.eu/rapid/press-release_IP-17-1784_en.htm.
The Emerging Threats (https://knightcolumbia.org/emergingthreats) series invites leading thinkers to identify and grapple with newly arising or intensifying structural threats to the system of free expression.

Olivier Sylvain
Discriminatory Designs on User Data
(Download as a PDF (/sites/default/files/content/Sylvain_Emerging_Threats.pdf))

Introduction: Intermediary Immunity and Discriminatory Designs (https://knightcolumbia.org/content/intermediary-immunity-and-discriminatory-designs) | David Pozen

Responses
Section 230’s Challenge to Civil Rights and Civil Liberties (https://knightcolumbia.org/content/section-230s-challenge-civil-rights-and-civil-liberties) | Danielle Keats Citron

To Err Is Platform (https://knightcolumbia.org/content/err-platform) | James Grimmelmann

Toward a Clearer Conversation About Platform Liability (https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability) | Daphne Keller
In June 2016, Facebook Vice President Andrew Bosworth circulated an internal memorandum describing the company’s mission. “We connect people,” he wrote. “Maybe someone finds love. Maybe [Facebook] even saves the life of someone on the brink of suicide. . . . Maybe it costs a life by exposing someone to bullies. Maybe someone dies in a terrorist attack coordinated on our tools.”

Bosworth was not speaking metaphorically. The month after he distributed the memo, Facebook was sued by the parents of five victims of terrorist attacks conducted by Hamas, which had made extensive use of the company’s platform to recruit and organize. Days before, the social media platform had inadvertently broadcast a murder over the streaming app Facebook Live.

Yet the outrage Facebook weathered in the aftermath of this violence was nothing compared to the scathing criticism it endured two years later after Buzzfeed News published Bosworth’s memo. When the memo leaked in March 2018, Facebook was neck-deep in controversies over its influence on the 2016 presidential election, including the use of the platform’s ad-targeting function by Russian trolls to sow political discord and the harvesting of user data under false pretenses by the Trump-affiliated political consultancy Cambridge Analytica. Meanwhile, Facebook’s competitors, Twitter and Google, faced mounting criticism over their failure to curb disinformation and abusive content.

In 2016, Bosworth’s memo might have been an awkward attempt at self-reflection. In 2018, it was received with shock as a clear acknowledgment of a Faustian bargain. Now the public could see that platforms understood well the destructive ways their services were being used. Platforms designed their services in a way that facilitated abuse—and even if they did so unintentionally, they could have changed the design once they realized the scope of the problem, but chose not to do so.

Of course, this had never been a secret to begin with. For years, scholars and advocates had been working with platforms to address serious abuses, from nonconsensual pornography and cyber stalking to terrorist activity and impersonations. Even Facebook CEO Mark Zuckerberg’s 2017 manifesto on “Building Global Community” recognized the use of Facebook for harmful ends like “bullying and harassment.”
Nonetheless, the public’s reaction was strong. Ordinary Facebook users and politicians alike felt that the company should have done more to prevent abuse, because it had had fair warning all along and structured its services knowing it was enabling such abuse.6

In the United States, the role of dominant content platforms like Facebook and Twitter in facilitating Russian election interference has precipitated a backlash against “big tech.” These entities are facing unprecedented scrutiny—and not just from the media.

If Congress was previously hesitant to regulate Silicon Valley for fear of stifling innovation and free expression, the pendulum is now swinging in the other direction. Shortly after news broke of the use of Facebook advertisements by Russian provocateurs, senators Mark Warner (D-VA) and Amy Klobuchar (D-MN) introduced the Honest Ads Act, which would require platforms to make “reasonable efforts” to bar foreign nationals from purchasing certain categories of political advertisements during campaign seasons.7

The push for regulation goes far beyond foreign manipulation of platforms’ advertising capabilities. The week before Bosworth’s memo leaked, the Senate passed the Allow States and Victims to Fight Online Sex Trafficking Act (known as FOSTA), writing into law the first-ever statutory exception to Section 230 of the Communications Decency Act, which provides technology companies with immunity from most liability concerning their publication of third-party speech.8 Its supporters credit Section 230 with enabling the growth of online platforms as safe havens for speech.9 In the view of many, Section 230 is a foundational legal protection of arguably ultra-constitutional status for a free and open Internet.10

As we will describe below, FOSTA is deeply flawed legislation. But its successful passage emphasizes the extent to which the political mood has shifted against technology companies. That FOSTA enjoyed overwhelming support from a dysfunctional Congress demonstrates the darkness of the mood.11 Indeed, when Mark Zuckerberg testified before Congress shortly after the bill was passed, Senator John Thune informed the Facebook CEO that he should understand FOSTA as “a wake-up call for the tech community.”12

Silicon Valley seems to recognize its perilous position too. In a striking move, the Internet Association (which represents Facebook, Google, Microsoft, and other big tech companies) ultimately chose to endorse FOSTA, perhaps because the writing was on the wall and perhaps because the climate on the Hill was growing increasingly inhospitable to social media companies given the controversy over their role in Russian election interference.13

The ground has shifted, and FOSTA is likely the first step to a far more regulated Internet. As one of the drafters of Section 230 (now US Senator Ron Wyden, D-OR) recently acknowledged, the law’s safe harbor was meant to incentivize efforts to clean up the
Internet—not to provide a free pass for ignoring or encouraging illegality.\textsuperscript{14} While Section 230 has always had costs—Jonathan Zittrain describes “the collateral damage . . . visited upon real people whose reputations, privacy, and dignity have been hurt in ways that defy redress”—the misuse of these platforms during the 2016 election makes clear the extent and scale of the harm.\textsuperscript{15}

Zuckerberg did not directly address Section 230 in his appearance before Congress. But over and over, he made a startling admission: Facebook, he said, needed to “take a more proactive role and a broader view of our responsibility. . . . It’s not enough to just build tools. We need to make sure that they’re used for good.”\textsuperscript{16}

Likewise, in the weeks before his testimony and in the midst of the Cambridge Analytica crisis, he told CNN, “I actually am not sure we shouldn’t be regulated.”\textsuperscript{17} Whether through law or ethical obligations—Zuckerberg flatly noted Facebook’s “broader responsibility than what the law requires”—it was striking to see the CEO of one of the most powerful technology companies in the world acknowledge some level of culpability for what took place on his platform.

We find ourselves in a very different moment now than we were in five or ten years ago, let alone twenty years ago when Section 230 was passed into law. No matter one’s opinion of the statute, the pressing question now is not whether the safe harbor will be altered, but to what extent. That is astounding—to say the least.

Even if one takes the view that Section 230 is due for an overhaul—as we do—FOSTA is not an ideal approach. Congress should proceed with care, eschewing both overly broad legislation that encourages censorship and piecemeal laws securing partial solutions. FOSTA manages to combine these weaknesses in one piece of legislation. It is over-inclusive and under-inclusive, a legislative failure but the law nonetheless.

We explore alternative possibilities as a way to start—and not to end—a conversation about how to reform the immunity provision. Our paper notes potential pitfalls for future legislation. In our view, efforts to fix the immunity provision should avoid both the problems of a piecemeal approach and the risks of an overly broad one.

To be clear, we do not aim here to solve those problems or offer a definitive legislative response. Our aim is far more modest—to note the problems with FOSTA and urge caution for future legislative endeavors. We hope that our suggestions provide a way to think critically about future proposals to fix Section 230.

Crucial to our paper are developments in corporate boardrooms that take place in the shadow of the law. For the past two years, European Union member states have warned platforms that if they fail to eliminate hate speech, terrorist material, and “fake news”
within twenty-four hours, they will face onerous legislation and penalties. Some member states (including Germany) have followed through on that promise. Platforms are turning to artificial intelligence as a salve, raising significant concerns about commitments to free expression. Machine-learning algorithms increasingly filter objectionable content, such as terrorist activity, before it ever appears. They are flagging hate speech for moderators, but with limited accuracy.

Without doubt, algorithmic responses can be beneficial by quickly heading off problems. But they can also exact significant costs to expression, which must be addressed as well. As companies seek to find their footing in this new political and regulatory environment, corporate speech policies and practices, including algorithmic moderation, should be subject to what one of us (Citron) has called “technological due process”—procedural and substantive commitments designed to protect free expression and safety alike.

**Legislative Solutions**

This part considers potential legislative solutions that would force platforms to reckon with abuses of their services. As Senator Wyden recently explained, “The key to Section 230 was making sure that companies in return for that protection—that they wouldn’t be sued indiscriminately—were being responsible in terms of policing their platforms.” But the senator himself acknowledged that platforms have not lived up to this promise. Under the current sweeping interpretation of Section 230, platforms have no reason to take down illicit material outside of advertising of sex trafficking under FOSTA, copyright violations, ECPA (Electronic Communications Privacy Act) violations, and activities constituting federal crimes. Victims have no leverage to insist that platforms respond to reports of abuse. Platforms can continue to argue that their advertising systems facilitating illegal discrimination are immunized from liability.

Although Section 230 has secured breathing space for the development of online services and countless opportunities to work, speak, and engage with others, it has also given platforms a free pass to ignore destructive activities, to deliberately repost illegal material, and to solicit unlawful activities while ensuring that abusers cannot be identified. As Rebecca Tushnet put it well, Section 230 “allows Internet intermediaries to have their free speech and everyone else’s too.” As the events of the last two years have shown, among those entities that Section 230 enables to speak freely are purveyors of systematic disinformation—state sponsored and others.

In the present political climate, lawmakers are set to change the current immunity enjoyed by online platforms. Current legislative solutions, however, are ill suited to the task. We critique a recently adopted amendment to Section 230 and offer modest suggestions for future reform efforts.
FOSTA: The Narrow Approach

FOSTA (at one point also referred to as the Stop Enabling Sex Traffickers Act or SESTA) went through numerous iterations over the course of months of legislative wrangling between numerous congressional offices and interest groups. While technology companies and Internet activists emphasized what they saw as the bill’s damage to Section 230 and ultimately to online freedom, FOSTA’s proponents played down the legislation’s effect on Section 230 immunity and framed it instead as a narrow crackdown on sex trafficking. The Senate passed FOSTA with a near-unanimous vote; the House did as well.26 The bill has now been signed into law.

FOSTA opens with a “Sense of Congress” introduction meant to reflect the original meaning of Section 230.27 It states that Section 230 was “never intended to provide legal protection to websites . . . that facilitate traffickers in advertising the sale of unlawful sex acts with sex trafficking victims.” That provision continues, “It is the sense of Congress that websites that promote and facilitate prostitution have been reckless in allowing the sale of sex trafficking victims and have done nothing to prevent the trafficking of children and victims of force, fraud, and coercion.”28

FOSTA goes on to provide that technology companies will not enjoy immunity from civil liability if they knowingly assist, support, or facilitate advertising activity that violates federal sex trafficking law, specifically 18 U.S.C. 1591. Currently, advertisers are liable under Section 1591(a)(2) if they knowingly benefit from outlawed ads. FOSTA not only carves out an exception to Section 230 immunity for violations of Section 1591, but it also redefines what constitutes a Section 1591 violation to include “knowingly assisting, supporting, or facilitating” advertising.

Under the new law, state attorneys general, as parens patriae, can seek civil penalties for such activity. Additionally, technology companies could face state criminal charges if the conduct charged would constitute a violation of Section 1591. Companies could also be criminally liable in state court for violations of 18 U.S.C. 2421A, a new section added by FOSTA to the Mann Act, which prohibits transporting a person across state lines “with intent that such individual engage in prostitution” or other criminal sexual activity.29 Section 2421A criminalizes the owning, management, or operation of a platform “with the intent to promote or facilitate” prostitution.

The law’s status as the first legislative incursion against Section 230 led to intense controversy over its drafting and passage. Initially, the arguments proceeded predictably. Anti-sex-trafficking advocates offered their strong support, portraying technology companies like Google as “allies” of human traffickers.30 Leading the opposition, Internet freedom advocates argued that limiting Section 230 immunity would “endanger . . . free expression and innovation” online.31 Sex workers and some advocates for sex trafficking victims and
survivors voiced concerns that clamping down on advertisements for sex online could place women in danger by forcing them to find work on the street and curtailing their ability to vet clients.32

At first, tech companies lined up in lockstep against the bill. But then the seemingly impossible happened: the platforms’ opposition receded. The climate on the Hill had decidedly changed for the worse, and tech companies seemed to recognize this shift. The Internet Association eventually endorsed the legislation after Senate staff changed the bill to head off some of its excesses.33

Amid clear congressional support for FOSTA, major platforms began doing damage control. Two days after the bill passed in the Senate, Craigslist closed its personal ads section, which was often used to post solicitations for sex. Pointing to FOSTA, the advertising hub wrote, “Any tool or service can be misused. We can’t take such risk without jeopardizing all our other services.”34 Likewise, Reddit announced a new site-wide policy prohibiting users from “solicit[ing] or facilitat[ing] any transaction or gift involving certain goods and services, including . . . [p]aid services involving physical sexual contact.” Reddit went on to invoke the typical catchall denial of responsibility for transactions users might undertake, adding, without apparent irony, “Always remember: you are dealing with strangers on the internet.”35

Even before the bill was signed into law, sex workers began tallying lists of websites that had shut down or removed US-based advertisements.36 As of several months later, however, the sex work market has not totally collapsed: other websites remain open, at least for now.37 Some sex workers have turned to Switter, a social network built on the open-source Twitter-style platform Mastodon, in response to what they believed to be censorship of their accounts by other social media platforms following FOSTA’s passage.38 Yet Switter itself briefly vanished when the content delivery network service Cloudflare pulled its services from the social network as a result of FOSTA.39 In an interview with the technology news site The Verge, Cloudflare’s general counsel described FOSTA as a “bad law” but said that “we have an obligation to comply” with it. His comments suggested the company was acting out of an excess of caution in responding to vague legislation: “We’re trying to figure out how law enforcement is going to apply it.” 40

Craigslist and Cloudflare were among the few companies that explicitly cited FOSTA in removing their services. But though neither Reddit nor the shuttered adult-advertising websites pointed to FOSTA, the timing of it all suggests a direct connection. House Judiciary Committee Chairman Bob Goodlatte (R-VA), who pushed FOSTA forward in the House of Representatives, appeared to claim credit for the culling: “Proud that passage of H.R.1865 is having an IMMEDIATE impact on websites advertising prostitution, several shutting down in the past few days,” read a tweet sent from the representative’s official account.41
On the other end of the spectrum, the Electronic Frontier Foundation announced, “Today was a dark day for the Internet.” The VPN service Private Internet Access took out full-page ads in the New York Times pleading with President Trump: “If you sign this bill, every website will be affected . . . Free speech dies.” And, in contrast to his recent comments about online platforms needing to exercise their power responsibly, Wyden countered Goodlatte on Twitter: “Today we take a real step backward, backward, and down a path we will regret.”

Critics argue, and with good reason, that FOSTA creates exactly the “moderator’s dilemma” that Section 230 sought to avoid. By immunizing platforms from liability both for under-filtering under Section 230(c)(1) and for “good faith” over-filtering under Section 230(c)(2), Congress aimed to incentivize self-regulation. Lawmakers wanted to shield companies from liability for incomplete moderation or, as the law’s title describes it, “Good Samaritan Protection for Filtering and Blocking Offensive Content.” (Whether courts have interpreted Section 230 in line with this position, supported by its legislative history and purpose, is another matter.)

FOSTA's detractors argue that FOSTA's unclear “knowingly facilitating” language could perversely push platforms to engage in no moderation at all. Companies might sit on their hands for fear their incomplete removal of ads for sex trafficking might be used as evidence of their “knowing facilitation” of distribution of that content. In this view, the “moderator’s dilemma” would push the platform to simply avoid all moderation so it could disclaim any “knowledge,” especially because it is not clear what would constitute “knowing facilitation” under current law.

The flip side, opponents underscore, is that companies might instead engage in over-the-top moderation to prove their anti-sex-trafficking bona fides and to strengthen their argument that they did not knowingly facilitate such activity in any given case. Overly aggressive moderation would likely involve shutting down hubs devoted to sex advertising or even websites that are known to host such advertising, even if the majority of users turn to the platform for other purposes (as is the case with Craigslist’s personal ads section).

Alternatively, as the next part of this paper explores, such moderation could result in the use of machine-learning algorithms to filter and block anything that relates to sex, including activities that have nothing to do with illegal sex trafficking. Given the bad publicity that could result from a refusal to moderate sex advertisements, aggressive over-removal seems the most likely danger.

The law's critics are onto something. FOSTA is poorly drafted—perhaps because the bill went through so many different revisions and because so many different offices had a hand
in drafting it. The language is confusing, especially the “knowingly assisting, supporting, or facilitating” standard. State and local prosecutors will be unwilling to expend scarce resources on enforcement if they worry about jury confusion and potential legal challenges on vagueness grounds. Even the Justice Department voiced concerns about the bill in a letter to Goodlatte.

But more troubling is that FOSTA endorses a piecemeal approach to a problem that should be solved more comprehensively. Carving out exceptions risks leaving out other destructive activities that should be addressed. In this case, FOSTA deals with sex trafficking but doesn’t touch on the numerous other ills that Section 230 has shielded technology companies from having to grapple with, such as stalking by cyber mobs and individuals, the use of platforms by designated foreign terrorist organizations, violations of federal election law (a topic of new importance given the ongoing investigations into Russian election interference), or other illegal activity.

A piecemeal approach would enable an assessment of whether the approach is working at all (and thus whether it should be extended to other areas), but it precludes experimentation with far better ideas. There is now a long record of how Section 230 has operated with narrow exceptions—and the answer is “not well.” Adding another exception does not address that broader problem.

If FOSTA had been better drafted, it could have been a significant, if modest, step. Of course, taking a modest step means that Congress would need to act again to address other problems as they arose. And given the pace at which Congress now works, realistically speaking, those updates would not be made. In light of the current legislative dysfunction, therefore, a more holistic approach would be beneficial. Yet it remains crucial that legislators proceed cautiously so as to avoid FOSTA’s missteps—or others.

Despite FOSTA’s flaws, congressional reexamination of Section 230 is a sign of changing times. As Zuckerberg’s testimony shows, content intermediaries are beginning to understand themselves as having a moral responsibility for what takes place on their platforms. The question is what form that regulation should take—which is the issue to which we now turn.

**Striking a Better Balance**

In this section, we consider alternative approaches that would avoid the piecemeal approach of FOSTA. In describing these potential responses, our aim is decidedly modest. We do not mean to suggest that these possibilities are the best or the only approach. Rather, we note them here as a starting place for future conversations.

One possibility suggested by noted free speech scholar Geoffrey Stone would be to deny the safe harbor to bad actors. Specifically, that exemption would apply to online service
providers that “knowingly and intentionally leave up unambiguously unlawful content that clearly creates a serious harm to others.”\textsuperscript{50} This would ensure that bad actors, including Backpage (which inspired FOSTA), could not claim immunity if they knowingly and intentionally leave up illegal content causing serious harm. Although this intervention would be narrow, it would not require constant updating to include other bad actors, as does FOSTA.

A variant on this theme would deny the immunity to online service providers that intentionally solicit or induce illegality or unlawful content. This approach takes a page from trademark intermediary liability rules. As Stacey Dogan urges in that context, the key is the normative values behind the approach.\textsuperscript{51} Under this approach, providers that profit from illegality—which surely can be said of sites that solicit illegality—would not enjoy immunity from liability and would risk potential lawsuits if they kept up harmful, illegal content. At the same time, other online service providers would not have a reason to broadly block or filter lawful speech to preserve the immunity. In other words, the approach would provide broad breathing space for protected expression.\textsuperscript{52}

Still another approach would amend Section 230 in a more comprehensive manner. As Benjamin Wittes and one of us (Citron) have argued, platforms should enjoy immunity from liability if they can show that their response to unlawful uses of their services is reasonable.\textsuperscript{53} A better revision to Section 230(c)(1) would read (revised language is italicized):

\begin{quote}
“No provider or user of an interactive computer service that \textit{takes reasonable steps to prevent or address unlawful uses of its services} shall be treated as the publisher or speaker of any information provided by another information content provider \textit{in any action arising out of the publication of content provided by that information content provider}.”
\end{quote}

The immunity would hinge on the reasonableness of providers’ (or users’) content moderation practices as a whole, rather than whether specific content was removed or allowed to remain in any specific instance. The determination of what constitutes a reasonable standard of care would take into account differences among online entities. Internet service providers (ISPs) and social networks with millions of postings a day cannot plausibly respond to complaints of abuse immediately, let alone within a day or two. On the other hand, they may be able to deploy technologies to detect content previously deemed unlawful. The duty of care will evolve as technology improves.

A reasonable standard of care would reduce opportunities for abuses without interfering with the further development of a vibrant Internet or unintentionally turning innocent platforms into involuntary insurers for those injured through their sites. Approaching the problem as one of setting an appropriate standard of care more readily allows for differentiating among various kinds of online actors, setting different rules for large ISPs
linking millions to the Internet versus websites designed to facilitate mob attacks or enable illegal discrimination.54

To use an example from earlier in this post, as a content delivery network provider and security service, Cloudflare would be held to a standard of care that would allow it to provide service to Switter without being potentially held liable for every post on the network, so long as on the main it took reasonable steps to prevent or address unlawful activity happening on Switter. A reasonable duty of care might require Cloudflare to notify Switter about illegality of which it knew; it would not necessarily mean that Cloudflare would be required to withdraw its services from Switter.

We emphasize that these possibilities for amendments to Section 230 are just that—possibilities, rather than fully realized recommendations. In a political environment suddenly tilting toward greater regulation, our point is that modest adjustments to Section 230 could conceivably maintain a robust culture of free speech online without extending the safe harbor to bad actors or, more broadly, to platforms that do not respond to illegality in a reasonable manner.55 The choice, in other words, need not be between the status quo—an option appearing increasingly unlikely—and further carve-outs to the safe harbor along the lines of the deeply flawed model represented by FOSTA.

### Algorithmic Content Moderation

In response to potential liability, many platforms will change their content moderation practices. In the wake of FOSTA, we have already seen the closure of sex advertising hubs, though by no means all such sites. We have also seen a turn toward automation, driven in large part by pressure from European lawmakers. Global content platforms like Facebook and Twitter are increasingly relying on algorithms to filter, block, or obscure troubling material, specifically hate speech and extremist, terrorist content.56

The turn toward automation in content moderation deserves serious study. As Ed Felten has astutely said of hard tech policy questions, lawyers often point to technical solutions and technologists often point to legal solutions.57 As this aphorism—which Paul Ohm has deemed “Felten's Third Law”—suggests, both approaches constitute magical thinking.58 The decision to automate speech decisions should be foregrounded with a careful consideration of the costs and benefits.

This part hopes to spark a dialogue about the perils of technical solutionism. Before vesting technology with the power to automate speech decisions, companies should engage in careful discussion of the costs and benefits to free speech and safety. After describing early congressional approval of automated content moderation, this part highlights some of automation’s successes and its problems. It ends by considering potential safeguards so that platforms can harness automation’s benefits while carefully addressing its potential costs.
Early Legislative Approval

Automated content moderation was contemplated and encouraged by Section 230’s drafters. *Stratton Oakmont v. Prodigy*, the case that prompted the law’s enactment, involved online service provider Prodigy’s use of software to filter profanity. Prodigy’s software, however, failed to catch alleged defamation appearing on a bulletin board, over which the plaintiff sued Prodigy. The court refused to consider Prodigy a distributor, which would have protected it from liability because it had no knowledge of the defamation. The court instead deemed Prodigy a publisher—and thus strictly liable for the alleged defamation—because it was engaged in content moderation.

To the early online service providers, the court’s ruling could not have been clearer. If service providers refrained from filtering content, they would not incur liability for defamation of which they had no knowledge. On the other hand, if they filtered content but did so incompletely, they would be strictly liable for users’ defamatory comments.

*Stratton Oakmont* was anathema to federal lawmakers, who wanted to encourage platforms (and parents) to filter as much noxious material as possible. Although protected speech like vulgarity and pornography could not be regulated under the First Amendment, it could be blocked or removed by private actors—and lawmakers wanted to ensure that online service providers had every incentive to do so. Hence, Section 230’s safe harbor provision, which immunized online service providers for inexact screening.

Pros and Cons

Over the past twenty years, content moderation tools have grown in sophistication and variety. Automation has certainly made it cheaper and easier to detect, block, or remove illegality. In certain circumstances, it is the right—if not ideal—response, offering these upsides with little downside to expression and other values. Yet context is a crucial factor in automation’s success.

When it comes to combating child pornography, hash technology is the “killer app.” Hashing is “a mathematical operation that takes a long stream of data of arbitrary length, like a video clip or string of DNA, and assigns it a specific value of a fixed length, known as a hash. The same files or DNA strings will be given the same hash, allowing computers to quickly and easily spot duplicates.” In essence, hashes are unique digital fingerprints.

Computer scientist Hany Farid, in conjunction with Microsoft, developed PhotoDNA hash technology that enables the blocking of content before it appears. The National Center for Missing and Exploited Children (NCMEC) has put this technology to work by collecting hashes of content that it deems to constitute child pornography. Its database has been resoundingly effective: with access to the NCMEC database, platforms can prevent hashed
images from reappearing online. Moreover, the database has not proved vulnerable to “censorship creep,” a term that one of us has used to describe the expansion of initially limited content moderation to block or filter other types of material. This is because child pornography is easily defined and identified—and because NCMEC controls the database, their experts ensure that only child pornography is included.

This approach is being used to tackle nonconsensual pornography—nude images posted without individuals’ consent. Consider Facebook’s recent efforts. Since 2014, the company has banned nonconsensual pornography in its terms of service (TOS) agreement. Users would report images as TOS violations and the company would react to those requests, removing images where appropriate. Yet abusers would routinely repost the material once it had been removed, leading to a game of whack-a-mole. In April 2017, Facebook announced its adoption of hash techniques to prevent the cycle of reposting: users would report images as nonconsensual pornography as before, but now, the company’s “specially trained representative[s]” would determine if the images violate the company’s terms of service and then designate the images for hashing. Photo-matching technology would block hashed images from reappearing on any of the platforms owned by Facebook.

This program has great promise to mitigate the damage suffered by victims of nonconsensual pornography. Preventing the reappearance of nonconsensual pornography is a relief to victims, who can rest easy knowing that at least on Facebook and its properties, friends, family, and coworkers will not see their nude images without their consent. Of course, this solution is confined to Facebook—but its success might portend wider adoption as in the case of child pornography moderation efforts.

To prevent censorship creep, Facebook is providing special training to moderators that will enable them to distinguish legitimate from illegitimate claims (such as requests to hash disfavored images that do not constitute nonconsensual pornography). The company is also working with victims’ organizations like Cyber Civil Rights Initiative (CCRI) and the National Network to End Domestic Violence (NNEDV) to enhance its training efforts. In this way, Facebook is emulating the child pornography model: leveraging the expertise of CCRI and NNEDV to ensure that its moderators know what nonconsensual pornography is and what it isn’t.

Other contexts, however, run a considerable risk of censorship creep and are ill suited to a hash solution. The dominant tech companies—Facebook, Microsoft, and YouTube—have constructed an industry database of hashed terrorist, extremist content. Yet the companies lack a clear consensus about the meaning of “violent extremist content.” Currently, platforms have a range of definitions of terrorist content that violates their terms of service, from “content that promotes terrorist acts, incites violence, or celebrates terrorist attacks” to “specific threats of violence or wish for the serious physical harm, death, or disease of an
individual or group of people.” Whether content amounts to violent and egregious terrorist material depends on the overall context, including the message and precise wording. Violent terrorist speech may be precisely that—or it may be news or advocacy against violent ideologies. In August 2017, for example, YouTube put in place new machine-learning systems to remove terrorist propaganda from its platform—and immediately generated outrage when that effort led to the purging of videos gathered by journalism and advocacy groups to document war crimes in Syria.

Compounding these concerns, governments will surely capitalize on the lack of clarity in the meaning of terrorist content. Along these lines, a United Kingdom security and immigration minister has argued that platforms should block terrorist content even if it is not illegal because people do not want to see “unsavoury” material. Likewise, government authorities could suggest inclusion of hashed videos of pornography or political protests. Although companies are ultimately in charge of the databases at hand, a government request might lead them to include content that otherwise they would not.

More generally, platforms are poised to deploy AI technologies to filter all manner of speech. Automation, after all, makes it easy to scale up solutions. Yet algorithms may be prone to both false positives and false negatives. Platforms’ faith in technology must be tempered by careful evaluation of those problems.

Platforms should assess whether the costs to expression and user support are worth the benefits—speed and reduced costs, for instance—of employing algorithms to detect terms-of-service violations. Take, for example, the problem of hate speech. Researchers have found that using algorithms to detect hate speech results in far more false positives than false negatives because it is difficult to capture context—tone, speaker, and audience. Although natural language processing algorithms can be trained to detect combinations and collections of words, they cannot distinguish jokes, sarcasm, or rebuttals of hate speech from hateful statements. Algorithms also replicate bias in training data, and for that reason have been shown to perform less accurately when analyzing the language of female speakers and African American speakers.

Right now, it appears that Facebook is using automation to flag hate speech for content moderators rather than to remove it automatically. But not so for extremist, terrorist speech. As Zuckerberg testified at the House hearing, the company is using filtering technologies to tackle terrorist content and more than 99 percent of extremist, terrorist content is blocked before ever appearing. He suggested that in “five or ten years,” AI will be able to perform a similar role in flagging hate speech. For now, the rate of false positives (or false negatives) in flagging extremist content is unknown to the public, but we hope it is carefully considered internally.
In 2010, Paul Ohm, who is both a technologist and a legal scholar, wisely warned:

Technical solutions too often lead to unintended consequences. Anyone who has ever struggled to use a computer with an Internet filter, cursing at the false positives and giggling at the false negatives, can breathe a sigh of relief that the anti-porn crusaders never convinced anyone to place filters deep inside the network itself. Likewise, we should worry about the recording industry’s plans for ISP filtering and three strikes laws as overbroad, disproportionate measures. If anything, technical solutions may be even less likely to succeed against the problem of online harassment than in the earlier battles. Music and porn travel through the Internet as large files, which can be easy to identify through fingerprinting and filtering. In contrast, Cyber Civil Rights harms often involve threats and harassment buried in small snippets of text whose threatening nature must be judged by a person not a machine. For all of these reasons, we should be deploying surgical strikes, not napalm.

If companies are going to automate private speech policies, they ought to heed Ohm’s warning. The next section explores potential safeguards for companies to employ as they operationalize speech policies.

**Safeguards**

Platforms should voluntarily commit to a notion of “technological due process” to guide their thinking as they develop systems of automated content regulation. A model of technological due process would explicitly address the trade-offs of “automation and human discretion.” As companies respond to regulatory pressures and an increased internal sense of responsibility for the content on their platforms, they should nevertheless take care to interrogate the extent to which automation of speech policies would result in false positives and thus unwarranted private censorship. They should also consider whether over-censorship can be reduced.

Where automation is sure to result in excessive false positives, as in the case of hate speech, a more careful approach would keep humans in the loop, with AI techniques identifying content for moderators to review. In 2017, Mark Zuckerberg explained that Facebook was “researching [AI] systems that can look at photos and videos to flag content our team should review.” The company wanted to “use AI to tell the difference between news stories about terrorism and actual terrorist propaganda so [it] can quickly remove anyone trying to use our services to recruit for a terrorist organization.” Zuckerberg admitted then that the project was fraught with problems: “This is technically difficult as it requires building AI that can read and understand news, but we need to work on this to help fight terrorism worldwide.” Nonetheless, all signs suggest that Facebook has fully automated its detection, removal, and blocking of extremist, terrorist content.

If terrorist content is vulnerable to censorship creep, aggressive over-moderation also poses a serious risk of suppressing newsworthy content. Companies could consider hiring or
consulting ombudsmen whose life’s work is the news-gathering process. Ombudsmen, who are often called public editors, work to “protect press freedom” and to promote “high-quality journalism.” Their role is to act “in the best interests of news consumers.” Ombudsmen could review contributions to hash databases with the public interest in mind and oversee the use of AI techniques to moderate content, weighing their costs and benefits. This will become especially important as platforms fully automate the filtering process.

Then, too, companies should be transparent about their speech policies. They should be clear what they mean when they ban certain content and why such content is banned. They should provide accountability over speech decisions. Users should be notified that their content has been removed (or blocked) and given a meaningful chance to challenge the decision.

All this raises the question of whether an amendment to Section 230 (along the lines that we support) would press companies to automate more and more of their content moderation. Would platforms adopt AI technologies that filter speech essential for a healthy democracy? Would requiring platforms to respond reasonably to illegality lead to AI filtering techniques that impoverish online expression? Would platforms wield their enormous power over online expression in ways that undermine the spirit of New York Times v. Sullivan—that public discourse be “uninhibited, robust, and wide-open”?

While platforms have arguably under-moderated themselves over the two decades since the passage of Section 230, disclaiming responsibility for abuse propagated on their services, we now risk over-moderation in response to shifting political moods—a shift manifesting most dramatically in the United States but present on both sides of the Atlantic.

Widespread recognition of the destructive uses of technology platforms is overdue, but legislators and the platforms themselves must be wary of overcorrecting or correcting too swiftly and sloppily. The powerful consequences of platform misuse—not only on an individual scale in enabling harassment, for example, but on a societal scale by facilitating election interference and even genocide—should make clear the dangers of technology companies’ overly limited view of their own responsibilities.

Yet these systematic misuses also show that the design and regulation of major platforms can have dramatic ripple effects that go well beyond what the designers anticipated. For this reason, overhasty regulation, in the form of either legislation or platform self-moderation, is dangerous as well. “Magical thinking” poses risks in either direction.

Recent scholarship has explored how platforms can have effects on free speech that profoundly shape civic discourse and yet go beyond the scope of the First Amendment as usually understood. We face the problem of how to conceptualize the role of technology...
companies as “private gatekeepers” to speech in the context of a legal tradition that has
developed to constrain state action.85 Even in the absence of legal vocabulary or doctrine
with which to fully understand this question, legislators and platforms should keep in mind
the danger of both silencing speech and allowing it to be drowned out by louder, abusive
speech.

This is not a reason to hold off on efforts to improve the “speech environment” online.
Rather, it is a reason to “move cautiously and with intellectual and epistemic modesty” so
we are guided both by a clear sense of the realistic capabilities of existing technology and
by normative and legal commitments to enabling and protecting the free expression of
ideas.86

Conclusion

In an environment of increasing skepticism toward big technology companies, the
likelihood of stricter legislative regulation of online platforms increases as well. We point
to possibilities for revising Section 230 that avoid the trap of encouraging overly aggressive
moderation while addressing abusive and illegal uses of platforms more broadly than
piecemeal approaches like FOSTA will allow.

While platforms are also turning to algorithmic moderation in response to threatened
regulation by EU member states, we also caution against the danger that use of AI could
lead to censorship creep. Successfully addressing the problems that have generated the
current crisis of confidence in technology will require not only solutions to the specific
issues at hand but also a deeper understanding of the power of these platforms to shape our
discourse and of the need to proceed carefully.

NOTES

1 Ryan Mac, Charlie Warzel, and Alex Kantrowitz, “Top Facebook Executive Defended Data Collection in 2016
Memo—And Warned That Facebook Could Get People Killed,” Buzzfeed, March 29, 2018, accessed August 23, 2018,
.xvXDvrv8P#.pcXbRXRLG.

-scandal-fallout.html.

-wing-channels.html; Kurt Wagner, “Twitter is Wondering Whether Twitter is Bad for Society—And Jack Dorsey is

4 Danielle Keats Citron, Hate Crimes in Cyberspace (Cambridge, MA: Harvard University Press, 2014); Danielle Keats


10 We say “ultra-constitutional” to suggest the perceived importance of the statutory immunity and to recognize that the statute provides even more protection than the First Amendment would secure. For instance, under First Amendment doctrine, content providers could be sued for defamation for publishing factual falsehoods about public officials knowing or reckless to the falsity of the facts. See *New York Times v. Sullivan*, 376 U.S. 254 (1964). Under Section 230, however, platforms enjoy immunity from liability for publishing statements about public officials even if they knew, or were reckless as to the fact, that the statements were false.


18 Citron, “Extremist Speech.”


20 Selyukh, “Section 230.”


23 Citron, “Cyber Civil Rights.”


27 “Sense of” resolutions are generally not considered binding parts of legislation and are often included for expressive purposes.


38 Switter describes itself as “a sex work-friendly social space.” The front page of the website reads: “In light of the FOSTA/SESTA bill and the recent shadow-banning of our accounts, we’ve decided to take social into our own hands,” Switter, accessed August 23, 2018, https://twitter.at/about; Megan Farokhmanesh, “Amid FOSTA Crackdown, Sex
According to Assembly Four, which runs Mastodon, Twitter was down for a total of about seven hours. The company switched to another content delivery network service after the outage. As of August 23, 2018, Twitter remains live. See Assembly Four, “Cloudflare and FOSTA/SESTA,” accessed August 23, 2018, https://assemblyfour.com/twitter/cloudflare-fosta-sesta.


46 The Electronic Frontier Foundation and others have challenged this portion of the statute on overbreadth grounds, contending that the confusion would lead companies to censor too much speech, thereby chilling freedom of expression. Complaint, Woodhull Freedom Foundation v. United States, 18 CV 11552 (D. D.C. June 28, 2018). Although our criticism of FOSTA does not focus on First Amendment grounds, we are sympathetic to the litigants’ argument that the “knowingly facilitating” language raises vagueness concerns. In line with our concern here, we worry that FOSTA undermines Section 230’s goal to encourage Good Samaritan monitoring. FOSTA does the opposite, driving platforms to filter in an overly aggressive and counter productive way or to do no filtering at all.


49 Hwang, “Dealing With Disinformation.”

50 Email of Geoffrey Stone to Danielle Citron, April 8, 2018 (on file with authors).


52 Ibid.

53 Citron and Wittes, “The Internet Will Not Break.”

54 Ibid.
Indeed, while we were finishing this paper, the office of Senator Mark Warner (D-VA) proposed an alternative revision to the statute, under which “users who have successfully proved that sharing of particular content by another user constituted a dignitary tort” under state law could “give notice to a platform; with this notice, platforms would be liable in instances where they did not prevent the content in question from being re-uploaded in the future.” The senator’s proposal relies on the use of hashing technology to identify the offending content, which we discuss below. US Senator Mark R. Warner, “Potential Policy Proposals for Regulation of Social Media and Technology Firms” (8–9), accessed August 23, 2018, https://graphics.axios.com/pdf/PlatformPolicyPaper.pdf.

Citron, “Extremist Speech.”


Citron, “Extremist Speech.”

Ibid.

Nonconsensual pornography is now a crime in forty states and in the District of Columbia. The Cyber Civil Rights Initiative generally and Professor Mary Anne Franks specifically (the organization’s legislative director) were the inspiration and moving force behind those developments. See https://www.cybercivilrights.org, accessed August 23, 2018. See also Mary Anne Franks, “Revenge Porn’ Reform: A View from the Front Lines,” Florida Law Review 69 (2018): 1251. Federal lawmakers—nominally Senator Kamala Harris and Representative Jackie Speier (both California Democrats)—have proposed federal law to criminalize the practice.


“What Is the Facebook Safety Advisory Board and What Does This Board Do?” Facebook Help Center, accessed August 23, 2018, https://www.facebook.com/help/222332597793306. One of us (Citron) serves on CCRi’s Board of Directors and is on Facebook’s task force on non-consensual intimate images (an unpaid position).


74 Citron, Hate Crimes in Cyberspace. For the original piece, see Danielle Keats Citron, “Technological Due Process,” Washington University Law Review 85 (2008): 1249 (proposing strategies to honor commitments of due process and rulemaking when federal and state agencies use automated systems to make decisions about important rights).


76 Zuckerberg, “Building Global Community.”


78 Ibid.

79 Ibid.

80 See Citron, “Extremist Speech.”


85 Ibid.

86 Ibid.
The Working Group on National Security, Technology, and Law brings together national and international specialists with broad interdisciplinary expertise to analyze how technology affects national security and national security law and how governments can use that technology to defend themselves, consistent with constitutional values and the rule of law.

The group focuses on a broad range of interests, from surveillance to counterterrorism to the dramatic impact that rapid technological change—digitalization, computerization, miniaturization, and automaticity—are having on national security and national security law. Topics include cybersecurity, the rise of drones and autonomous weapons systems, and the need for—and dangers of—state surveillance. The group’s output will also be published on the Lawfare blog, which covers the merits of the underlying legal and policy debates of actions taken or contemplated to protect the nation and the nation’s laws and legal institutions.

Jack Goldsmith and Benjamin Wittes are the cochairs of the National Security, Technology, and Law Working Group.

For more information about this Hoover Institution Working Group, visit us online at http://www.hoover.org/research-teams/national-security-technology-law-working-group.
Every day, people come to Facebook to share their stories, see the world through the eyes of others, and connect with friends and causes. The conversations that happen on Facebook reflect the diversity of a community of more than two billion people communicating across countries and cultures and in dozens of languages, posting everything from text to photos and videos.

We recognize how important it is for Facebook to be a place where people feel empowered to communicate, and we take our role in keeping abuse off our service seriously. That’s why we have developed a set of Community Standards that outline what is and is not allowed on Facebook. Our Standards apply around the world to all types of content. They’re designed to be comprehensive – for example, content that might not be considered hate speech may still be removed for violating our bullying policies.

The goal of our Community Standards is to encourage expression and create a safe environment. We base our policies on input from our community and from experts in fields such as technology and public safety. Our policies are also rooted in the following principles:
Safety: People need to feel safe in order to build community. We are committed to removing content that encourages real-world harm, including (but not limited to) physical, financial, and emotional injury.

Voice: Our mission is all about embracing diverse views. We err on the side of allowing content, even when some find it objectionable, unless removing that content can prevent a specific harm. Moreover, at times we will allow content that might otherwise violate our standards if we feel that it is newsworthy, significant, or important to the public interest. We do this only after weighing the public interest value of the content against the risk of real-world harm.

Equity: Our community is global and diverse. Our policies may seem broad, but that is because we apply them consistently and fairly to a community that transcends regions, cultures, and languages. As a result, our Community Standards can sometimes appear less nuanced than we would like, leading to an outcome that is at odds with their underlying purpose. For that reason, in some cases, and when we are provided with additional context, we make a decision based on the spirit, rather than the letter, of the policy.
Everyone on Facebook plays a part in keeping the platform safe and respectful. We ask people to share responsibly and to let us know when they see something that may violate our Community Standards. We make it easy for people to report potentially violating content, including Pages, Groups, profiles, individual content, and/or comments to us for review. We also give people the option to block, unfollow, or hide people and posts, so that they can control their own experience on Facebook.

The consequences for violating our Community Standards vary depending on the severity of the violation and a person's history on the platform. For instance, we may warn someone for a first violation, but if they continue to violate our policies, we may restrict their ability to post on Facebook or disable their profile. We also may notify law enforcement when we believe there is a genuine risk of physical harm or a direct threat to public safety.

Our Community Standards, which we will continue to develop over time, serve as a guide for how to communicate on Facebook. It is in this spirit that we ask members of the Facebook community to follow these guidelines.
PART I.

Violence and Criminal Behavior

1. Credible Violence

We aim to prevent potential real-world harm that may be related to content on Facebook. We understand that people commonly express disdain or disagreement by threatening or calling for violence in facetious and non-serious ways. That's why we try to consider the language, context and details in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information like a targeted person's public visibility and vulnerability. We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.

READ MORE

2. Dangerous Individuals and Organizations

In an effort to prevent and disrupt real-world harm, we do not allow any organizations or individuals that are engaged in the following to have a presence on Facebook:

- Terrorist activity
- Organized hate
- Mass or serial murder
- Human trafficking
- Organized violence or criminal activity

We also remove content that expresses support or praise for groups, leaders, or individuals involved in these activities.
3. Promoting or Publicizing Crime

We prohibit people from promoting or publicizing violent crime, theft, and/or fraud because we do not want to condone this activity and because there is a risk of copycat behavior. We also do not allow people to depict criminal activity or admit to crimes they or their associates have committed. We do, however, allow people to debate or advocate for the legality of criminal activities, as well as address them in a rhetorical or satirical way.

4. Coordinating Harm

In an effort to prevent and disrupt real-world harm, we prohibit people from facilitating or coordinating future criminal activity that is intended or likely to cause harm to people, businesses, or animals. People can draw attention to harmful activity that they may witness or experience as long as they do not advocate for or coordinate harm.

5. Regulated Goods

To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, or trade non-medical drugs, pharmaceutical drugs, and marijuana. We also prohibit the purchase, sale, gifting, exchange, and transfer of firearms, including firearm parts or ammunition, between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may promote items available for sale off of our services as long as those retailers comply with all applicable laws and regulations. We allow discussions about sales of firearms...
and firearm parts in stores or by online retailers and advocating for changes to firearm regulation. Regulated goods that are not prohibited by our Community Standards may be subject to our more stringent Commerce Policies.
2. Dangerous Individuals and Organizations

Policy Rationale

In an effort to prevent and disrupt real-world harm, we do not allow any organizations or individuals that are engaged in the following to have a presence on Facebook:

- Terrorist activity
- Organized hate
- Mass or serial murder
- Human trafficking
- Organized violence or criminal activity

We also remove content that expresses support or praise for groups, leaders, or individuals involved in these activities.

We do not allow the following people (living or deceased) or groups to maintain a presence (for example, have an account, Page, Group) on our platform:

Terrorist organizations and terrorists

- A terrorist organization is defined as:
  - Any non-governmental organization that engages in premeditated acts
  - A member of a terrorist organization or any person who commits a terrorist act in its name
  - A terrorist act is defined as a premeditated act of violence against persons or property

Hate organizations and their leaders and prominent members

- A hate organization is defined as:
  - Any association of three or more people that is organized under a name

Mass and serial murderers
We consider a homicide to be a mass murder if it results in four or more deaths in one incident.

We consider any individual who has committed two or more murders over multiple incidents or locations a serial murderer.

We make these assessments based upon the information available to us. A mass or serial murderer who meets any of the following criteria:

- They were convicted of mass or serial murder.
- They were killed by law enforcement during commission of the mass or serial murder.
- They killed themselves at the scene or in the aftermath of the mass or serial murder.
- They were identified by law enforcement with images from the crime.

Human trafficking groups and their leaders

- Human trafficking groups are organizations responsible for any of the following:
  - Prostitution of others, forced/bonded labor, slavery, or the removal of others;
  - Recruiting, transporting, transferring, detaining, providing, harboring, or assisting in the commission of these crimes.

Criminal organizations and their leaders and prominent members

- A criminal organization is defined as:
  - Any association of three or more people that is united under a name, common purpose, and organization.

  - Homicide
  - Drug trafficking
  - Arms trafficking
  - Identity theft
  - Money laundering
  - Extortion or trafficking
  - Assault
  - Kidnapping
  - Sexual exploitation (covered in section 7 and section 8)

We do not allow symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content.

We do not allow content that praises any of the above organizations or individuals or any acts committed by them.

We do not allow coordination of support for any of the above organizations or individuals or any acts committed by them.
<table>
<thead>
<tr>
<th>1. Credible Violence</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Promoting or Publicizing Crime</td>
</tr>
</tbody>
</table>

---

Facebook © 2018

Report Abuse
3. Promoting or Publicizing Crime

Policy Rationale

We prohibit people from promoting or publicizing violent crime, theft, and/or fraud because we do not want to condone this activity and because there is a risk of copycat behavior. We also do not allow people to depict criminal activity or admit to crimes they or their associates have committed. We do, however, allow people to debate or advocate for the legality of criminal activities, as well as address them in a rhetorical or satirical way.

Do not post:

Content depicting, admitting, or promoting the following criminal acts committed by you or your associates

- Acts of physical harm committed against people
- Acts of physical harm committed against animals except in cases of hunting, fishing, religious sacrifice, or food preparation/processing
- Poaching or selling endangered species or their parts
- Staged animal vs. animal fights
- Theft
- Vandalism or property damage
- Fraud
- Trafficking as referenced in section 2
- Sexual violence or sexual exploitation, including sexual assault, as referenced in section 7 and section 8
4. Coordinating Harm

Policy Rationale

In an effort to prevent and disrupt real-world harm, we prohibit people from facilitating or coordinating future criminal activity that is intended or likely to cause harm to people, businesses, or animals. People can draw attention to harmful activity that they may witness or experience as long as they do not advocate for or coordinate harm.

Do not post:

Statements of intent, calls to action, or advocation for the following:

- Acts of physical harm committed against people
- Acts of physical harm committed against animals except in cases of hunting, fishing, religious sacrifice, or food preparation/processing
- Poaching or selling endangered species and their parts
- Staged animal vs. animal fights
- Theft
- Vandalism/property damage
- Fraud
- Arranged marriages with refugees or internally displaced persons
- Trafficking as referenced in section 2
- Sexual violence or sexual exploitation, including sexual assault, as referenced in section 7 and section 8

Offers of services to smuggle or assist in smuggling people.
5. Regulated Goods

Policy Rationale

To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, or trade non-medical drugs, pharmaceutical drugs, and marijuana. We also prohibit the purchase, sale, gifting, exchange, and transfer of firearms, including firearm parts or ammunition, between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may promote items available for sale off of our services as long as those retailers comply with all applicable laws and regulations. We allow discussions about sales of firearms and firearm parts in stores or by online retailers and advocating for changes to firearm regulation. Regulated goods that are not prohibited by our Community Standards may be subject to our more stringent Commerce Policies.

Do not post:

Content about non-medical drugs (other than alcohol or tobacco) that

• Coordinates or encourages others to sell non-medical drugs
• Depicts, admits to, or promotes sales of non-medical drugs by the poster of the content or their associates
• Promotes, encourages, coordinates, or provides instructions for use of non-medical drugs
• Admits, either in writing or verbally, to personal use of non-medical drugs unless posted in a recovery context

Content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs. This includes content that
· Mentions or depicts marijuana or pharmaceutical drugs
· Makes an attempt to sell or trade, by which we mean any of the following:
  · Explicitly mentioning the product is for sale or trade or delivery
  · Asking the audience to buy
  · Listing the price
  · Encouraging contact about the product either by explicitly asking to be contacted
  · Attempting to solicit the product, defined as:
    · Stating interest in buying the product, or
    · Asking if anyone has the product for sale/trade
· This applies to both individual pieces of content and Pages and Groups primarily dedicated to the sale of marijuana or pharmaceutical drugs

Content that attempts to sell, gift, exchange, or transfer firearms, firearm parts, ammunition, or explosives between private individuals. This includes content that

· Mentions or depicts firearms, firearm parts, ammunition, or explosives and a product unrelated to firearms, and
· Makes an attempt to sell or transfer including any of the following, unless posted by an entity representing a brick-and-mortar store, legitimate website, or brand:
  · Explicitly mentioning the product is for sale or trade or delivery
  · Asking the audience to buy
  · Listing the price or noting that the product is free
  · Encouraging contact about the product either by
    · Explicitly asking to be contacted
    · Including any type of contact information
  · Making an attempt to solicit the item for sale, defined as
    · Stating that they are interested in buying the good, or
    · Asking if anyone else has the good for sale/trade

Content that attempts to sell, gift, exchange, transfer, promote or otherwise provide access to instructions for 3D printing or computer-aided manufacturing of firearms or firearm parts.

⚠️

For the following content, we restrict visibility to adults twenty-one years of age and older:

Content posted by a brick-and-mortar store, legitimate website, or brand, which coordinates or promotes the sale or transfer of...
firearms, firearm parts, ammunition, or explosives. This includes content that

- Explicitly mentions the product is for sale or transfer and
- Asks the audience to buy the product, or
- Lists the price or notes that the product is free, or
- Encourages contact about the product either by explicitly asking to be c
4. Coordinating Harm

6. Suicide and Self-Injury
PART II.

Safety

6. Suicide and Self-Injury

In an effort to promote a safe environment on Facebook, we remove content that encourages suicide or self-injury, including real-time depictions that might lead others to engage in similar behavior. Self-injury is defined as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We want Facebook to be a space where people can share their experiences, raise awareness about these issues, and support each other through difficult experiences, and so we allow people to discuss suicide and self-injury. We encourage people to offer and to seek support from one another in connection with these difficult topics.

We work with organizations around the world to provide assistance to people in distress. We also talk with experts in suicide and self-injury to help inform our policies and enforcement. For example, we have been advised by experts that we should not remove live videos of self-harm while there is an opportunity for loved ones and authorities to provide help or resources.

We remove any content that identifies and negatively targets victims or survivors of self-injury or suicide seriously, humorously, or rhetorically. People can, however, share information about self-injury and suicide to draw attention to the issue and allow for discussion so long as they do not promote or encourage self-injury or suicide.

Learn more about our suicide and self-injury policies and the resources that we provide.
We do not allow content that sexually exploits or endangers children. When we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. We know that sometimes people share nude images of their own children with good intentions; however, we generally remove these images because of the potential for abuse by others and to help avoid the possibility of other people reusing or misappropriating the images.

We also work with external experts, including the Facebook Safety Advisory Board, to discuss and improve our policies and enforcement around online safety issues, especially with regard to children.

**8. Sexual Exploitation of Adults**

We recognize the importance of Facebook as a place to discuss and draw attention to sexual violence and exploitation. We believe this is an important part of building common understanding and community. In an effort to create space for this conversation while promoting a safe environment, we remove content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation, while also allowing space for victims to share their experiences. We remove content that displays, advocates for, or coordinates sexual acts with non-consenting parties or commercial sexual services. We do this to avoid facilitating transactions that may involve trafficking, coercion, and non-consensual sexual acts. "Sexual services" include prostitution, escort services, sexual massages, and filmed sexual activity.

To protect victims and survivors, we also remove images that depict incidents of sexual violence and intimate images shared without permission from the people pictured. For additional information on these efforts, please visit Using Technology to Protect Intimate Images and Help Build a Safe Community, as well as our guide to reporting and removing intimate images shared without your consent.
9. Bullying

Bullying happens in many places and comes in many different forms from making statements degrading someone's character to posting inappropriate images to threatening someone. We do not tolerate bullying on Facebook because we want the members of our community to feel safe and respected.

We will remove content that purposefully targets private individuals with the intention of degrading or shaming them. We recognize that bullying can be especially harmful to minors, and our policies provide heightened protection for minors because they are more vulnerable and susceptible to online bullying. In certain instances, we require the individual who is the target of bullying to report content to us before removing it.

Our bullying policies do not apply to public figures because we want to allow discourse, which often includes critical discussion of people who are featured in the news or who have a large public audience. Discussion of public figures nonetheless must comply with our Community Standards, and we will remove content about public figures that violates other policies, including hate speech or credible threats.

Our Bullying Prevention Hub is a resource for teens, parents, and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including on how to start important conversations for people being bullied, parents who have a child being bullied or accused of bullying, and educators who have students involved with bullying.

10. Harassment

We do not tolerate harassment on Facebook. We want people to feel safe to engage and connect with their community. Our harassment policy applies to both public and private individuals because we want to prevent unwanted or malicious contact on the platform. Context and intent matter, and we allow people to share and re-share posts if it is clear that something was shared in order to condemn or draw attention to harassment. In addition to
reporting such behavior and content, we encourage people to use tools available on Facebook to help protect against it.

11. Privacy Violations and Image Privacy Rights

Privacy and the protection of personal information are fundamentally important values for Facebook. We work hard to keep your account secure and safeguard your personal information in order to protect you from potential physical or financial harm. You should not post personal or confidential information about others without first getting their consent. We also provide people ways to report imagery that they believe to be in violation of their privacy rights.
6. Suicide and Self-Injury

Policy Rationale

In an effort to promote a safe environment on Facebook, we remove content that encourages suicide or self-injury, including real-time depictions that might lead others to engage in similar behavior. Self-injury is defined as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We want Facebook to be a space where people can share their experiences, raise awareness about these issues, and support each other through difficult experiences, and so we allow people to discuss suicide and self-injury. We encourage people to offer and to seek support from one another in connection with these difficult topics.

We remove any content that identifies and negatively targets victims or survivors of self-injury or suicide seriously, humorously, or rhetorically. People can, however, share information about self-injury and suicide to draw attention to the issue and allow for discussion so long as they do not promote or encourage self-injury or suicide.

Learn more about our suicide and self-injury policies and the resources that we provide.

Do not post:

Content that promotes, encourages, coordinates, or provides instructions for
Content about self-injury that contains promotional slogans without clear, anti-self-injury disclaimers

Except in limited situations of newsworthiness, it is against our policies to post content depicting a person who engaged in a suicide attempt or death by suicide.

For the following content, we include a warning screen so that people are aware the content may be disturbing:

- Photos or videos depicting a person’s death by suicide that are determined to be newsworthy
- Photos or videos depicting a person who engaged in euthanasia/assisted suicide in a medical setting

We provide resources to people who post written or verbal admissions of engagement in self injury, including:

- Suicide
- Euthanasia/assisted suicide
- Self mutilation
- Eating disorders
- Images where more than one cut of self mutilation is present on a body part and the primary subject of the image is one or more unhealed cuts
7. Child Nudity and Sexual Exploitation of Children

Policy Rationale

We do not allow content that sexually exploits or endangers children. When we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. We know that sometimes people share nude images of their own children with good intentions; however, we generally remove these images because of the potential for abuse by others and to help avoid the possibility of other people reusing or misappropriating the images.

We also work with external experts, including the Facebook Safety Advisory Board, to discuss and improve our policies and enforcement around online safety issues, especially with regard to children.

Do not post:

Content that depicts participation in or advocates for the sexual exploitation of children, including (but not limited to)

- Engaging in any sexual activity involving minors
- Soliciting, displaying, sharing, or viewing imagery of nude, sexualized, or sexual activity with minors
- Arranging real-world sexual encounters or obtaining sexual material from a minor directly
- Adults soliciting minors
- Minors soliciting minors
- Displaying nudity to minors
- Minors soliciting adults
- Using our products and site functionality with the intention of sexualizing minors
Content (including photos, videos, real-world art, digital content, and text) that depicts

• Any sexual activity involving minors
• Minors in a sexual fetish context
• Minors with sexual elements, including (but not limited to):
  • Restraints
  • Focus on genitals
  • Presence of aroused adult
  • Presence of sex toys
  • Sexualized costume
  • Stripping
  • Staged environment (for example, on a bed) or professionally shot (quality or location)
  • Open-mouth kissing with minor or adult

Content (including photos, videos, real-world art, digital content, and verbal depictions) that shows minors in a sexualized context

Content that depicts child nudity where nudity is defined as

• Visible genitalia (even when covered or obscured by transparent clothing)
• Visible anus and/or fully nude close-up of buttocks
• Uncovered female nipples for children older than toddler-age
• No clothes present from neck to knee for children older than toddler-age
• Digitally-created depictions of nude minors, unless the image is for health or educational purposes
8. Sexual Exploitation of Adults

Policy Rationale

We recognize the importance of Facebook as a place to discuss and draw attention to sexual violence and exploitation. We believe this is an important part of building common understanding and community. In an effort to create space for this conversation while promoting a safe environment, we remove content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation, while also allowing space for victims to share their experiences. We remove content that displays, advocates for, or coordinates sexual acts with non-consenting parties or commercial sexual services. We do this to avoid facilitating transactions that may involve trafficking, coercion, and non-consensual sexual acts. "Sexual services" include prostitution, escort services, sexual massages, and filmed sexual activity.

To protect victims and survivors, we also remove images that depict incidents of sexual violence and intimate images shared without permission from the people pictured. For additional information on these efforts, please visit Using Technology to Protect Intimate Images and Help Build a Safe Community, as well as our guide to reporting and removing intimate images shared without your consent.

Do not post:

In instances where content consists of any form of non-consensual sexual touching, crushing, necrophilia or bestiality, including:

- Depictions (including real photos/videos), or
- Advocacy (including aspirational and conditional statements), or
- Statements of intent, or
- Calls for action, or
-
Participation by yourself or others to engage in any form of the above mentioned sexual acts.

Content that attempts to exploit people by any of the following:

- Coercing money, favors, or images from people by threats of exposure of their naked or semi-naked photos/videos
- Sharing imagery that fulfills all three of the following conditions:
  - Image is non-commercial or produced in a private setting
  - Person in the image is (near) nude, engaged in sexual activity, or in a sexualized pose
  - Lack of consent to share the image is indicated by
    - Vengeful context (for example, caption, comments, or page title)
    - Independent sources (for example, media coverage or law enforcement)
    - A visible match between the person depicted in the image and the person who reported the content to us shares the same name as
- Sharing images of a real person focusing on sexualized areas of the body such as the breasts, groin, or buttocks (also known as creepshots or upskirts). The following elements need to be present:
  - the focal point is on a sexualized area of the body
  - the person in the image is clearly unaware
  - Threatening or stating an intent to share intimate imagery without consent
  - Soliciting intimate imagery to view or share without consent
  - Threatening or stating an intent to share private sexual conversations

Attempts to coordinate adult sexual services or engaging in sexual solicitation including (but not limited to)

- Escort services
- Prostitution
- Filmed sexual encounters
- Sexualized massage
- Requesting rates on images of escorts
- Offering contact information with escort images or sexual solicitation slang terms
- Paid domination service
- Offering or soliciting sex or sexual fetish partners
For the following content, we include a warning screen so people are aware the content may be disturbing:

Narratives and statements that contain a depiction of non-consensual sexual touching (written or verbal) that includes details beyond mere naming or mentioning the act if:

- Shared by the victim, or
- Shared by a third party (other than the victim) in support of victim or condemnation of act or for general awareness to be determined by context/caption
9. Bullying

Policy Rationale

Bullying happens in many places and comes in many different forms from making statements degrading someone's character to posting inappropriate images to threatening someone. We do not tolerate bullying on Facebook because we want the members of our community to feel safe and respected.

We will remove content that purposefully targets private individuals with the intention of degrading or shaming them. We recognize that bullying can be especially harmful to minors, and our policies provide heightened protection for minors because they are more vulnerable and susceptible to online bullying. In certain instances, we require the individual who is the target of bullying to report content to us before removing it.

Our bullying policies do not apply to public figures because we want to allow discourse, which often includes critical discussion of people who are featured in the news or who have a large public audience. Discussion of public figures nonetheless must comply with our Community Standards, and we will remove content about public figures that violates other policies, including hate speech or credible threats.

Our Bullying Prevention Hub is a resource for teens, parents, and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including on how to start important conversations for people being bullied, parents who have a child being bullied or accused of bullying, and educators who have students involved with bullying.

Do not post:

Content about another private individual that reflects...
• Claims about sexual activity
  • High-severity physical descriptions
  • Ranking individuals on physical appearance or personality
  • Threats of non-consensual sexual touching
  • Sexualized text targeting another individual
  • An individual in a context that is intended to degrade, for example, menstruating, urinating, vomiting, or defecating
  • Physical bullying where the context further degrades the individual
  • Comparison to animals that are culturally perceived as intellectually or physically inferior or to an inanimate object

Content that has been photoshopped to target and demean an individual, including by highlighting specific physical characteristics or threatening violence in text or with imagery

Content that specifies an individual as the target of

• Statements of intent to commit violence
• Calls for action of violence
• Statements advocating violence
• Aspirational and conditional statements of violence
• Physical bullying
• Claims about religious identity or blasphemy

In addition, we may remove Pages or Groups that are dedicated to attacking individual(s) by, for example

• Cursing at an individual or individuals
• Making negative character claims
• Making negative ability claims
• Claims about blasphemy
• Appearing to be first person but is actually posted by a different individual than the person referenced and targets more than one individual

We also remove content that is targeted at minors when it contains:

• Cursing
• Claims about romantic involvement or sexual orientation
• Allegations about criminal or illegal behavior
• Coordinating, advocating, or promoting exclusion
• Negative character claims
• Negative ability claims
• Expressions of contempt or disgust
• Calls for death or serious disease or disability
In some cases, content is written in the first person but is actually posted by a different individual than the person referenced in the content. This may be done to target the person in the content with the intention of degrading or shaming them. We remove:

Content that contains the following and is reported by the individual depicted:

- Claims about sexual activity
- Comparisons to animals that are culturally perceived as intellectually or physically inferior or to an inanimate object
- High-severity physical descriptions
- Ranking individuals on physical appearance or personality
- Cursing at a person
- Claims about romantic involvement or sexual orientation
- Negative character or ability claims

For the following content, we include a warning screen so that people are aware the content may be disturbing:

Videos of physical bullying or violence against minors in a fight context shared with a condemning caption
| Community Standards |  
|---------------------|---
| 8. Sexual Exploitation of Adults |  
| 10. Harassment |  

Facebook © 2018
10. Harassment

Policy Rationale

We do not tolerate harassment on Facebook. We want people to feel safe to engage and connect with their community. Our harassment policy applies to both public and private individuals because we want to prevent unwanted or malicious contact on the platform. Context and intent matter, and we allow people to share and re-share posts if it is clear that something was shared in order to condemn or draw attention to harassment. In addition to reporting such behavior and content, we encourage people to use tools available on Facebook to help protect against it.

Do not:

Repeatedly contact a single person despite that person's clear desire and action to prevent that contact

Repeatedly contact large numbers of people with no prior solicitation

Send messages to any individual that contain

- Targeted cursing aimed at an individual or group of individuals in the thread
- Calls for death, serious disease or disability, or physical harm aimed at an individual or group of individuals in the thread
- Bullying policy violations
- Claims that a victim of a violent tragedy is lying about being a victim, acting/pretending to be a victim of a verified event, or otherwise is paid or employed to mislead people about their role in the event when sent directly to a survivor and/or immediate family member of a survivor or victim
Send messages to a group that contain any bullying policy violations, regardless of whether the person being targeted is a public or private individual.

Target anyone maliciously, including public figures, by:

- Attacking them based on their status as a victim of sexual assault or sexual exploitation
- Threatening any participant in public discourse with violence in an attempt to intimidate or silence them
- Calling for self-injury or suicide of a specific person, or group of people

Target victims or survivors of violent tragedies by name or by image, with claims that they are:

- Lying about being a victim of an event
- Acting/pretending to be a victim of an event
- Otherwise paid or employed to mislead people about their role in the event

Target a minor who is a public figure with:

- Claims about sexual activity or sexually transmitted disease(s)
- Content has been photoshopped to include threats of violence either in text or image (for example, adding bullseye, dart, gun to head)
- Calls for death or serious disease or disability
- Statements of intent to commit violence or low severity harm in an attempt to silence someone
- Objects created to attack through:
  - Targeted cursing
  - High-severity physical description
  - Claims about blasphemy
  - Expressions of contempt
  - Expressions of disgust
11. Privacy Violations and Image Privacy Rights
11. Privacy Violations and Image Privacy Rights

Policy Rationale

Privacy and the protection of personal information are fundamentally important values for Facebook. We work hard to keep your account secure and safeguard your personal information in order to protect you from potential physical or financial harm. You should not post personal or confidential information about others without first getting their consent. We also provide people ways to report imagery that they believe to be in violation of their privacy rights.

Do not post:

Content that facilitates identity theft by posting or soliciting personally identifiable information, including (but not limited to)

- National identification numbers, Social Security numbers, passport numbers, or exam numbers
- Government IDs
- School and education IDs featuring two of the following: (1) name, (2) photo, or (3) ID number
- Digital identities, including passwords

Content that contains medical/psychological, biometric, or genetic details of others

Content that facilitates identity theft by sharing personally identifiable information via an external link

Content that facilitates identity theft by sharing private financial information of an organization or business

Content that facilitates identity theft by disclosing the following personal financial information (of either the self or others)
• Bank account and/or card information
• Financial records paired with account information

Content that facilitates identity theft by sharing the private contact information of others defined as
• Private phone numbers or addresses
• Email, Messenger, and chat identities
• The above information may be shared to promote charitable causes, non-violating services, or to facilitate finding missing people or animals

Except in limited cases of newsworthiness, content claimed or confirmed to come from a hacked source, regardless of whether the affected person is a public figure or a private individual.

Content that identifies individuals by name and depicts their personal information, including:
• Driver's licenses, Government IDs other than driver's licenses, Green Cards, or immigration papers
• Marriage, birth, and name change certificates
• Digital identities, including passwords
• License plates

Content that includes photographs that display the external view of private residences if the following conditions apply:
• The residence is a single-family home, or the resident's unit number is identified in the image/caption
• The city or neighborhood is identified
• A resident is mentioned or depicted
• That same resident objects to the exposure of their private residence

Content that exposes the undercover status of law enforcement personnel if
• The content contains the agent's full name or other explicit identification and explicitly mentions their undercover status, or
• The content contains images identifying the faces of the law enforcement personnel and explicitly mentions their undercover status

Content that exposes information about safe houses by sharing any of the below, unless the safe house is actively promoting its
location, contact information, or the type of service and protection it offers through comments, posts, Pages or Groups:

- Actual address of the safe house (post box only is allowed),
- Images of the safe house,
- Identifiable city/neighborhood of the safe house, or
- Information outing residents of the safe house

The following content also may be removed

- A reported photo or video of people where the person depicted in the image
  - A minor under thirteen years old, and the content was reported by the person depicted in the image
  - A minor between thirteen and eighteen years old, and the content was reported by the person depicted in the image
  - An adult, where the content was reported by the adult from outside the minors
  - Any person who is incapacitated and unable to report the content on their own
PART III.

Objectionable Content

12. Hate Speech

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Sometimes people share content containing someone else’s hate speech for the purpose of raising awareness or educating others. Similarly, in some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. When this is the case, we allow the content, but we expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

We allow humor and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

Click here to read our Hard Questions Blog and learn more about our approach to hate speech.
We remove content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation. We allow graphic content (with some limitations) to help people raise awareness about issues. We know that people value the ability to discuss important issues like human rights abuses or acts of terrorism. We also know that people have different sensitivities with regard to graphic and violent content. For that reason, we add a warning label to especially graphic or violent content so that it is not available to people under the age of eighteen and so that people are aware of the graphic or violent nature before they click to see it.

**14. Adult Nudity and Sexual Activity**

We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content. Additionally, we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or satirical purposes.

Our nudity policies have become more nuanced over time. We understand that nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause, or for educational or medical reasons. Where such intent is clear, we make allowances for the content. For example, while we restrict some images of female breasts that include the nipple, we allow other images, including those depicting acts of protest, women actively engaged in breast-feeding, and photos of post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures.

**15. Cruel and Insensitive**

We believe that people share and connect more freely when they do not feel targeted based on their vulnerabilities. As such, we have...
higher expectations for content that we call cruel and insensitive, which we define as content that targets victims of serious physical or emotional harm.
12. Hate Speech

Policy Rationale

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Sometimes people share content containing someone else’s hate speech for the purpose of raising awareness or educating others. Similarly, in some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. When this is the case, we allow the content, but we expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

We allow humor and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

Click here to read our Hard Questions Blog and learn more about our approach to hate speech.

Do not post:
**Tier 1** attacks, which target a person or group of people who share one of the above-listed characteristics or immigration status (including all subsets except those described as having carried out violent crimes or sexual offenses), where attack is defined as

- Any violent speech or support in written or visual form
- Dehumanizing speech such as reference or comparison to:
  - Insects
  - Animals that are culturally perceived as intellectually or physically inferior
  - Filth, bacteria, disease and feces
  - Sexual predator
  - Subhumanity
  - Violent and sexual criminals
  - Other criminals (including but not limited to "thieves," "bank robbers," or "violent and sexual criminals")
  - Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image
  - Designated dehumanizing comparisons in both written and visual form

**Tier 2** attacks, which target a person or group of people who share any of the above-listed characteristics, where attack is defined as

- Statements of inferiority or an image implying a person's or a group's physical, mental, or moral deficiency
  - Physical (including but not limited to "deformed," "undeveloped," "hideous," "ugly")
  - Mental (including but not limited to "retarded," "cretin," "low IQ," "stupid,
  - Moral (including but not limited to "slutty," "fraud," "cheap," "free riders")
- Expressions of contempt or their visual equivalent, including (but not limited to)
  - "I hate"
  - "I don't like"
  - "X are the worst"
- Expressions of disgust or their visual equivalent, including (but not limited to)
  - "Gross"
  - "Vile"
  - "Disgusting"
  - Cursing at a person or group of people who share protected characteristics

**Tier 3** attacks, which are calls to exclude or segregate a person or group of people based on the above-listed characteristics. We do allow criticism of immigration policies and arguments for restricting those policies.
Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.
13. Violence and Graphic Content

Policy Rationale

We remove content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation. We allow graphic content (with some limitations) to help people raise awareness about issues. We know that people value the ability to discuss important issues like human rights abuses or acts of terrorism. We also know that people have different sensitivities with regard to graphic and violent content. For that reason, we add a warning label to especially graphic or violent content so that it is not available to people under the age of eighteen and so that people are aware of the graphic or violent nature before they click to see it.

Do not post:

Imagery of violence committed against real people or animals with comments or captions by the poster that contain

- Enjoyment of suffering
- Enjoyment of humiliation
- Erotic response to suffering
- Remarks that speak positively of the violence; or
- Remarks indicating the poster is sharing footage for sensational viewing pleasure

Videos of dying, wounded, or dead people if they contain

- Dismemberment unless in a medical setting
- Visible internal organs
- Charred or burning people
- Victims of cannibalism

Videos that show child abuse, defined as
• Repeated kicking, beating, slapping, or stepping on by an adult or animal
• Strangling or suffocating by an adult or animal
• Drowning by an adult or animal
• Biting through skin by an adult or animal
• Poisoning by an adult
• Forcible restraint by an adult
• Inflicting of burn or cut wounds by an adult
• Forcible smoking
• Tossing, rotating, or shaking of an infant (too young to stand) by their wrists/ankles, arms/legs, or neck

For the following content, we include a warning screen so that people are aware the content may be disturbing. We also limit the ability to view the content to adults, ages eighteen and older:

Imagery featuring mutilated people if it contains the following in a medical setting

• Dismemberment
• Visible internal organs
• Charred or burning people
• Victims of cannibalism
• Throat-slitting

Videos of self-immolation when that action is a form of political speech or newsworthy

Photos of wounded or dead people if they show

• Dismemberment
• Visible internal organs
• Charred or burning people
• Victims of cannibalism
• Throat-slitting

Imagery featuring animals that shows

• Dismemberment
• Visible internal organs
• Charred or burning animals
Videos of animal abuse, defined as:

- Repeated kicking or beating of an animal by a person
- Acts of torture by a person committed against animals
- Biting of an animal for abusive purposes
- Repeated animal-to-animal attacks
- Animals being killed in a hunting, manufacturing, or food preparation/processing context

Videos or photos that show child abuse, defined above, when accompanied by captions or context condemning or raising awareness about the abuse

Videos that show the violent death of a person or people by accident or murder

Still images depicting the violent death of a person or people

Videos that show acts of torture committed against a person or people

Videos of physical bullying or violence against minors in a fight context shared with a condemning caption

Photos and videos of non-medical foreign objects (such as metal objects, knives, nails) involuntarily inserted or stuck inside people or animals causing grievous injury

Videos of charred or burning people in the context of cremation

⚠️

For the following content, we include a label so that people are aware the content may be sensitive:

Imagery of fetuses that show:

- Dismemberment
- An abortion or abandonment context
- Imagery of newborn babies in an abandonment context

Photos and videos of animals in a ritual slaughter context showing dismemberment, or visible innards, or charring and burning
**14. Adult Nudity and Sexual Activity**

**Policy Rationale**

We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content. Additionally, we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or satirical purposes.

Our nudity policies have become more nuanced over time. We understand that nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause, or for educational or medical reasons. Where such intent is clear, we make allowances for the content. For example, while we restrict some images of female breasts that include the nipple, we allow other images, including those depicting acts of protest, women actively engaged in breast-feeding, and photos of post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures.

**Do not post:**

Images of

- Real nude adults, where nudity is defined as
  - Visible genitalia except in the context of birth giving and after-birth moments
  - Visible anus and/or fully nude close-ups of buttocks unless photoshopped
  - Uncovered female nipples except in the context of breastfeeding, birth giving
- Sexual activity, including
  - Sexual intercourse
  - Explicit sexual intercourse, defined as mouth or genitals entering or in contact
  - Implied sexual intercourse, defined as mouth or genitals entering or in contact
  - Implied stimulation of genitalia/anus, defined as stimulating genitalia/anus

--

https://www.facebook.com/communitystandards/adult_nudity_sexual_activity
Other sexual activities including (but not limited to):

- Erections
- Presence of by-products of sexual activity
- Stimulating genitals or anus, even if above or under clothing
- Use of sex toys, even if above or under clothing
- Stimulation of naked human nipples
- Squeezing naked female breast except in breastfeeding context
- Fetish content that involves
  - Acts that are likely to lead to the death of a person or animal
  - Dismemberment
  - Cannibalism
  - Feces, urine, spit, snot, menstruation, or vomit

Digital content that meets our definition of sexual activity unless any of the following conditions exist:

- Content was posted in a satirical or humorous context
- Content was posted in an educational or scientific context
- Imagery is not sufficiently detailed and only body shapes or contours are visible
- Content meets the definition of implied sexual activity.

Sexually explicit language, defined as description that goes beyond mere mention of:

- A state of sexual arousal
- An act of sexual intercourse or sexual activity, unless posted in an attempt at humor or satire, or is educational in nature
15. Cruel and Insensitive

Policy Rationale

We believe that people share and connect more freely when they do not feel targeted based on their vulnerabilities. As such, we have higher expectations for content that we call cruel and insensitive, which we define as content that targets victims of serious physical or emotional harm.

Do not post:

Content that depicts real people and mocks their implied or actual serious physical injuries, disease, or disability, non-consensual sexual touching, or premature death.
PART IV.

Integrity and Authenticity

16. Spam

We work hard to limit the spread of commercial spam to prevent false advertising, fraud, and security breaches, all of which detract from people's ability to share and connect. We do not allow people to use misleading or inaccurate information to collect likes, followers, or shares.

17. Misrepresentation

Authenticity is the cornerstone of our community. We believe that people are more accountable for their statements and actions when they use their authentic identities. That's why we require people to connect on Facebook using the name they go by in everyday life. Our authenticity policies are intended to create a safe environment where people can trust and hold one another accountable.

18. False News

Reducing the spread of false news on Facebook is a responsibility that we take seriously. We also recognize that this is a challenging and sensitive issue. We want to help people stay informed without stifling productive public discourse. There is also a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, significantly reduce its distribution by showing it lower in the News Feed.
19. Memorialization

When we learn that someone has passed away, we memorialize the account by adding "Remembering" above the name on the person’s profile. This makes it clear that the account is now a memorial site and protects against attempted logins and fraudulent activity. We do not remove, update, or change anything about the profile or the account because we want to respect the choices someone made while still alive. We have also made it possible for people to identify a legacy contact to look after their account after they pass away. People can let us know in advance if they want their account permanently deleted when they die. We may delete profiles when the next of kin tells us that the person who passed would have preferred that we delete the account rather than memorialize it. Visit Hard Questions for more information about our memorialization policy and process.
16. Spam

Policy Rationale

We work hard to limit the spread of commercial spam to prevent false advertising, fraud, and security breaches, all of which detract from people's ability to share and connect. We do not allow people to use misleading or inaccurate information to collect likes, followers, or shares.

Do not:

- Artificially increase distribution for financial gain
- Create or use fake accounts or compromise other people's accounts to
  - Impersonate or pretend to be a business, organization, public figure, or
  - Attempt to create connections, create content, or message people
- Restrict access to content by requiring people to like, share, or recommend before viewing
- Encourage likes, shares, or clicks under false pretenses
- Maliciously use login credentials or personally identifiable information by:
  - Attempting to gather or share login credentials or personally identifiable
  - Using another person's login credentials or personally identifiable inform
  - Promise non-existent Facebook features
Community Standards

Introduction

I. Violence and Criminal Behavior

II. Safety

III. Objectionable Content

IV. Integrity and Authenticity

17. Spam

18. False News

19. Memorialization

V. Respecting Intellectual Property

VI. Content-Related Requests

17. Misrepresentation

Policy Rationale

Authenticity is the cornerstone of our community. We believe that people are more accountable for their statements and actions when they use their authentic identities. That's why we require people to connect on Facebook using the name they go by in everyday life. Our authenticity policies are intended to create a safe environment where people can trust and hold one another accountable.

Do not:

Misrepresent your identity by

• Using a name that does not abide by our name policies
• Providing a false date of birth

Misuse our profiles product by

• Creating a profile for someone under thirteen years old
• Maintaining multiple accounts
• Creating inauthentic profiles
• Sharing an account with any other person
• Creating another account after being banned from the site
• Evading the registration requirements outlined in our Terms of Service

Impersonate others by

• Using their images with the explicit aim to deceive people
• Creating a profile assuming the persona of or speaking for another person or entity
• Creating a Page assuming to be or speak for another person or entity for whom the user is not authorized to do so.
• Posting imagery that is likely to deceive the public as to the content's origin, if:
The entity or an authorized representative objects to the content, and
Can establish a risk of harm to members of the public.

Engage in inauthentic behavior, which includes creating, managing,
or otherwise perpetuating

- Accounts that are fake
- Accounts that have fake names
- Accounts that participate in, or claim to engage in, coordinated inauthentic
any of the following:
- Mislead people about the origin of content
- Mislead people about the destination of links off our services (for example
- Mislead people in an attempt to encourage shares, likes, or clicks
- Mislead people to conceal or enable the violation of other policies under
16. Spam

18. False News
18. False News

Policy Rationale

Reducing the spread of false news on Facebook is a responsibility that we take seriously. We also recognize that this is a challenging and sensitive issue. We want to help people stay informed without stifling productive public discourse. There is also a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, significantly reduce its distribution by showing it lower in the News Feed.

We are working to build a more informed community and reduce the spread of false news in a number of different ways, namely by

- Disrupting economic incentives for people, Pages, and domains that propagate misinformation
- Using various signals, including feedback from our community, to inform a machine learning model that predicts which stories may be false
- Reducing the distribution of content rated as false by independent third-party fact-checkers
- Empowering people to decide for themselves what to read, trust, and share by informing them with more context and promoting news literacy
- Collaborating with academics and other organizations to help solve this challenging issue
19. Memorialization

Policy Rationale

When we learn that someone has passed away, we memorialize the account by adding “Remembering” above the name on the person's profile. This makes it clear that the account is now a memorial site and protects against attempted logins and fraudulent activity. We do not remove, update, or change anything about the profile or the account because we want to respect the choices someone made while still alive. We have also made it possible for people to identify a legacy contact to look after their account after they pass away. People can let us know in advance if they want their account permanently deleted when they die. We may delete profiles when the next of kin tells us that the person who passed would have preferred that we delete the account rather than memorialize it. Visit Hard Questions for more information about our memorialization policy and process.

Once we become aware that a person has passed away:

- Their account is secured and memorialized by adding “Remembering” above the name on the person's profile. The profile or account is not updated or otherwise changed in any way
- Immediate family members can also request that we remove a loved one’s profile
PART V.

Respecting Intellectual Property

20. Intellectual Property

Facebook takes intellectual property rights seriously and believes they are important to promoting expression, creativity, and innovation in our community. You own all of the content and information you post on Facebook, and you control how it is shared through your privacy and application settings. However, before sharing content on Facebook, please be sure you have the right to do so. We ask that you respect other people’s copyrights, trademarks, and other legal rights. We are committed to helping people and organizations promote and protect their intellectual property rights. Facebook’s Terms of Service do not allow people to post content that violates someone else’s intellectual property rights, including copyright and trademark. We publish information about the intellectual property reports we receive in our bi-annual Transparency Report, which can be accessed at https://transparency.facebook.com/
20. Intellectual Property

Policy Rationale

Facebook takes intellectual property rights seriously and believes they are important to promoting expression, creativity, and innovation in our community. You own all of the content and information you post on Facebook, and you control how it is shared through your privacy and application settings. However, before sharing content on Facebook, please be sure you have the right to do so. We ask that you respect other people’s copyrights, trademarks, and other legal rights. We are committed to helping people and organizations promote and protect their intellectual property rights. Facebook’s Terms of Service do not allow people to post content that violates someone else’s intellectual property rights, including copyright and trademark. We publish information about the intellectual property reports we receive in our bi-annual Transparency Report, which can be accessed at https://transparency.facebook.com/

Upon receipt of a report from a rights holder or an authorized representative, we will remove or restrict content that engages in:

- Copyright infringement
- Trademark infringement
PART VI.

Content-Related Requests

21. User Requests

We comply with:

• User requests for removal of their own account
• Requests for removal of a deceased user’s account from a verified immediate family member or executor
• Requests for removal of an incapacitated user’s account from an authorized representative

READ MORE

22. Additional Protection of Minors

We comply with:

• Requests for removal of an underage account
• Government requests for removal of child abuse imagery depicting, for example, beating by an adult or strangling or suffocating by an adult
• Legal guardian requests for removal of attacks on unintentionally famous minors

READ MORE
21. User Requests

We comply with:

- User requests for removal of their own account
- Requests for removal of a deceased user’s account from a verified immediate family member or executor
- Requests for removal of an incapacitated user’s account from an authorized representative
22. Additional Protection of Minors

We comply with:

- Requests for removal of an underage account
- Government requests for removal of child abuse imagery depicting, for example, beating by an adult or strangling or suffocating by an adult
- Legal guardian requests for removal of attacks on unintentionally famous minors