A HOOVER INSTITUTION ESSAY ON

Platform Justice

CONTENT MODERATION AT AN INFLECTION POINT

DANIELLE CITRON & QUINTA JURECIC

Aegis Series Paper No. 1811

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."[1]

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.[2] Meanwhile, Facebook's competitors, Twitter and Google, faced mounting criticism over their failure to curb disinformation and abusive content.[3]

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.[4] Even Facebook CEO Mark Zuckerberg's 2017 manifesto on "Building Global Community" recognized the use of Facebook for harmful ends like "bullying and harassment."[5]

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.[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.[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."[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."[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.[18] 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.[19]

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."[20] 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.[21] Platforms can continue to argue that their advertising systems facilitating illegal discrimination are immunized from liability.[22]

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.[23] As Rebecca Tushnet put it well, Section 230 "allows Internet intermediaries to have their free speech and everyone else's too."[24] 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.[25]

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."[42] 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."[43] 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."[44]

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.[45] 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.[46] 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.[47]

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,[48] violations of federal election law (a topic of new importance given the ongoing investigations into Russian election interference), or other illegal activity.[49]

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."[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.[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.[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.[53] A better revision to Section 230(c)(1) would read (revised language is italicized):

> "No provider or user of an interactive computer service that *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 *in any action arising out of the publication of content provided by that information content provider.*"

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.[59] 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.[60] 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."[61] 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.[62] 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.[63]

This approach is being used to tackle nonconsensual pornography—nude images posted without individuals' consent.[64] 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.[65] 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.[66] 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.[67] Yet the companies lack a clear consensus about the meaning of "violent extremist content."[68] 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."[69] 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.[70]

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.[71] Likewise, government authorities could suggest inclusion of hashed videos of pornography or political protests.[72] 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.[73]

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.[74] A model of technological due process would explicitly address the trade-offs of "automation and human discretion."[75] 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."[76] 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.[77] Ombudsmen, who are often called public editors, work to "protect press freedom" and to promote "high-quality journalism."[78] Their role is to act "in the best interests of news consumers."[79] 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.[80]

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"?[81]

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.[82]

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.[83] "Magical thinking" poses risks in either direction.[84]

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, https://www.buzzfeed.com/ryanmac/growth-at-any-cost-top-facebook-executive-defended-data?utm_term=.xvXDvrv8P#.pcXbRXRLG.

2 Nicholas Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," *New York Times,* April 4, 2018, accessed August 23, 2018, https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

3 Jonah Engel Bromwich, "YouTube Cracks Down on Far-Right Videos as Conspiracy Theories Spread," *New York Times,* March 3, 2018, accessed August 23, 2018, https://www.nytimes.com/2018/03/03/technology/youtube-right-wing-channels.html; Kurt Wagner, "Twitter is Wondering Whether Twitter is Bad for Society—And Jack Dorsey is Starting New Research to Find Out," *Recode,* March 1, 2018, accessed August 23, 2018, https://www.recode.net/2018/3/1/17067070/twitter-tweets-abuse-harassment-health-measurement-safety-jack-dorsey.

4 Danielle Keats Citron, *Hate Crimes in Cyberspace* (Cambridge, MA: Harvard University Press, 2014); Danielle Keats Citron, "Extremist Speech, Compelled Conformity, and Censorship Creep," *Notre Dame Law Review* 93 (2018): 1035;

Danielle Keats Citron and Helen Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age," *Boston University Law Review* 91 (2011): 1435.

5 Mark Zuckerberg, "Building a Global Community," Facebook, February 16, 2017, accessed August 23, 2018, https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634.

6 Will Oremus, "Speed Kills," *Slate,* April 2, 2018, accessed August 23, 2018, https://slate.com/technology/2018/04/silicon-valleys-obsession-with-rapid-growth-literally-killing-us.html; Senator Ed Markey (D-MA) (@SenMarkey), "Death from bullying cannot be the cost of doing business. Terrorist attacks cannot be the cost of doing business. It is @facebook's moral obligation to maintain the integrity and safety of their platform. When they fail to do so, Congress must act," Twitter, March 30, 2018, accessed August 23, 2018, https://twitter.com/SenMarkey/status/979760217025077249.

7 S. 1989, 115th Cong. (2017).

8 H.R. 1865, 115th Cong. (2017–2018).

9 Nitasha Tiku, "How a Controversial New Sex-Trafficking Law Will Change the Web," *Wired,* March 22, 2018, accessed August 23, 2018, https://www.wired.com/story/how-a-controversial-new-sex-trafficking-law-will-change-the-web.

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.

11 All Actions: H.R. 1865, Congress.gov, accessed August 23, 2018, https://www.congress.gov/bill/115th-congress/house-bill/1865/all-actions?overview=closed&q=%7B%22roll-call-vote%22%3A%22all%22%7D.

12 Emily Stewart, "The Next Big Battle over Internet Freedom is Here," *Vox,* April 23, 2018, accessed August 23, 2018, https://www.vox.com/policy-and-politics/2018/4/23/17237640/fosta-sesta-section-230-internet-freedom.

13 Internet Association, "Statement in Support of the Bipartisan Compromise to The Stop Enabling Sex Traffickers Act," *IA News,* November 3, 2017, accessed August 23, 2018, https://internetassociation.org/statement-in-support-of-the-bipartisan-compromise-to-stop-enabling-sex-trafficking-act-sesta.

14 Alina Selyukh, "Section 230: A Key Legal Shield for Facebook, Google is about to Change," NPR, March 21, 2018, accessed August 23, 2018, http://www.wbur.org/npr/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change.

15 Jonathan Zittrain, "CDA 230 Then and Now: Does Intermediary Immunity Keep the Rest of Us Healthy?" *The Recorder,* November 10, 2017, accessed August 23, 2018, https://www.law.com/therecorder/sites/therecorder/2017/11/10/cda-230-then-and-now-does-intermediary-immunity-keep-the-rest-of-us-healthy.

16 "Transcript of Mark Zuckerberg's Senate Hearing," *Washington Post,* April 10, 2018, accessed August 23, 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/?utm_term=.63fc016546dc.

17 "Mark Zuckerberg in His Own Words: The CNN Interview," CNN, March 21, 2018, accessed August 23, 2018, http://money.cnn.com/2018/03/21/technology/mark-zuckerberg-cnn-interview-transcript/index.html.

18 Citron, "Extremist Speech."

19 Richard Allan, "Who Should Decide What Is Hate Speech in an Online Global Community?" *Hard Questions* (blog), accessed August 23, 2018, https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech.

20 Selyukh, "Section 230."

21 Danielle Keats Citron, "Cyber Civil Rights," *Boston University Law Review* 89 (2009): 61; Mark Lemley, "Rationalizing Internet Safe Harbors," *Journal of Telecommunications and High Technology Law* (2007): 101, accessed August 23, 2018, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=979836; Douglas Gary Lichtman and Eric Posner, "Holding Internet Service Providers Accountable," John M. Olin Program in Law and Economics Working Paper No. 217, 2004, accessed August 23, 2018, https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1235&context=law_and_economics.

22 Olivier Sylvain, "Discriminatory Designs on User Data," Knight First Amendment Institute at Columbia University, April 2018, accessed August 23, 2018, https://knightcolumbia.org/content/discriminatory-designs-user-data; Olivier Sylvain, "Intermediary Design Duties," *Connecticut Law Review* 50 (2018): 203.

23 Citron, "Cyber Civil Rights."

24 Rebecca Tushnet, "Power without Responsibility: Intermediaries and the First Amendment," *George Washington Law Review* 76 (2008): 101, 117.

25 Tim Hwang, "Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment," MIT Media Laboratory, December 17, 2017, accessed August 23, 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442.

26 Steven Overly and Ashley Gold, "How Tech Lost on the Sex Trafficking Bill," *Politico,* March 22, 2018, accessed August 23, 2018, https://www.politico.com/story/2018/03/22/how-tech-lost-on-the-sex-trafficking-bill-423364.

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

28 H.R. 1865, 115th Cong. (2017–2018).

29 18 U.S.C. § 2421 (2015).

30 Nicholas Kristof, "Google and Sex Traffickers Like Backpage.com," *New York Times,* September 7, 2017, accessed August 23, 2018, https://www.nytimes.com/2017/09/07/opinion/google-backpagecom-sex-traffickers.html.

31 Elliot Harmon, "Internet Censorship Bill Would Spell Disaster for Speech and Innovation," Electronic Frontier Foundation, August 2, 2017, accessed August 23, 2018, https://www.eff.org/deeplinks/2017/08/internet-censorship-bill-would-spell-disaster-speech-and-innovation.

32 "Why Do Sex Workers Need Online Spaces?" Survivors Against SESTA, March 2018, accessed August 23, 2018, https://survivorsagainstsesta.files.wordpress.com/2018/03/onlinespaces_impact-003.pdf.

33 Tom Jackman, "Internet Companies Drop Opposition to Bill Targeting Online Sex Trafficking," *Washington Post,* November 7, 2017, accessed August 23, 2018, https://www.washingtonpost.com/news/true-crime/wp/2017/11/07/internet-companies-drop-opposition-to-bill-targeting-online-sex-trafficking/?utm_term=.ddd78c47eff3.

34 *FOSTA,* Craigslist, March 23, 2018, accessed August 23, 2018, https://www.craigslist.org/about/FOSTA.

35 "New Addition to Site-wide Rules Regarding the Use of Reddit to Conduct Transactions," Reddit, March 23, 2018, accessed August 23, 2018, https://www.reddit.com/r/announcements/comments/863xcj/new_addition_to_sitewide_rules_regarding_the_use.

36 "US Sex Workers Affected #AfterFosta—Next Steps," Coyote RI, March 23, 2018, accessed August 23, 2018, http://coyoteri.org/wp/us-sex-workers-affected-2-days-after-fosta-passes-the-senate; Liara Roux, "Post-SESTA/FOSTA Self-Censoring for Twitter, Reddit, and other Social Media," *Tits and Sass,* March 23, 2018, accessed August 23, 2018, http://titsandsass.com/post-sesta-fosta-self-censoring-for-twitter-reddit-and-other-social-media.

37 Sugar Daddie Online, accessed August 23, 2018, https://sugardaddie.com.

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://switter.at/about; Megan Farokhmanesh, "Amid FOSTA Crackdown, Sex

Workers Find Refuge on Mastodon," *The Verge,* April 11, 2018, accessed August 23, 2018, https://www.theverge.com/2018/4/11/17188772/trump-sesta-fosta-bill-switter-sex-workers-mastodon.

39 According to Assembly Four, which runs Mastodon, Switter 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, Switter remains live. See Assembly Four, "Cloudflare and FOSTA/SESTA," accessed August 23, 2018, https://assemblyfour.com/switter/cloudflare-fosta-sesta.

40 Megan Farokhmanesh, "Switter, One of the Last Online Spaces Friendly to Sex Workers, Was Just Banned by Its Network," *The Verge,* April 19, 2018, accessed August 23, 2018, https://www.theverge.com/2018/4/19/17256370/switter-cloudflare-sex-workers-banned.

41 Representative Bob Goodlatte (@RepGoodlatte), Twitter, March 23, 2018, accessed August 23, 2018, https://twitter.com/RepGoodlatte/status/977237099089887232.

42 Elliot Harmon, "How Congress Censored the Internet," Electronic Frontier Foundation, March 21, 2018, accessed August 23, 2018, https://www.eff.org/deeplinks/2018/03/how-congress-censored-internet.

43 Kate D'Adamo (@KDAdamo), "Well. That's a thing," Twitter, March 25, 2018, accessed August 23, 2018, https://twitter.com/KateDAdamo/status/977919166702288896; Private Internet Access (@buyvpnservice), "Digital Copies of Our #HaveAVoice Ads in the New York Times Today," Twitter, March 25, 2018, accessed August 23, 2018, https://twitter.com/buyvpnservice/status/977910770859200512.

44 Senator Ron Wyden (@RonWyden), Twitter, March 21, 2018, accessed August 23, 2018, https://twitter.com/RonWyden/status/976528191044182020.

45 Mike Godwin, "Why Internet Advocates Are Against the Anti-Sex Trafficking Bill," *Slate,* March 14, 2018, accessed August 23, 2018, https://slate.com/technology/2018/03/the-antisex-trafficking-bill-sesta-fosta-will-hurt-the-internet.html; Daphne Keller, "Toward a Clearer Conversation about Platform Liability," Knight First Amendment Institute at Columbia University, April 6, 2018, accessed August 23, 2018, https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability.

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.

47 Letter from Stephen E. Boyd, assistant attorney general, to Robert W. Goodlatte, chairman of House Committee on the Judiciary, February 27, 2018, accessed August 23, 2018, https://assets.documentcloud.org/documents/4390361/Views-Ltr-Re-H-R-1865-Allow-States-and-Victims.pdf.

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

49 Hwang, "Dealing With Disinformation."

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

51 Stacey Dogan, "Principled Standards vs. Boundless Discretion: A Tale of Two Approaches to Intermediary Trademark Liability Online," *Columbia Journal of Law & the Arts* 37 (2014): 502, 508.

52 Ibid.

53 Citron and Wittes, "The Internet Will Not Break."

54 Ibid.

55 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 dignatory 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.

56 Citron, "Extremist Speech."

57 Paul Ohm, "Breaking Felten's Third Law: How Not to Fix the Internet," *Denver Law Review Online,* February 22, 2010, accessed August 23, 2018, http://www.denverlawreview.org/how-to-regulate/2010/2/22/breaking-feltens-third-law-how-not-to-fix-the-internet.html.

58 Ed Felten, "A Free Internet, If We Can Keep It," Center for Information Technology Policy, Princeton University, January 28, 2010, accessed August 23, 2018, https://freedom-to-tinker.com/2010/01/28/free-internet-if-we-can-keep-it.

59 Stratton Oakmont v. Prodigy Servs. Co., 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

60 Jeffrey Koseff, *The 26 Words that Created the Internet* (Ithaca, NY: Cornell University Press, forthcoming).

61 Jamie Condliffe, "Facebook and Google May Be Fighting Terrorist Videos with Algorithms," *MIT Technology Review,* June 27, 2016, accessed August 23, 2018, https://www.technologyreview.com/s/601778/facebook-and-google-may-be-fighting-terrorist-videos-with-algorithms.

62 Citron, "Extremist Speech."

63 Ibid.

64 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—notably Senator Kamala Harris and Representative Jackie Speier (both California Democrats)—have proposed federal law to criminalize the practice.

65 Antigone Davis, "The Facts: Non-Consensual Intimate Image Pilot," Facebook Newsroom, November 9, 2017, accessed August 23, 2018, https://newsroom.fb.com/news/h/non-consensual-intimate-image-pilot-the-facts.

66 "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).

67 Casey Newton, "Facebook, Microsoft, Twitter, and YouTube are Creating a Database of 'Terrorist Content,'" *The Verge,* December 5, 2016, accessed August 23, 2018, https://www.theverge.com/2016/12/5/13849570/facebook-microsoft-twitter-google-terrorist-content-database.

68 Ibid. Each company will apply its own definition. See "Partnering to Help Curb Spread of Online Terrorist Content," Facebook, December 5, 2016, accessed August 23, 2018, https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content.

69 See "Violent or Graphic Content," YouTube Help, accessed August 23, 2018, https://support.google.com/youtube/answer/2802008?hl=en&ref_topic=2803176. Also see "The Twitter Rules," Twitter Help Center, accessed August 23, 2018, https://help.twitter.com/en/rules-and-policies/twitter-rules.

70 Malachy Browne, "YouTube Removes Videos Showing Atrocities in Syria," *New York Times,* August 22, 2017, accessed August 23, 2018, https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.

71 Liat Clark, "UK Gov Wants 'Unsavoury' Web Content Censored," *Wired UK,* March 15, 2014, accessed August 23, 2018, http://www.wired.co.uk/article/government-web-censorship.

72 See Geoffrey R. Stone, *Perilous Times: Free Speech in Wartime: From the Sedition Act of 1798 to the War on Terrorism* (New York: Norton, 2004), 555–56.

73 Natasha Duarte, Emma Llanso, and Anna Loup, "Mixed Messages? The Limits of Automated Social Media Content Analysis," Center for Democracy & Technology, November 2017, accessed August 23, 2018, https://cdt.org/files/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

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).

75 Citron, *Hate Crimes in Cyberspace.* Kate Klonick takes up this model in her groundbreaking work on the speech rules and practices of content platforms, which she calls the "New Speech Governors." Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review* 131 (2018): 1598, 1668–69.

76 Zuckerberg, "Building Global Community."

77 See "About ONO," Organization of News Ombudsmen & Standards Editors, accessed August 23, 2018, http://newsombudsmen.org/about-ono.

78 Ibid.

79 Ibid.

80 See Citron, "Extremist Speech."

81 N.Y. Times Co. v. Sullivan, 376 U.S. 254, 270 (1964).

82 Megan Specia and Paul Mozur, "A War of Words Puts Facebook at the Center of Myanmar's Rohingya Crisis," *New York Times,* October 27, 2017, accessed August 23, 2018, https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html.

83 See James Grimmelman, "To Err Is Platform," Knight First Amendment Institute at Columbia University, April 2018, accessed August 23, 2018, https://knightcolumbia.org/content/err-platform.

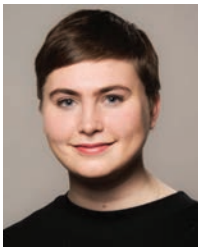
84 Danielle Keats Citron and Neil M. Richards, "Four Principles for Digital Expression," *Washington University Law Review* 95 (2018): 1353.

85 Ibid.

86 Ibid.

The preferred citation for this publication is Danielle Citron and Quinta Jurecic, "Platform Justice: Content Moderation at an Inflection Point," Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1811 (September 5, 2018), available at https://www.lawfareblog.com/platform-justice-content-moderation-inflection-point.

Author bios



QUINTA JURECIC

Quinta Jurecic is the managing editor of *Lawfare*. She previously served as an editorial writer for the *Washington Post* and as *Lawfare's* associate editor.

DANIELLE CITRON

Danielle Citron is the Morton & Sophia Macht Professor of Law at the University of Maryland Carey School of Law, where she writes and teaches about information privacy, free speech, and civil rights. She is the author of *Hate Crimes in Cyberspace* (Harvard University Press) and more than twenty-five law review articles. She is an affiliate fellow at the Yale Information Society Project and the Stanford Center for Internet and Society.

Working Group on National Security, Technology, and Law

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.

Hoover Institution, Stanford University
434 Galvez Mall
Stanford, CA 94305-6003
650-723-1754

Hoover Institution in Washington
The Johnson Center
1399 New York Avenue NW, Suite 500
Washington, DC 20005
202-760-3200

HOOVER INSTITUTION