A HOOVER INSTITUTION ESSAY

Internet Platforms

OBSERVATIONS ON SPEECH, DANGER, AND MONEY

DAPHNE KELLER

Aegis Series Paper No. 1807

I. Introduction

Public demands for internet platforms to intervene more aggressively in online content are steadily mounting. Calls for companies like YouTube and Facebook to fight problems ranging from "fake news" to virulent misogyny to online radicalization seem to make daily headlines. Some of the most emphatic and politically ascendant messages concern countering violent extremism (CVE).[1] As British prime minister Theresa May put it, "Industry needs to go further and faster" in removing prohibited content,[2] including by developing automated filters to detect and suppress it automatically.

The public push for more content removal coincides with growing suspicion that platforms are, in fact, taking down too much. Speakers across the political spectrum charge that platforms silence their speech for the wrong reasons. Over seventy social justice organizations wrote to Facebook in 2017, saying that the platform enforces its rules unfairly and removes more speech from minority speakers.[3] Conservative video educator Dennis Prager says that YouTube suppressed his videos in order to "restrict non-left political thought,"[4] and pro-Trump video bloggers Diamond and Silk told the House Judiciary Committee that Facebook had censored them.[5] Prager is suing YouTube and demanding reinstatement. As he points out, speech that disappears from the most important platforms loses much of its power because many potential listeners simply don't find it. In extreme cases—as with Cloudflare's banishment of the Daily Stormer[6]—disfavored voices may disappear from the internet completely.

One thing these opposing public pressures tell us is that platforms really are making both kinds of mistakes. By almost anyone's standards, they are sometimes removing too much speech, and sometimes too little. Well-publicized hiring sprees[7] on content moderation teams might help with this problem. Increased public transparency[8] into those teams' rules and processes almost certainly will as well.

The other thing the conflicting public sentiments about platforms and speech illuminate, though, is a set of fundamental problems with delegating complex decisions about free expression and the law to private companies. As a society, we are far from consensus about legal or social speech rules. There are still enough novel and disputed questions surrounding even long-standing legal doctrines, like copyright and defamation, to keep law firms in

business. If democratic processes and court rulings leave us with such unclear guidance, we cannot reasonably expect private platforms to do much better. However they interpret the law, and whatever other ethical rules they set, the outcome will be wrong by many people's standards.

The study of intermediary liability tells us more about what to expect when we delegate interpretation and enforcement of speech laws to private companies. Intermediary liability laws establish platforms' legal responsibilities for content posted by users. Twenty years of experience with these laws in the United States and elsewhere tells us that when platforms face legal risk for user speech, they routinely err on the side of caution and take it down. This pattern of over-removal becomes more consequential as private platforms increasingly constitute the "public square" for important speech. Intermediary liability law also tells us something about the kinds of rules that can help avoid over-removal.

In this essay, I will describe the lessons learned from existing intermediary liability laws and the foreseeable downsides of requiring platforms to go "further and faster" in policing internet users' speech. Policy makers must decide if these costs are justified by the benefits of a more regulated and safer internet.

The first cost of strict platform removal obligations is to internet users' free expression rights. We should expect over-removal to be increasingly common under laws that ratchet up platforms' incentives to err on the side of taking things down. Germany's new NetzDG law, for example, threatens platforms with fines of up to €50 million for failure to remove "obviously" unlawful content within twenty-four hours' notice.[9] This has already led to embarrassing mistakes. Twitter suspended a German satirical magazine for mocking a politician,[10] and Facebook took down a photo of a bikini top artfully draped over a double speed bump sign.[11] We cannot know what other unnecessary deletions have passed unnoticed.

Such a burden on individual speech raises constitutional questions. Does the First Amendment limit laws that incentivize private platforms to silence legal speech? If so, what obligations can the government impose on platforms before encountering a constitutional barrier? In this paper's first analytical section, I discuss precedent on this question. Courts in the United States have spent little time considering it because our speech-protective intermediary liability statutes largely render constitutional analysis unnecessary. But Supreme Court cases about "analog intermediaries" like bookstores provide important guidance. In addition, courts outside the United States have wrestled with these questions in the internet context, often drawing on US precedent. Based on the US cases and international experience, I will suggest four considerations that would make any new US intermediary liability laws more likely—or less—to survive constitutional review.

The second cost I will discuss is to security. Online content removal is only one of many tools experts have identified for fighting terrorism. Singular focus on the internet, and overreliance on content purges as tools against real-world violence, may miss out on or even undermine other interventions and policing efforts.

The cost-benefit analysis behind CVE campaigns holds that we must accept certain downsides because the upside—preventing terrorist attacks—is so crucial. I will argue that the upsides of these campaigns are unclear at best, and their downsides are significant. Over-removal drives extremists into echo chambers in darker corners of the internet, chills important public conversations, and may silence moderate voices. It also builds mistrust and anger among entire communities. Platforms straining to go "faster and further" in taking down Islamist extremist content in particular will systematically and unfairly burden innocent internet users who happened to be speaking Arabic, discussing Middle Eastern politics, or talking about Islam. Such policies add fuel to existing frustrations with governments that enforce these policies, or platforms that appear to act as state proxies. Lawmakers engaged in serious calculations about ways to counter real-world violence—not just online speech—need to factor in these unintended consequences if they are to set wise policies.

The third cost is to the economy. There is a reason why the technology-driven economic boom of recent decades happened in the United States. As publications with titles like "How Law Made Silicon Valley" point out,[12] our platform liability laws had a lot to do with it. These laws also affect the economic health of ordinary businesses that find customers through internet platforms—which, in the age of Yelp, Grubhub, and eBay, could be almost any business. Small commercial operations are especially vulnerable when intermediary liability laws encourage over-removal, because unscrupulous rivals routinely misuse notice and takedown to target their competitors.

The point here is not that platforms should never take content down. They already do take things down, and this is often very much in the public interest. The upsides of content removal are not hard to see. They include everything from reducing movie piracy to eliminating horrific material like child pornography. The point is that these upsides come with well-known downsides. Neither platforms nor lawmakers can throw a switch and halt the flow of particular kinds of speech or content. Artificial intelligence and technical filters cannot do it either—not without substantial collateral damage. Delegating speech law enforcement to private platforms has costs. Lawmakers need to understand them, and plan accordingly, when deciding when and how the law tells platforms to take action.

II. What Platforms Do with Prohibited Content

This section will describe platforms' real-world content removal operations and will then examine more closely the laws governing those operations. Platform behavior is in part a

product of national intermediary liability laws. These laws generally require platforms to remove content in at least some circumstances, with obligations usually triggered when the platform learns about illegal material. In the United States, for example, a platform that does not remove pirated videos upon becoming aware of them could be liable for copyright infringement.[13] Platforms operate notice and takedown programs as a way both to comply with such laws and to enforce their own discretionary content policies.

Most well-known platforms take down considerably more content than the law requires, enforcing their own community guidelines or terms of service. Some do so because of their core business goals—LinkedIn will take down material that's irrelevant to users' professional profiles, for example.[14] More broadly, most prohibit things like nudity, bullying, or racial slurs[15]—speech that is legal in the United States but widely considered offensive. A social media service that does not successfully prune such material from users' day-to-day experience would risk losing both users and advertisers. Advertisers' power to drive platform content policies was well illustrated in 2017, when media reports about hateful or extremist content on YouTube led advertisers to threaten withdrawal.[16] YouTube changed its policies almost overnight,[17] only to face a different public backlash[18]—this time from independent media creators economically devastated by the policy changes, which they called the "adpocalypse."

In recent years, concerns have mounted about the power of online platforms to enforce their own biases through discretionary content policies. Legal analysis of that issue would easily consume another thirty pages. But as a social and policy matter, the outrage from those evicted, seemingly unfairly, from the new "public square" is an important consideration as governments push platforms to take on ever greater roles in regulating online speech.

A. The Big Picture

1. Removal and Over-removal Speech available on the internet is, in the Supreme Court's words, "as diverse as human thought."[19] Realism about humans and their thoughts tells us that this will not always be a good thing. The internet enables everything from the viral generosity of the ALS Ice Bucket Challenge[20] to the dangerous stupidity of the Tide Pod Challenge;[21] and from outpourings of solidarity like the #MeToo movement to brutal invasions of privacy like revenge pornography. Offensive, dangerous, and illegal speech has been an issue since day one.

The same range of legitimate and troll-like behavior shows up in content removal demands sent to intermediaries. Abusive requests, including efforts to silence commercial and ideological rivals, are common. Notice and takedown systems for popular platforms routinely receive these requests and employ ever-growing teams of workers to sort them from legitimate complaints—with only partial success.

Legally mandated notice and takedown systems unavoidably stack the deck in favor of accusers and against online speakers. In the face of potential liability, a platform's easiest and cheapest course is to take accusations at face value. Speedy compliance avoids both legal risk and the cost of hiring a lawyer—or even a minimum-wage employee—to assess whether a claim is valid. And while potential claimants can afford to pull their punches in legal gray areas—to tolerate satire, edgy humor, or gloves-off disagreement online—platforms cannot. Once they know of any legal risk, their incentive is to err on the side of removal.

Improper removal demands take many forms. Some—like HBO's for an autistic teenager's artwork that used the phrase "winter is coming"[22]—are presumably mistakes. Others clearly are not. One recent requester impersonated a journalist to suppress coverage of a federal fraud investigation;[23] another tried to hide web pages listing his name on state sex-offender registries.[24] A former operator of a revenge porn website, who is now running for public office, has used both copyright and privacy claims to hide evidence of his past,[25] including video interviews in which he discusses his former business. Eugene Volokh has uncovered and reported on a creative new tactic: falsifying court orders in hopes of tricking Google into removing search results.[26]

Data and detailed information about platforms' removal practices are hard to come by, but over-removal is clearly widespread.[27] The standout work in the field, by researchers at UC Berkeley and Columbia University, documents significant problems under the US Digital Millennium Copyright Act (DMCA)—even though that law has provisions intended to prevent over-removal.[28] The study also reveals a marked divergence between the small handful of major platforms and the remaining, overwhelming majority of American internet intermediaries.[29]

Smaller and lower-profile platforms in the study received removal requests numbering in the tens, hundreds, or thousands per year. They relied on employees to review and respond to them. Some reported pushing back on illegitimate claims. Others said they simply honored 100 percent of requests in order to avoid legal exposure. Most took users' content down "even when they [were] uncertain about the strength of the underlying claim."[30]

For today's handful of mega-platforms, the picture is very different. The increased scale of notice and takedown operations from a decade ago is remarkable: Google web search went from getting a few hundred DMCA notices per year in 2006[31] to two million each day in 2016.[32] The Berkeley–Columbia report found debatable legal claims in 28 percent of web search removal requests, and clear error—removal requests targeting works the rights holders did not actually own—in 4.2 percent.[33] This seemingly small percentage would affect 4.5 million individual web pages.[34]

2. Automation and Filters To handle the expanded volume of takedowns, both major notifiers and major platforms rely increasingly on automation rather than human review.[35]

Automation makes more effective rights enforcement possible at scale, but it also introduces new kinds of errors. For example, representatives of the musician Usher asked Google to remove a movie version of *The Fall of the House of Usher*, presumably because they relied on keyword searches to generate removal notices.[36] For the same likely reason, requests by film or TV rights holders regularly include the works' IMDB or Wikipedia pages. Google's review process catches some of these mistakes but by no means all.

Automation and algorithmic content removal has expanded on the platform side as well—at least for the most successful platforms. Industry groups and individual platforms have developed filtering or monitoring technologies to identify and remove unwanted content.[37] These technologies feature prominently in current debates about platform responsibility. In his April 2018 Congressional testimony, Facebook CEO Mark Zuckerberg said that filters aided by Artificial Intelligence would one day block harmful content ranging from fake news to terrorist propaganda. But the technology behind content filters is often poorly understood.[38]

Some of the earliest internet content filters were simple text-matching programs used to filter spam or pornography. These were somewhat effective, but they also generated notorious errors by blocking innocent content.[39] More sophisticated language filters today use natural language processing (NLP) and sentiment analysis, which attempt to identify objectionable text without relying on a blocklist of forbidden phrases.[40] Recent research has shown accuracy rates in the 70–80 percent range for commercially available NLP tools, meaning that one time in four or five they take down the wrong thing.[41] Unsurprisingly, these filters miss sarcasm and jokes, and they perform poorly in languages not spoken by their developers.[42]

Somewhat more sophisticated filters rely on hashes (digital thumbprints) to identify images or videos. Platforms first widely used filters of this sort to combat child pornography.[43] Separate efforts across different companies gradually converged, and the National Center for Missing and Exploited Children now administers a database of more advanced PhotoDNA hashes.[44] Using these, a platform can search for duplicates and automatically remove them or flag them for human evaluation.

For copyright, many UGC hosts use hashing tools to find exact duplicates of video or audio files identified as infringing. YouTube and, later, Facebook also developed far more sophisticated technology capable of identifying even modified copies. YouTube's Content ID[45] cost the company a reported $60 million to develop and has become the cornerstone of both enforcement efforts and commercial deals with rights holders.[46] Human or technical errors in the Content ID system, despite its industry-leading technology, routinely lead to the removal of noninfringing content—like when Ariana Grande's benefit concert disappeared midstream from the artist's own YouTube account.[47]

Most recently, platforms have adopted filtering technology as part of CVE efforts. In 2016, four of the biggest platforms announced a shared database of hashes for filtering

out "extremist content," as defined under their individual terms of service.[48] In 2017, they launched the Global Internet Forum to Counter Terrorism, which will share this technology with smaller companies.[49] According to Senate testimony in 2018, twelve platforms now use the database,[50] effectively shrinking the portion of the internet where information barred by the filters may be shared. Facebook and YouTube representatives report that 98 percent or more of CVE removals are now instigated by automated tools, rather than users.[51] Both companies also use human reviewers to check the machines' decisions. Facebook currently employs over seventy-five hundred people in content moderation and review roles;[52] YouTube expects to exceed ten thousand in 2018.[53]

Platforms that rely on filters have political reasons to trumpet the technologies' capabilities, as a means to stave off regulatory threats. They also have commercial reasons, since the same tools may underlie anything from agreements with rights holders to ad-targeting to new artificial intelligence–based services. But scientists, including Princeton's Nick Feamster, point out that the technology is not all that one might hope,[54] and a recent report suggests that commercially available filters may greatly overstate their efficacy.[55]

Neither long-standing filters nor new ones labeled as artificial intelligence (AI), machine learning, or other whizbang technology can look at a new video or web page and say whether it violates the law. Some can, clumsily, identify things like patterns or bare flesh, and companies say that AI is increasingly effective at identifying potential terrorist content.[56] But no reputable experts suggest that filters are good enough to be put in charge of deciding what is illegal in the first place. What filters like Content ID or the terrorism hash system do is find duplicates of specific material that a human previously flagged. Even in that narrow technical task, filters generate false positives (flagging the wrong thing) and false negatives (failing to flag the right thing). A report by Feamster and Evan Engstrom lists numerous technical bases for filter error.[57] Platforms that cannot make investments like YouTube's $60 million for Content ID will, if forced to build filters, presumably be forced to tolerate high rates of false positives in order to avoid liability for false negatives.

More fundamentally, filters are no substitute for human judgment, particularly for legal questions. To an algorithm, an album cover image used to promote illegal downloads is indistinguishable from the same image in a concert review. An ISIS video looks the same, whether used in recruiting or in news reporting. This is presumably why YouTube's filters took down the channel of a UK-based human rights organization documenting war crimes in Syria.[58] This context-blindness may be acceptable in policing child pornography, which has no legal context and is not considered protected speech under the law. For complex speech, though, it is a real source of problems.

Human review to correct for the limitations of filtering technology, as discussed by platform witnesses in the Senate hearing, is better than relying on machines entirely. But it cannot ultimately solve the over-removal problem, which is already rampant in existing,

human-operated content moderation. And once human errors feed into a filter's algorithm, they will be amplified, turning a one-time mistake into an every-time mistake and making it literally impossible for users to share certain images or words.

3. Global Pressures Another issue unique to the major platforms is the complexity and geopolitical scope of content removal pressures. Platforms with international presence must comply not only with US laws but also with those of other countries. To some extent, they may do so by offering different versions of their product in different countries. A French Twitter user, for example, will typically see a French version of the product, with tweets or accounts excised based on French law.[59] But product differentiation of this sort can be inconvenient and costly. An easier path is to find an acceptable amalgam—or the lowest common denominator—of national laws and comply with these as a matter of "voluntary" policy.[60]

The reach of many countries' national laws has increased in recent years. Courts and regulators have ordered platforms to comply globally with laws ranging from French "Right to Be Forgotten" regulations to Canadian trade secret rules to Austrian hate speech law.[61] Many of these laws or orders would not be enforceable in the United States under the SPEECH Act,[62] which was unanimously passed by Congress in 2010 to limit enforcement of certain foreign judgments if they violate US intermediary liability law or the First Amendment. But US enforceability may not matter at the end of the day. Courts in important foreign markets have significant enforcement power of their own. Noncompliant American companies may find their assets seized, their employees arrested (as has happened to platforms in Brazil[63] and India[64]), or markets disrupted by service blockages (as has happened in China,[65] Russia,[66] Turkey,[67] and Malaysia,[68] among other places). That makes local legal pressures, including demands for global compliance, very real.

Of course, not all pressure comes from laws. Bad press and strained relationships with advertisers have their own effect. The mere threat of new legal measures may be as important as laws that actually get passed. In 2016, for example, four major platforms reached an agreement with the European Commission to voluntarily enforce EU law–based hate speech policies[69] in their global terms of service.[70] The move was widely perceived as a compromise to stave off legislation.[71] Other "self-regulatory" efforts may also have roots in less conspicuous government pressure. Critics charge that such backroom negotiations with platforms allow governments to avoid democratic processes and accountability. Some also argue that governments may violate free expression rights by strong-arming platforms to remove offensive but legal speech.[72]

Whatever the mechanism, the upshot is that pressure coming from just one or a few countries can have global impact. In the past, the United States very much played this role: European critics have long charged that the internet generally, and American tech companies in particular, "export" First Amendment values by disseminating speech that is legal here but illegal elsewhere. Now, increasingly, the EU finds itself in a position to export

its own preferred balance between speech and other values, such as privacy—whether through direct legal enforcement, as in the recently enacted General Data Protection Regulation, or soft power. Pressure in Europe and elsewhere also indirectly shapes outcomes in other countries by changing underlying internet technologies. Once platforms build out technical capacity to do things like filter user speech, governments around the world will want to use them too, for anything from Saudi blasphemy laws to Russian antigay laws to Thai laws against insulting the king.[73]

B. What the Law Says

National intermediary liability laws vary but share a basic architecture. They typically immunize platforms for legal claims based on the content of user-generated communications, and they often pair this immunity with conditions or obligations. To preserve immunity, a platform must maintain a sufficiently hands-off relationship to user speech. A platform that creates its own content, or collaborates with a user to do so, is likely to be legally responsible for it under any country's law.

National intermediary liability laws often reflect lawmakers' priorities, particularly with respect to free speech and technological growth. At a high level, though, the purpose of these laws is to take bad content down from the internet and keep good content up. Where doctrinal differences arise is usually in the mechanism for doing so. US law, uniquely, relies on pragmatic incentives rather than legal mandates for many claims, as will be discussed below. Most other countries require platforms to take down illegal content once they "know" about it, though laws vary significantly in what kind of knowledge triggers liability. In countries with stronger free expression laws, a platform may only have to take content down once it knows that a court has held it illegal. In other countries, a mere allegation may suffice.

Intermediary liability laws typically start from the recognition that platforms cannot possibly spot every instance of defamation, copyright infringement, or hate speech in the torrent of human communication traversing their servers. Laws that hold them liable and expose them to meaningful damages would doom many online businesses and hinder innovation. This pragmatic concern is likely the reason why so many countries, despite divergent legal cultures, reject strict liability for intermediaries. China, for example, generally does not hold intermediaries liable for content they do not know about, although a new Chinese Copyright Monitoring Center now scans the internet for infringement.[74]

1. US Law Intermediary liability in the United States is mostly governed by three laws: the Communications Decency Act (CDA) for most civil claims, the DMCA for copyright claims, and Title 18 of the US Code for criminal claims.

a. The Communications Decency Act The first law, Communications Decency Act Section 230 (CDA 230), immunizes platforms from traditional speech torts, such as

defamation, and other civil claims that effectively treat a platform as the publisher of a user's speech.[75] It also bars most state, but not federal, criminal claims.[76] The statute, which passed as part of the 1996 Telecommunications Act overhaul, states at some length Congress's observation that the internet has "flourished, to the benefit of all Americans, with a minimum of government regulation," and its intention to uphold this legal framework.[77]

CDA 230 provides a defense for almost any claim that would hold a platform responsible for its users' speech. Platforms lose the immunity if they help create or develop that speech. The statute does not apply, or provide any defense, for claims under federal criminal law, intellectual property law, certain laws involving prostitution or trafficking, or the Electronic Communications Privacy Act (ECPA).[78] And importantly, the CDA not only protects platforms from liability for user content they leave online but also protects them when they take content down.[79] This "Good Samaritan" rule frees platforms to experiment with proactive measures, like the evolving community guidelines and hashing systems described above. Without it, American platforms would have reason to fear—as platforms commonly do in other countries—that by trying to moderate speech on their platforms, they will be deemed insufficiently neutral and will face liability.

Congress's decision to rely on immunities and incentives, instead of legal mandates, reflects a pragmatic calculation: that companies would, on the whole, try harder to weed out bad content if those efforts didn't expose them to legal risk. A pair of recent cases at the time of the legislation prompted their concern. In one case, a platform undertook to moderate inappropriate user content. As a result, the court treated it as a publisher that could be held liable for defamation, whether or not it knew of a specific post.[80] In the other, a platform that made no such efforts escaped liability.[81] With the Good Samaritan rule, Congress removed these perverse legal incentives.

CDA 230 also embodies a policy judgment of sorts: to prioritize economic development and free expression on the internet at the cost of imperfect enforcement. It leaves plaintiffs with remedies against the people who actually create unlawful content. But it leaves little recourse for those harmed by anonymous online speech, and it eliminates the potentially more effective (and remunerative) path of suing intermediaries instead of speakers.

CDA 230 became controversial in recent years because of legal victories by Backpage .com, a site that hosted prostitution ads. After the First Circuit upheld the company's CDA 230 defense against claims that it was complicit in sex trafficking, a Senate investigation revealed that Backpage employees had actually helped traffickers create ads.[82] Congress responded in 2018 by amending CDA 230 for the first time in two decades. The Allow States and Victims to Fight Online Sex Trafficking Act, commonly known as FOSTA, curtails platform immunities in sex-trafficking cases.[83]

b. The Digital Millennium Copyright Act The second major US law is the Digital Millennium Copyright Act.[84] The DMCA is often compared to notice and takedown laws in Europe and other countries, but it has some very important differences. It was closely negotiated by commercial interests on all sides and by civil liberties advocates.[85] As a result of their efforts, the final bill included groundbreaking protections for online speech.

One essential protection for both speech and privacy is DMCA 512(m), which says that platforms do not have to monitor their users' speech in order to avoid liability. The opposite rule—making every user utterance a source of legal risk, subject to review by anxious platform lawyers—would have yielded a very different internet today. Among other things, that internet would presumably include few or no American companies offering open-access internet platforms for unmoderated speech. This part of the DMCA has come under political pressure in recent years, with rights holders arguing that content-filtering technologies have evolved to the point that law should mandate their use.[86] The US Copyright Office undertook a public study on this issue beginning in 2015.[87]

The DMCA also protects speech using procedural rules for the platforms that "adjudicate" disputes between users. These function roughly like civil procedure in a courtroom, to increase fairness between a plaintiff and defendant. The DMCA's procedures include a "counter-notice" process, allowing people accused of infringement to defend themselves against mistaken or malicious claims.[88] The DMCA also provides penalties for bad-faith accusations.[89]

An important but unmeasurable impact of the DMCA's procedures lies in deterring false claims. In the early 2000s, researchers experimented by posting well-known public domain essays—including John Stuart Mill's 1869 "On Liberty"—online, then submitting false copyright removal requests. Most European companies took the essays down without question. The sole American one did not, citing the DMCA's required penalty-of-perjury statement.[90]

As the Ninth Circuit has noted, the DMCA's procedural protections have First Amendment ramifications:

> Accusations of alleged infringement have drastic consequences: A user could have content removed, or may have his access terminated entirely. If the content infringes, justice has been done. But if it does not, speech protected under the First Amendment could be removed. We therefore do not require a service provider to start potentially invasive proceedings if the complainant is unwilling to state under penalty of perjury that he is an authorized representative of the copyright owner, and that he has a good-faith belief that the material is unlicensed.[91]

In practice, not all of the DMCA's bulwarks against over-removal have been as effective as drafters hoped.[92] But the basic idea—that procedural hurdles can reduce over-removal—is

a sound one. It has been embraced by civil liberties advocates around the world, many of whom endorse a longer list of procedural protections, the Manila Principles,[93] as a basis for legislation in their own countries.

c. Federal Criminal Law Platforms are bound by the same federal criminal laws as any potential defendant. Of particular relevance in recent years have been the laws on child pornography and terrorism.

Federal child pornography law has been updated several times to reflect the evolving role of internet intermediaries. Intermediaries that learn of child pornography on their services follow detailed preservation and reporting requirements, in addition to processing removals on an urgent basis.[94] Federal law provides platforms with immunities for this process. Since 2008, federal law has authorized the National Center for Missing and Exploited Children to maintain and share with intermediaries a database of hashes for known child pornography images.[95] Platforms use the hashes to find, delete, and report matching material. The law specifies that platforms are not *required* to filter[96] and that the law may not be construed to require platforms to monitor user communications or affirmatively seek out potential violations.[97]

Federal law also criminalizes knowing provision of material support to designated foreign terrorist organizations (FTOs).[98] The precise contours of material support law as applied to platforms—including whether providing social media accounts constitutes material support—have not been established. Ongoing civil litigation may provide some relevant precedent in the near future or may be resolved on CDA 230 grounds instead.[99] Clarification under criminal law may be longer in coming, since the major platforms have voluntarily adopted proactive efforts going beyond their legal obligations.

In practice, platforms—particularly small ones—may struggle to comply with these laws without silencing nonextremist speech in the process. As a not atypical example, a user or regional law enforcement agency may report that a video of armed men on horseback, shouting in Kurdish, supports terrorism.[100] For a platform without language or regional policy expertise, such a claim is hard to assess. But given uncertainty and a risk of jail time, by far the cheapest and safest course is to assume the worst and take down the video.

If the men in the video really are bad guys—whether or not members of designated FTOs—there may be little harm done. On the other hand, if they are regional leaders opposing radicalization, removing the video could strengthen extremist messages and alienate potential allies. Making the wrong choice can affect not only speech rights but also, as will be discussed below, US security interests.[101] And in complex and volatile situations, platforms may be caught in the middle—and perceived as de facto voices of US priorities and foreign policy. YouTube, for example, has suspended and reinstated the Kurdish YPG militia from the platform;[102] meanwhile, the United States has supported the group as an

important partner in the fight against ISIS,[103] then withdrawn support,[104] seemingly in response to Turkey's insistence that the YPG is a terrorist organization.

Situations like this are complex from moral and policy points of view, even before looking at the law. As a legal matter, assuming that providing open-access services like hosting constitutes material support, the issue is one of "knowledge." When does a platform know that a group is engaged in terrorism and that online content supports it? For examples like the Kurdish video, one of the two criminal material support statutes, 18 USC 2239B, at least partially answers these questions. It bars support only for organizations designated by the secretary of state as FTOs. So if a platform can figure out which organization the video "supports," its analysis is done. The other statute, 2239A, however, potentially requires much more difficult factual determinations because it covers terrorist acts by organizations not on the FTO list. Assuming the statute applies to online content and platforms, it provides strong legal incentive to err on the side of deletion, and very little protection for lawful—or strategically important—speech.

2. Law outside the United States Outside the United States, the legal picture is very different. For the most part, laws are far friendlier to claimants or regulators and less protective of online speakers and businesses. Many jurisdictions, particularly in less advanced economies, have no specific statutes on intermediary liability.[105] If claims against platforms come up, they are assessed under preexisting statutes or doctrines. This leads to considerable uncertainty. Counsel in these countries may advise that, for platforms, removal is the only safe course, even for highly disputable claims.

The most expansive statutory scheme outside the United States comes from the European Union's eCommerce Directive. The directive covers all speech-related claims against platforms, from crimes to small-scale copyright infringement. It requires internet hosts that know about unlawful content to remove it, or face liability themselves.[106] Notice and takedown systems under European law implementing the directive look something like the DMCA. Because companies must remove content even for complex and fact-based claims such as defamation, however, they often have to make difficult legal judgment calls with little or no information about the underlying dispute. In addition, European laws rarely prescribe specific notice and takedown procedures, so corrective measures like counter-notice are rare.

The directive also states that EU member states may not impose "general" obligations for intermediaries to monitor user-generated content.[107]

The meaning of the prohibited "general" monitoring is disputed, in part because the directive also says that courts may order platforms to "prevent" future infringements.[108] So far, the Court of Justice of the EU (CJEU) has considered and rejected monitoring orders four times, in one case spelling out a possible limiting principle: national courts cannot

order "active monitoring of *all* the data of *each*" of a platform's users, but they may be able to order measures targeting specific users.[109]

A new case pointedly raising this issue is pending before the CJEU now. It will review an Austrian court's order for Facebook to remove user posts calling the leader of the country's Green Party a "lousy traitor" and "corrupt bumpkin."[110] Because the lower court concluded that these terms were illegal hate speech, it ordered Facebook to remove the posts and monitor to ensure they would not reappear in the future—for Facebook users anywhere in the world.

The Facebook case could turn, not on the directive, but on internet users' free expression rights. The broad question is, do free expression rights of internet users create limits on intermediary liability laws that incentivize private platforms to remove internet users' legal speech? The narrower version of the question, in the Facebook case, is whether the government violates internet users' rights by assigning platforms a proactive policing obligation.

In Europe, as in the United States, the answer does not turn entirely on substantive speech laws, such as statutes defining illegal hate speech. (Although the state's interest in enacting a particular law, and the clarity of the legal prohibition platforms must apply, could certainly matter.) Rather, it turns on the likely consequences of delegating enforcement of the speech-restrictive law to a private platform. Yale law professor Jack Balkin has called this a question of "collateral censorship."[111]

The European answer to this question has been mixed. In the strongest pro-speech ruling, the European Court of Human Rights (ECHR) found that compelling a news platform to police user comments in search of defamatory ones would have a "chilling effect on the freedom of expression on the Internet."[112] Monitoring obligations would both incentivize over-removal and discourage private development of open online forums. Accordingly, the court overruled a Hungarian court that had held the platform strictly liable for user comments. The ECHR, however, reached the opposite conclusion in a very similar case—also about news forum comments—because the unlawful content at issue was hate speech, rather than defamation. Given the state's exceptionally strong interest in regulating hate speech under European law, the court said, it could permissibly burden free expression rights by requiring a news site to constantly review and erase internet users' comments.[113]

Legal questions about intermediary liability and internet user rights in Europe are arising in the legislative context as well. Mounting political pressures are straining the EU's existing intermediary liability laws and may soon change them significantly. Germany's NetzDG is the most extreme law actually enacted so far, but it is unlikely to remain so. Both the United Kingdom[114] and France[115] have considered stiff penalties, including jail time, for individuals who visit Jihadi websites. And leaders of both countries, along with Italy, say

platforms should face fines if they do not find and remove terrorist content within two hours after it is uploaded.[116] Theresa May and Emanuel Macron have also been outspoken in insisting that platforms can and should use filters more aggressively.

One of the most politically significant demands for online speech filters to date was the European Commission's 2017 Communication on Illegal Content.[117] It calls for pervasive, "fully automated deletion or suspension of content" for material ranging from child pornography to "xenophobic and racist speech that publicly incites hatred."[118] The Communication suggests that human review is unnecessary before platforms suppress "known" illegal content, including videos reported as such by the police but not reviewed by courts.[119] While this document itself has no force of law, it is a strong signal of politically ascendant approaches to online speech in the EU. And real legislative filtering requirements may be imminent in another area, with the Commission's strong support. The controversial new Copyright Directive[120] would, if enacted in its current form, require services hosting a "large amount" of user content to implement filters.[121]

European pressure stems in part from concerns about terrorism, immigration, and hate speech.[122] A significant current of the European conversation, however, involves competition and the power of US-based internet companies.[123] New content removal obligations for platforms are seen as long overdue curbs on the companies' power—even as critics point out that platforms' power only increases if they are made the de facto interpreters and enforcers of national speech laws.

Europe is not alone in demanding increased liability or removal efforts by platforms. Vietnam, for example, recently reached an agreement with Facebook to prioritize removal requests from state ministries[124]—and flexed its economic muscle by dropping state-controlled companies' advertising on YouTube when ads appeared next to videos critical of the government.[125] Russia has passed increasingly stringent laws governing online content, including its own expansive version of the "right to be forgotten,"[126] and has used anti-extremism laws to target political satire.[127] Turkey has repeatedly demanded that services including Twitter and YouTube remove user speech, including both comments disparaging of Mustafa Kemal Ataturk, the founder of modern Turkey, and material related to government corruption investigations.[128] When platforms refuse, Turkey sometimes blocks them entirely.[129] The relationship between US platforms and Turkish takedown demands may become more complex, particularly in the CVE area, with fluctuating political and military relationships.

This welter of overlapping and competing demands for ratcheted-up enforcement is unlikely to subside soon. The 2017 meeting of the G7 countries, for example, led to a statement urging tech companies to build tools for the "automatic detection of content promoting incitement to violence"[130]—without addressing states' widely varying definitions of those terms. A later G20 statement similarly called for "appropriate filtering, detecting and

removing of content that incites terrorist acts."[131] As I will suggest in the following sections, arriving at interpretations consistent with US law and aligned with US interests may be a significant challenge.

III. Unintended Consequences of Removal Efforts

This section will discuss likely unintended consequences from badly designed intermediary liability laws in three areas: speech rights, national security, and the economy. In practice, the three overlap considerably. For example, US courts considering a law that regulated extremist content online would ask both about the law's consequences for speech rights and whether it achieved its security goals. Similarly, legal regimes that deter investment in speech platforms affect both the economy and internet users' exercise of free expression rights. Policy makers should look to this big picture—recognizing both upsides and downsides—in considering proposed internet content regulations in the coming years.

A. Speech Consequences and the First Amendment

The most obvious problem with poorly crafted intermediary liability rules is the one discussed above: platforms will erase lawful online speech. For US policy makers looking at proposals like those currently circulating in Europe, this could be a showstopper. This section will review at a high level how courts have applied the First Amendment to laws that regulate speech indirectly, by placing responsibility on entities other than speakers themselves. It will then discuss takeaways for internet regulation.

The Supreme Court has set a high First Amendment bar for laws affecting online speech, starting with the seminal 1997 *Reno v. ACLU* ruling.[132] Most recently, in *Packingham v. North Carolina*, the court unanimously rejected a law barring sex offenders from social media sites.[133] To date, the court has not accepted any medium-specific constraints on internet speech and has rejected analogies to regulated media like radio or broadcast. Future cases arising from the pressing issues of today may present the justices with more pessimistic arguments about internet communication. For example, statutes responding to foreign election interference or "fake news" could build on the idea that social networks' novel ability to amplify certain messages and to narrowly target audiences justifies more restrictive legislation. Or in the CVE context, the Court might review arguments made by Cass Sunstein and others that the internet requires courts to broaden the categories of speech considered to present an imminent danger.[134]

In *Packingham*, the Court spoke to the role of social media in today's society. Justice Kennedy wrote that sites like Twitter and Facebook are "integral to the fabric of our modern society and culture," effectively serving as the "modern public square."[135] Barring sex offenders from them completely, the Court concluded, violated the "well established" general rule that "the Government may not suppress lawful speech as the means to suppress unlawful speech."[136]

Although *Packingham* involved direct government regulation of speakers, its rule against overreaching speech suppression is relevant for laws regulating intermediaries as well. The Court fleshed out constitutional parameters for such laws in cases of past decades involving "analog intermediaries" such as bookstores or newspapers that sell advertising space to third parties.[137] While the precise application of these rules to today's intermediaries remains to be seen, the First Amendment clearly sets outside limits on laws that will foreseeably, and avoidably, lead platforms to silence lawful speech.

The bookstore cases *Smith v. California* and *Bantam Books v. Sullivan* both overturned laws holding booksellers liable for obscene books on their shelves.[138] In *Smith*, the Court rejected a strict liability rule, noting that a bookseller who is liable for anything on the shelves "will tend to restrict the books he sells to those he has inspected; and thus the State will have imposed a restriction upon the distribution of constitutionally protected, as well as obscene literature."[139] The fact that booksellers, rather than the state, would choose what books to remove was immaterial. This "self-censorship, compelled by the State, would be a censorship affecting the whole public, hardly less virulent for being privately administered."[140]

Bantam Books found constitutional fault with a notice-based system in which state regulators delivered lists of allegedly obscene books to booksellers—who, as the Court noted, lacked the publishers' economic incentive to challenge overreaching removal remands.[141] State action that led private booksellers to silence speech without judicial review was, the Court found, an unconstitutional prior restraint.[142]

A lower court case, *CDT v. Pappert*, applied this precedent in the internet context. Reviewing a law that required internet service providers (ISPs) to block child pornography, the court found prior restraint operating there as well. The *Pappert* court found the law unconstitutional because it foreseeably led the ISPs to suppress too much speech: although the law did not require it, ISPs commonly blocked all content at a particular internet location, including legal speech, to avoid risk and compliance costs. The court concluded that the statute could have been drafted to achieve the state's goals without this collateral damage and that it therefore failed First Amendment review.[143]

Pappert and the Supreme Court bookseller cases involved criminal liability, but the limits they establish arise in civil cases as well. As the Court said in striking down state libel law in *New York Times v. Sullivan*, "what a State may not constitutionally bring about by means of a criminal statute is likewise beyond the reach of its civil law[.] The fear of damage awards . . . may be markedly more inhibiting than the fear of prosecution under a criminal statute."[144] *Sullivan*, like the bookstore cases, concerned a defendant acting as an intermediary. The plaintiffs sued the newspaper, not for its own reporting, but for allegations made in a paid ad placed by civil rights activists. If a newspaper had to investigate every claim made in paid third-party content, the court said, it "might shut off an important outlet for the promulgation of information and ideas by persons who do not themselves have access to

publishing facilities."[145] Following the same reasoning and relying on both *Sullivan* and *Smith*, a lower court in an important pre-CDA 230 case, *Cubby v. Compuserve*, similarly limited an intermediary's liability for defamation claims.[146]

As discussed earlier, platforms operating notice and takedown systems often remove user-generated content needlessly because they interpret unclear laws too cautiously or take bad-faith accusations at face value. Courts have spoken to this issue in the First Amendment context as well. The Supreme Court in *Reno* noted that the law overturned in that case "would confer broad powers of censorship, in the form of a 'heckler's veto,' upon any [person]" who brought a false claim to an internet company.[147] The Fourth Circuit made similar points in the first major CDA 230 ruling, *Zeran v. AOL*: "Liability upon notice has a chilling effect on the freedom of Internet speech" because of platforms' "natural incentive simply to remove messages upon notification."[148]

What kind of intermediary liability law might be sufficiently well designed to satisfy First Amendment review? In nonlegal terms, the answer might be "a law that solves a serious problem with a minimum of collateral damage." One legal version of this standard, used in *Reno*, is that a law's "burden on protected speech cannot be justified if it could be avoided by a more carefully drafted statute."[149]

Laws in the United States and around the world provide models—and experience with real-world outcomes—of what "carefully drafted" legislation designed to limit burdens on protected speech might look like.[150] This experience, and the First Amendment case law discussed above, suggest four important takeaways.

First, "rigorous procedural safeguards" matter.[151] Courts and legislatures in some parts of the world have concluded, in cases drawing substantially on US precedent, that only judicial review suffices to protect online speakers' rights.[152] Under those countries' laws, people asking platforms to take down speech must provide a court order to substantiate their claims. The US First Amendment would presumably not support a blanket court order requirement of this sort, given the notice and takedown framework already accepted in the DMCA and criminal laws. But US courts could potentially conclude, as some foreign ones have done, that prior judicial review is constitutionally required for complex or nonurgent claims.[153]

And importantly, procedure outside of courts, as part of a private notice and takedown process, can also help protect online expression. Protections like the DMCA's counter-notice process to rebut wrongful accusations may deter abusive removal demands and increase the likelihood that wrongful removals will be corrected. Procedures of this sort could be an important tool in "tailoring" intermediary liability laws.

A second important point is that internet users' speech is particularly threatened when platforms must proactively police it. Using flawed technical filters to automatically erase

speech poses a particularly obvious problem. As discussed earlier, filters often fail, whether by deleting the wrong thing entirely or by misunderstanding news reporting and other legitimate uses.[154] But whatever means platforms use to identify objectionable content, monitoring obligations give them reason to take down even more legal speech—as the Supreme Court discussed in rejecting strict liability for booksellers in *Smith* and as EU courts have addressed in internet filtering cases.[155] The United States has successfully steered clear of such obligations to date. As described earlier, Congress rejected monitoring requirements for internet platforms in the only two statutes—the DMCA and child pornography law—that address the issue.[156]

A third takeaway is that a great deal turns on the mental state requirement for liability. In both *Smith* and *Sullivan*, the Court said that laws holding defendants liable without sufficient awareness of unlawful speech violated the First Amendment. Extensive DMCA case law examines and turns on the question of what constitutes "knowledge" in the intermediary liability context. At one extreme, "knowledge" could mean simple awareness that particular content exists on the platform. At the other, it could mean awareness that a court has adjudicated the content unlawful in a fair proceeding—which, as discussed earlier, is the standard applied by some non-US courts based on their equivalents of the First Amendment. In between lie "knowledge" standards applied in the DMCA and other contexts, such as awareness of content that the reasonable nonlawyer would recognize as illegal. Whatever standard the law sets will shape the likely margin of over-removal carried out by cautious platforms against lawful speech.

A fourth and final conclusion is that the nature of the harmful content matters. As an initial matter, the state's interest in passing a law—and its tolerance for collateral damage to speech—may vary depending on the threats the law averts. But the kind of content at issue will also affect platforms' likely error rate and the value of procedural protections or other statutory "tailoring" to reduce such errors.

At one extreme end of the harmful content spectrum is child pornography, which presents the most urgent and universally recognized need for platform removal obligations. This is in part because it is so harmful. Importantly, though, it is also uniquely recognizable. Lawmakers can reasonably expect platforms to "know it when they see it" in most cases.[157] Because there is no legal context for child pornography, little legal or no judgment is called for. The margin of over-removal and collateral damage to lawful speech is likely to be small. Over-removal, when it happens, mostly results from concerned or cautious platform employees underestimating the ages of people depicted in sexual images or videos. In policy discussions, only the strongest free expression advocates tend to get exercised about laws that inadvertently suppress this material.[158] Finally, the kinds of procedural protections that make sense for less serious claims—such as notification for alleged copyright infringers—may reasonably be omitted from laws that concern, and help law enforcement investigate, serious crimes.

Elsewhere on the spectrum of potentially unlawful content, the kinds of tailoring that might avoid unnecessary harm to lawful speech are much more complex. When platforms must resolve privacy claims involving public figures, for example—as happens with the European Right to Be Forgotten—the risk that they will misjudge the law is much higher. The cost of that error, too, is very different. Removing unflattering news reports of a current or future political candidate, for example, may seriously harm the public interest.

Perhaps the most complex category of content, for First Amendment purposes, is terrorist recruitment material or propaganda. As will be discussed in the next section, the exact nature of the threat posed by this kind of online speech is difficult to assess. Platforms' error rates in sorting legal from illegal material in this area will be particularly high—and particularly consequential. They may easily silence political speech (like criticisms of governments), religious speech (like sermons by potentially extremist religious leaders), or news reporting (like broadcasts excerpting ISIS video content). Existing efforts too often lead to such outcomes already. Given these concerns, First Amendment tailoring for potential platform liability laws governing terrorist content could be particularly challenging.

B. Security Consequences

Demands for platforms to eliminate extremist content have increased exponentially in recent years.[159] This pressure reflects the increased online presence of terrorist organizations, as well as fear of new terrorist attacks in the United States and around the world. But the precise theory of harm reduction behind, for example, the UK government's demand that platforms identify extremist content within two hours, is not always clear. That makes the effectiveness of these takedowns difficult to measure. We can tally some numbers—like the nearly five hundred thousand accounts Twitter suspended in 2017, or the ten thousand Google employees who work on content policy enforcement—but it is unclear how these relate to enhanced security.[160] And disregarding the realities of notice and takedown can lead to untethered numbers and strange analysis. For example, the European Commission recently celebrated major platforms' 70 percent takedown rate for hate speech notices[161]—without knowing what portion of those notices accurately identified illegal material in the first place. Even more mathematically sound data about online content, though, can only be a proxy for real-world security wins.

One definition of success for platform CVE removal campaigns is the creation of safer spaces online, where the average user will not encounter offensive or frightening content. Platforms and advertisers, worried about reputation and loss of customers, are likely to be particularly motivated by this concern. Some governments may agree and see it as their job to protect citizens from disturbing ideas or speech. For the United States, though, such regulation of adults' speech and information access fits poorly with the Constitution and

cultural norms—including the expectation that a free and informed citizenry, not one shielded from bad speech, is a source of national strength.

For US purposes, the compelling goal is not to regulate the speech itself or to hide information from ordinary citizens. Rather, it is to limit genuinely dangerous consequences. Preventing terrorist attacks is, in the Supreme Court's words, "an urgent objective of the highest order" and one that can justify some constraints on speech.[162] The most imminently threatening speech in the CVE context may be communications between individuals planning or executing attacks. Private or encrypted messages of this sort are rarely at issue in the context of content takedown, however, and they tend to raise questions about surveillance law rather than intermediary liability. Far more typical, and at the center of most public discussion, is content that may lead to violence by cultivating, radicalizing, or recruiting adherents to extremist causes. This potentially radicalizing material runs the gamut from beheading videos to religious sermons to extravagant lies about daily life in the Caliphate.

What response to potentially radicalizing speech online can best prevent real-world harm? This is at base an empirical question and a massively complicated one. It requires a closer look at the radicalization process and the internet's role in it. It also requires considering the operational reality and consequences of platform content removals.

This is not—at all—to say that platforms should never remove extremist content. Rather, the point is that any pragmatic policy calculation must factor in real-world costs and weigh them against benefits. The costs side of the ledger includes problems already familiar from other platform content removal efforts.

1. Radicalization and Online Content: What Do We Know? As explained in a recent Brookings Institute report, many questions about online speech and harm reduction remain unanswered: "Further study is required to evaluate the unintended consequences of [social media] suspension campaigns and their attendant trade-offs. Fundamentally, tampering with social networks is a form of social engineering."[163]

For all the untold pages and grant dollars dedicated to the topic of online extremism, we still know remarkably little about when extremist speech leads to violence and how to prevent that from happening. The Brookings report described literature in the field as being rife with "anecdotal observations, strongly held opinions, and small data samples derived with relatively weak—or entirely undisclosed—methods."[164] Other critics call it an "explosion of speculation with little empirical grounding."[165]

Better substantiated conclusions typically relate to online speech, rather than off-line behavior. Few would dispute, for example, that online radicalization is a goal for ISIS and

other foreign terrorist organizations—or that the internet vastly increases access to their materials.[166] And the internet clearly has an important role in radicalization, though not one that can be cleanly separated from the outside world. A 2017 literature review identified an emerging "consensus that the internet alone is not generally a cause of radicalisation, but can act as a facilitator and catalyser of an individual's trajectory towards violent political acts."[167] In the words of a German government report, "The internet does not replace the real world influences but reinforces them."[168]

We do not know how seeing—or even sharing—this material affects an individual's chances of engaging in political violence.[169] As two Rand studies on radicalization found, this is a real shortcoming. Conflating online speech with real-world behavior can "lead policymakers in the wrong direction when it comes to counter-radicalization programs."[170] And to my knowledge, no empirical studies address what happens when online content is identified as extremist and disappears. We have very little sense of how particular removal policies (that is, what content is banned) or operational processes (how platforms decide what to take down and how they communicate with users) affect people at risk of radicalization.

Security experts drawing on the limited data and their own sensibilities arrive at different conclusions about the value of aggressive platform CVE campaigns, though the 2017 literature review found that a majority now doubt their efficacy.[171] The Brookings report, for example, concluded that "while it is possible to target suspensions in a manner that would be far more devastating to ISIS [Twitter] networks, we do not advise such an approach."[172] University of Maryland war studies professor Peter Neumann goes farther: "Approaches that are aimed at reducing the supply of violent extremist content on the Internet are neither feasible nor desirable."[173]

2. Radicalization and Content Removal: What Happens When We Take Things Down? Although we know little about the real-world consequences of CVE removal campaigns, we can easily extrapolate from twenty years of experience with other kinds of online content removal. That experience suggests that badly designed or executed CVE campaigns may be worse than merely ineffective. They could also make us less safe.

Experts, including the EU counter-terrorism coordinator, have emphasized the urgent need for CVE measures to rectify "the sense of social marginalisation which plagues Muslim communities across Europe."[174] Experience with notice and takedown systems tells us that any content removals, but especially erroneous ones, tend to do just the opposite. They make internet users feel outraged and powerless. In other words, platform CVE efforts may cultivate precisely the attitudes and animosities that counter-radicalization efforts are supposed to prevent.

The fantasy of internet content removal is that unwanted information vanishes without a trace. Perhaps a propagandist huddled over a laptop in Syria gnashes his teeth in

frustration, or a worried mom of teenagers in Molenbeek breathes a sigh of relief, and the world otherwise goes on. The reality is different. The act of deleting online content has consequences—just as it has had in other internet contexts for decades, and just as it does when other enforcers deny parade permits, seize film reels, or black out text in letters and magazines. In the "social engineering" experiment of CVE, those consequences matter.

a. Over-removal and Countering Violent Extremism Perhaps unsurprisingly, reports of over-removal resulting from platform CVE campaigns are now commonplace. As mentioned earlier, YouTube took down videos of Syrian atrocities posted by a UK human rights watchdog.[175] Facebook, similarly, accidentally deleted the page of a Chechen pro-independence group despite its opposition to terrorism.[176] It also removed posts documenting Rohingya ethnic cleansing in Myanmar,[177] reportedly because it had classified Rohingya organizations as dangerous militant groups.[178]

Individuals' unremarkable and innocuous online speech also frequently disappears, often with no remedy or acknowledgment of the error. A British Muslim woman known to the author, for example, found that a prayer she posted on Facebook had been removed for violating the platform's community standards. The prayer's text read, in Arabic, "God, before the end of this holy day forgive our sins, bless us and our loved ones in this life and the afterlife with your mercy, almighty."

The more platforms are pushed to instantaneously police the digital world, the more common such errors will become. We should expect to see them in particular for internet users speaking in Chechen, Farsi, Indonesian, and other languages common in Muslim-majority countries. Few tech company employees hired to review complaints or machine-generated flags will be fluent in these languages or able to grasp local political context or nuance. Platforms are less economically motivated to fine-tune their operations or address errors affecting small or remote markets and are more likely to simply hit the delete key. The result will be more mistakes and more understandably angry users.

Stories like the British woman's are unfair and troubling individually. They are even more troubling collectively. Systematic and uncorrected over-removal affecting internet-savvy Muslims, including immigrants and children of immigrants in places like Brussels, Paris, or New York, is a slap in the face of the very people those cities depend on to help resist radicalization in their communities.[179]

b. Impact on Speakers The human and psychological toll of internet content removal, and over-removal, can affect anyone living with a Swiss cheese version of the internet. Several studies show that internet users throughout the world self-censor when they are aware of potentially being watched—including by avoiding searches on sensitive health topics like eating disorders or depression.[180] More costs are to be expected among people who see themselves as the objects of suspicion and censorship. The implied message of overbroad

removal campaigns—that affected groups cannot be trusted to discuss their religion or public affairs unsupervised and that sacrificing ordinary people's speech rights is acceptable to the US and other governments—is, at the very minimum, an unwelcoming one.

The impact is worse for those individuals who open Facebook or Twitter one day to find their own posts gone or their friends, respected community leaders, or news sources banned. Indignant and angry reactions under these circumstances are common even in apolitical contexts like copyright infringement. They are understandably stronger when people feel they are the target of, in the words of attorneys suing YouTube for different removal decisions, "censorship based entirely on unspecified ideological objection to the message or on the perceived identity and political viewpoint of the speaker."[181]

The operational details of removal can make matters worse. People are often particularly frustrated when they are not told the reason for a removal or when platforms seem to have applied rules unfairly.[182] The possibility of appealing an apparent error also matters. Users who can find no redress from faceless internet platforms—and particularly platforms acting at the behest of governments—may feel all the more powerless and disenfranchised.[183]

Frustration and mistrust are not limited to individuals directly affected. Stories and outrage about bad content removals spread within communities as people blog, tweet, or talk to friends about them. In internet speech policy circles, outrage over actual or perceived injustices of this sort is a staple. These injustices fuel an ongoing cycle of blog posts, news articles, conferences, academic careers, public interest campaigns, and lawsuits.[184] Particularly striking examples—like Facebook's removal of an iconic Vietnam War photo[185] or Twitter's temporary suspension of President Trump's account[186]—become national news. The social damage from overzealous removal campaigns spreads.

By imposing costs on individuals and communities well beyond actual extremists, CVE efforts can reinforce the very problems they were meant to correct. Feelings of alienation and social exclusion are, security researchers say, important risk factors for radicalization,[187] as are frustration and moral outrage.[188] Knowing this, yet accepting aggressive CVE campaigns' likely impact, may be a serious miscalculation. If suppressing propaganda from real terrorists comes at the cost of high over-removal rates for innocent Arabic-language posts or speech about Islam generally, the trade-off may be not only disrespectful and unfair but dangerous.

c. Echo Chambers, Counter-speech, and Political Dialogue Beyond their downsides for social trust and alienation, overreaching removals in the CVE context can also distort important political conversations. One facet of this problem is the isolation of extremists in dark corners of the internet. As the Brookings study noted, being barred from Twitter may drive

potential ISIS recruits into increasingly concentrated and insular groups on other platforms. There, they enter a "much louder echo chamber" that may "speed and intensify the radicalization process."[189] Ironically, CVE campaigns may reinforce recruiters' own efforts to shift conversations with potential recruits out of the public eye.[190]

Moving these discussions into echo chambers does more than increase the power of extremist voices. It also decreases that of opposing voices, including from the potentially most effective sources: peers and community members. The idea, as researchers at London's International Centre for the Study of Radicalisation explained, is simple: "If allowed to fester in an uncensored internet, the narrative will become less appealing."[191] These "organic social pressures that could lead to deradicalization" are an important part of open platforms like Twitter—and may be reduced or eliminated in more insular settings.[192] Some studies suggest that pushback from respected members of a speaker's own social group can be the most effective means of de-escalating verbal aggression.[193] Governments and NGOs have launched affirmative internet counter-narrative campaigns on this basis.[194] Google's Jigsaw division has recently experimented with a variant, the "Redirect Method," which provides people with counter-messaging at the point of initial interest, as expressed in search queries.[195]

Of course, the political dialogue that shapes radicalization choices is not just a tug-of-war between speakers at two extremes. Any pressing public issue—and certainly those so compelling as to inspire acts of terrorism—is likely to generate a wide array of perspectives. Platform removals that do not recognize nuanced differences and gradations between politically engaged speakers can easily silence the wrong ones. A particular risk comes from suppressing speakers who act as voices of moderation within their own political spectrum, especially those who share experiences and grievances with potential extremists but who oppose violence. De-radicalized former ISIS recruits, for example, are believed to be among the most effective participants in counter-radicalization efforts.[196] When platforms silence little-known (to them) speakers addressing topics related to extremism, these important voices may be lost.

d. Law Enforcement Tools and Priorities A final security concern about platform removal campaigns is their impact on law enforcement and intelligence efforts. In part, this is a question about allocation of resources—between policing speech and policing off-line activity. London's Metropolitan Police, for example, led Europe in the development of Internet Referral Units charged with finding and reporting extremist content to online platforms.[197] At the same time, Theresa May as home secretary presided over a 13 percent reduction in police officers in England overall.[198] This prioritization may have had serious consequences. Reports suggest that attackers in both Manchester and London had been identified to police by concerned friends, for example, but overburdened law enforcement agencies were unable to act on the information.[199]

Counter-terrorism scholars Peter Neumann and Shiraz Maher described a related problem in the government's response to convicted British extremist Anjem Choudary, who is said to have inspired the London Bridge attackers.[200] Choudary had a YouTube channel, but "practically all of his followers were known to him personally and were recruited face to face," they explained. "It is one thing for the internet companies to pull down radical propaganda. But they face an uphill battle while preachers such as Choudary have spent years spreading their message virtually unchallenged on British streets."[201]

The narrower and in some sense simpler question is whether removal campaigns may hurt law enforcement efforts by "destroy[ing] valuable sources of intelligence."[202] Different agencies—domestically and internationally—may have very different strategies and priorities regarding online extremist activity. When they do not coordinate, platforms can be caught in the middle. A rare case in which such a conflict became public occurred in 2010, when the CIA unsuccessfully opposed Pentagon efforts to shut down Al Qaeda online forums. As one official said, this caused serious setbacks for CIA efforts that depended on the forums: "[We] understood that intelligence would be lost, and it was; that relationships with cooperating intelligence services would be damaged, and they were; and that the terrorists would migrate to other sites, and they did."[203]

C. Economic Consequences

America's strong intermediary liability laws are broadly credited as economic drivers. In his forthcoming book, US Naval Academy cybersecurity professor Jeff Kosseff calls CDA 230 "the twenty-six words that created the Internet."[204] This was, of course, precisely what Congress meant to do. In the statute's words, CDA 230 serves to "preserve the vibrant and competitive free market" for internet services, "unfettered by Federal or State regulation."[205]

While Congress in CDA 230 focused on the growth of the internet services themselves, the economic benefits of smart intermediary liability laws—and the harms of foolish ones—go well beyond the internet sector. The emergence of new platforms like Yelp or eBay, for example, have ancillary benefits for businesses that use them to find or transact with customers. And business developments made possible by intermediary immunities—like the emergence of mobile apps and app stores—create entirely new opportunities for entrepreneurs building everything from star-finder apps to makeup tutorials.

Laws that drive popular platforms to over-remove can hurt businesses as much as they hurt individual users. Competitors all too frequently abuse notice and takedown systems to target one another. In 2006, fully half of Google web search removal demands fell into this category.[206] In a typical example, one driving school tried to exclude another from

search results by claiming copyright in an alphabetical list of cities.[207] Small businesses—from the seamstress who sells her work on Etsy or eBay to the mechanic whose customers find him through a web search, online ads, or Yelp—have a lot to lose. Being improperly removed from these platforms is something like disappearing from the Yellow Pages in past decades—but worse, because many small businesses now lack any physical storefront. And a climate of unpredictable removals, with resulting fluctuation in revenues, undermines small online businesses generally.

Intermediary liability laws are also essential for small platforms—from modest examples like a corner café with Wi-Fi or a daycare center with a blog for parents, to the next Snapchat, WhatsApp, or Instagram. Counting the number of US companies that depend on intermediary immunities would be an impossible task. As one measure, Engine, an advocacy organization representing start-ups, lists over a thousand members. As another, the Copyright Office's DMCA registration page lists hundreds of thousands of entities that self-identify as intermediaries—and were well-lawyered enough to register as such. Any of these could be wiped out if they faced liability whenever a user shared a pirated song or unseemly video of Hulk Hogan.

As for the next potential challenger to giants like Google or Facebook, many view companies like video-hosting site Veoh as cautionary tales. Veoh, a YouTube competitor with largely similar service, had millions of users and some $70 million in investment from sources like Goldman Sachs and Time Warner.[208] Nevertheless, it became, as *Wired* put it, one of a "long list of promising start-ups driven into bankruptcy by copyright lawsuits"[209]—against both the company and its investors. Although Veoh ultimately prevailed under the DMCA, it did not survive the litigation. Meanwhile, YouTube—backed by the resources of corporate parent Google—emerged intact from very similar, and nearly simultaneous, litigation.[210] The realistic consequence of stories like Veoh's may be that the next YouTube competitor simply never gets funded. This would be consistent with venture capitalists' own reports, in surveys, that unclear or weak intermediary liability laws deter them from investing.[211]

Laws of the sort currently under discussion in Europe, which create not only legal risk but the necessity for up-front spending, may have even starker effects. The UK's prosposed two-hour takedown requirement, or even Germany's twenty-four-hour requirement may be supportable for companies with multilingual, around-the-clock compliance teams.[212] But start-ups cannot pay lawyers to scrutinize every dubious legal request that comes in the door. Even having existing employees spend time guessing about the law is too expensive for many. Small internet registrars, for example, informally report that legal complaints are a growing and expensive problem. If governments go further and require companies to build technical filters, barriers to market entry and disadvantages for small players will become even greater. Few will have the option to do as YouTube did: invest $60 million in the problem.[213] If the European Union goes forward with recent proposals, the Facebooks

and YouTubes of the world will adapt and survive, if not in exactly the form we are used to. Smaller companies, most likely, will not. For European lawmakers, with relatively small domestic internet sectors, this may not be a major consideration. For the United States, though, it adds yet another reason to be careful and smart with internet regulation.

Operators of smaller platforms who track the political winds on these issues are well aware of the threat. In the 2016 Berkeley study, many US companies stated flatly that they could not afford to match the big players' content removal processes and technologies.[214] Even if they could, legal compliance could lock them into inefficient or rapidly obsolescing technical standards, adding another disadvantage compared to larger players.[215] It is little wonder that these smaller platforms describe feeling "left aside in policy debates and news accounts skewed by attention to the relatively few" larger actors.[216] Policy makers genuinely concerned with their well-being and with competition against giant incumbents can correct this by looking more closely at the economic consequences of intermediary liability law changes.

IV. Conclusion

Current attitudes toward intermediary liability, particularly in Europe, verge on "regulate first, ask questions later." I have suggested here that some of the most important questions that should inform policy in this area already have answers. We have twenty years of experience to tell us how intermediary liability laws affect, not just platforms themselves, but the general public that relies on them. We also have valuable analysis and sources of law from pre-internet sources, like the Supreme Court bookstore cases. The internet raises new issues in many areas—from competition to privacy to free expression—but none are as novel as we are sometimes told. Lawmakers and courts are not drafting on a blank slate for any of them.

Demands for platforms to get rid of all content in a particular category, such as "extremism," do not translate to meaningful policy making—unless the policy is a shotgun approach to online speech, taking down the good with the bad. To "go further and faster" in eliminating prohibited material, platforms can only adopt actual standards (more or less clear, and more or less speech-protective) about the content they will allow, and establish procedures (more or less fair to users, and more or less cumbersome for companies) for enforcing them.

On internet speech platforms, just like anywhere else, only implementable things happen. To make sound policy, we must take account of what real-world implementation will look like. This includes being realistic about the capabilities of technical filters and about the motivations and likely choices of platforms that review user content under threat of liability.

NOTES

1 Ellen Nakashima, "Obama's Top National Security Officials to Meet with Silicon Valley CEOs," *Washington Post,* January 7, 2016, accessed May 24, 2018, https://www.washingtonpost.com/world/national-security/obamas-top-national-security-officials-to-meet-with-silicon-valley-ceos/2016/01/07/178d95ca-b586-11e5-a842-0feb51d1d124_story.html?utm_term=.64e49be4ee75; Melissa Eddy and Mark Scott, "Facebook and Twitter Could Face Fines in Germany over Hate Speech Posts," *New York Times*, March 14, 2017, accessed May 24, 2018, https://www.nytimes.com/2017/03/14/technology/germany-hate-speech-facebook-tech.html.

2 Christopher Hope and Kate McCann, "Google, Facebook and Twitter Told to Take Down Terror Content within Two Hours or Face Fines," *Telegraph,* September 19, 2017, accessed May 27, 2018, http://www.telegraph.co.uk/news/2017/09/19/google-facebook-twitter-told-take-terror-content-within-two.

3 Sam Levin, "Civil Rights Groups Urge Facebook to Fix 'Racially Biased' Moderation System," *Guardian,* January 18, 2017, accessed May 27, 2018, https://www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter.

4 "PragerU Takes Legal Action against Google and YouTube for Discrimination," news release, accessed May 27, 2018, https://www.prageru.com/press-release-prager-university-prageru-takes-legal-action-against-google-and-youtube-discrimination.

5 Ashley Gold, "'We've been censored,' Diamond and Silk tell Congress," Politico, April 26, 2018, accessed June 4, 2018, https://www.politico.com/story/2018/04/26/diamond-and-silk-congress-hearing-1116887.

6 Natasha Lomas, "DigitalOcean and Cloudflare Ditch Neo-Nazi Client, The Daily Stormer," TechCrunch, August 16, 2017, accessed May 24, 2018, https://techcrunch.com/2017/08/16/digital-ocean-and-cloudflare-ditch-neo-nazi-client-the-daily-stormer.

7 April Glaser, "Want a Terrible Job? Facebook and Google May Be Hiring," *Slate,* January 18, 2018, accessed May 24, 2018, https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html.

8 Associated Press, "Facebook Announces New Ad Transparency before Russia Hearing," *Los Angeles Times,* October 27, 2017, accessed May 24, 2018, http://www.latimes.com/business/la-fi-tn-facebook-ad-transparency-20171027-story.html.

9 Act to Improve Enforcement of the Law in Social Networks, December 7, 2017, accessed May 24, 2018, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=099B20F1854667953970FB02EE7A0510.1_cid334?__blob=publicationFile&v=2; see also Philip Oltermann, "Tough New German Law Puts Tech Firms and Free Speech in Spotlight," *Guardian*, January 5, 2018, accessed May 24, 2018, https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

10 "German Hate Speech Law Tested as Twitter Blocks Satire Account," Reuters, January 3, 2018, accessed May 27, 2018, https://www.reuters.com/article/us-germany-hatecrime/german-hate-speech-law-tested-as-twitter-blocks-satire-account-idUSKBN1ES1AT.

11 Jefferson Chase, "Facebook Slammed for Censoring German Street Artist," DW.com, January 15, 2018, accessed May 27, 2018, http://www.dw.com/en/facebook-slammed-for-censoring-german-street-artist/a-42155218.

12 Anupam Chander, "How Law Made Silicon Valley," *Emory Law Journal* 63, 639 (2014).

13 See 17 USC § 501.

14 "Recognizing and Reporting Spam, Inappropriate, and Abusive Content," LinkedIn, updated May 15, 2018, accessed May 24, 2018, https://www.linkedin.com/help/linkedin/answer/37822; "Reporting Inappropriate Content, Messages, or Safety Concerns," LinkedIn, December 2017, accessed May 24, 2018, https://www.linkedin.com/help/linkedin/suggested/146/reporting-inappropriate-content-messages-or-safety-concerns?lang=en.

15 "Community Standards: Introduction," Facebook, accessed May 24, 2018, https://www.facebook.com/communitystandards#encouraging-respectful-behavior; see also "The Twitter Rules," Twitter, accessed May 24, 2018, https://help.twitter.com/en/rules-and-policies/twitter-rules; "Recognizing and Reporting Spam, Inappropriate, and Abusive Content."

16 Packingham v. North Carolina, 137 S. Ct. 1730 (2017); Olivia Solon, "Google's Bad Week: YouTube Loses Millions as Advertising Row Reaches US," March 25, 2017, accessed May 24, 2018, https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon.

17 Natasha Lomas, "After YouTube Boycott, Google Pulls Ads from More Types of Offensive Content," TechCrunch, March 21, 2017, accessed May 24, 2018, https://techcrunch.com/2017/03/21/after-youtube-boycott-google-pulls-ads-from-more-types-of-offensive-content.

18 Amanda Hess, "How YouTube's Shifting Algorithms Hurt Independent Media," *New York Times,* April 17, 2017, accessed May 27, 2018, https://www.nytimes.com/2017/04/17/arts/youtube-broadcasters-algorithm-ads.html.

19 521 U.S. § 844 (1997).

20 Amanda Trejos, "Ice Bucket Challenge: 5 Things You Should Know," *USA Today,* July 3, 2017, accessed May 24, 2018, https://www.usatoday.com/story/news/2017/07/03/ice-bucket-challenge-5-things-you-should-know/448006001.

21 "Teens Are Eating Laundry Detergent for the 'Tide Pod Challenge,'" CBS News, January 12, 2018, accessed May 24, 2018, https://www.cbsnews.com/news/tide-pod-challenge-ingesting-detergent-risks/?ftag=CNM-00-10aab7e&linkId=47101011.

22 Shaun Nichols, "HBO Slaps Takedown Demand on 13-Year-Old Girl's Painting Because It Used 'Winter Is Coming,'" Register, December 8, 2016, accessed May 27, 2018, http://www.theregister.co.uk/2016/12/08/winter_is_coming_hbo_dmca_trademark.

23 Tim Cushing, "Someone under Federal Indictment Impersonates a Journalist to File Bogus DMCA Notice," TechDirt, May 23, 2017, accessed May 24, 2018, https://www.techdirt.com/articles/20170518/09500537404/someone-under-federal-indictment-impersonates-journalist-to-file-bogus-dmca-notice.shtml

24 Eugene Volokh, "Another libel takedown order," *Washington Post,* May 30, 2017, accessed May 27, 2018, https://www.washingtonpost.com/news/volokh-conspiracy/wp/2017/05/30/another-libel-takedown-order/?utm_medium=twitter&utm_source=dlvr.it&utm_term=.b6dac13ae044.

25 Adam Steinbaugh, "Chance Trahan and Revenge Porn Site 'Is Anybody Down,'" YouTube, July 16, 2013, accessed May 27, 2018, https://www.youtube.com/watch?v=9LxZEzXFbBk; Tim Cushing, "Former Revenge Porn Site Operator Readies for Senate Run by Issuing Bogus Takedown Requests to YouTube," TechDirt, October 4, 2017, accessed May 24, 2018, https://www.techdirt.com/articles/20170929/17100738318/former-revenge-porn-site-operator-readies-senate-run-issuing-bogus-takedown-requests-to-youtube.shtml.

26 Eugene Volokh, "[libel takedown litigation articles]," *Washington Post,* various dates, accessed May 27, 2018, https://www.washingtonpost.com/news/volokh-conspiracy/wp/category/libel-takedown-litigation/?utm_term=.87c3336d4fdb.

27 Daphne Keller, "Empirical Evidence of 'Over-removal' by Internet Companies under Intermediary Liability Laws," Center for Internet and Society, October 12, 2015, accessed May 24, 2018, http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

28 Jennifer M. Urban, Joe Karaganis, and Brianna Schofield, "Notice and Takedown in Everyday Practice," UC Berkeley Public Law Research Paper no. 2755628, March 30, 2016. Like most empirical work on intermediary liability, this paper focuses on copyright because that is where data is available. Platforms disclose much more information about copyright claims than other kinds of claims.

29 Urban, Karaganis, and Schofield, "Notice and Takedown," 28–29, see also 73–74.

30 Urban, Karaganis, and Schofield, "Notice and Takedown," 41.

31 The entire set of notices from Google between March 2002 and August 2005 numbered 734, for an average of 210 per year. Jennifer Urban and Laura Quilter, "Efficient Process or 'Chilling Effects'?," *Santa Clara Computer and High Tech Law Journal* 22 (2006): 621, 641.

32 Gina Hall, "How Many Copyright Takedown Notices Does Google Handle Each Day? About 2 Million," *Silicon Valley Business Journal,* March 7, 2016, accessed May 27, 2018, https://www.bizjournals.com/sanjose/news/2016/03/07/how-many-copyright-takedown-notices-does-google.html.

33 Associated Press, "Facebook Announces New Ad Transparency."

34 Urban, Karaganis, and Schofield, "Notice and Takedown," 88.

35 Mark Fahey, "Blame the Robots for Copyright Notice Dysfunction," CNBC, March 29, 2016, accessed May 24, 2018, https://www.cnbc.com/2016/03/29/blame-the-robots-for-copyright-notice-dysfunction.html.

36 Mark Fahey, "Blame the Robots"; see also Urban, Karaganis, and Schofield, "Notice and Takedown," 91.

37 Kristina Cooke, "Facebook Developing Artificial Intelligence to Flag Offensive Live Videos," Reuters, December 1, 2016, accessed May 25, 2018, https://www.reuters.com/article/us-facebook-ai-video/facebook-developing-artificial-intelligence-to-flag-offensive-live-videos-idUSKBN13Q52M.

38 Drew Harwell, "AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how," *Washington Post,* April 11, 2018, accessed June 4, 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/?noredirect=on&utm_term=.90bb780f68ee

39 Declan McCullagh, "Google's Chastity Belt Too Tight," CNET, April 23, 2004, accessed May 25, 2018, https://www.cnet.com/news/googles-chastity-belt-too-tight.

40 John Rehling, "How Natural Language Processing Helps Uncover Social Media Sentiment," Mashable, November 8, 2011, accessed May 25, 2018, https://mashable.com/2011/11/08/natural-language-processing-social-media/#Pilz0fSl3Eq2.

41 Center for Democracy and Technology, "Mixed Messages? The Limits of Automated Social Media Content Analysis," November 2017, 5, accessed May 25, 2018, https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis.

42 Center for Democracy and Technology, "Mixed Messages?," 14, 19.

43 Sophie Curtis, "Facebook, Google, and Twitter Block 'Hash List' of Child Porn Images," *Telegraph,* August 10, 2015, accessed May 25, 2018, https://www.telegraph.co.uk/technology/internet-security/11794180/Facebook-Google-and-Twitter-to-block-hash-list-of-child-porn-images.html.

44 https://developer-westus.microsoftmoderator.com/docs/services/541884a75471830c285b26cf/operations/5418861b5471830c285b26d0.

45 "How Content ID Works," YouTube Help, accessed May 27, 2018, https://support.google.com/youtube/answer/2797370?hl=en.

46 Urban, Karaganis, and Schofield, "Notice and Takedown," 64 (reporting public figure of $60 million and private estimates several times higher).

47 Mike Masnick,"YouTube Takes Down Ariana Grande's Manchester Benefit Concert on Copyright Grounds," Techdirt, June 7, 2017, accessed May 27, 2018, https://www.techdirt.com/articles/20170606/17500637534/youtube-takes-down-ariana-grandes-manchester-benefit-concert-copyright-grounds.shtml.

48 Olivia Solon, "Facebook, Twitter, Google and Microsoft Team Up to Tackle Extremist Content," *Guardian,* December 5, 2016, accessed May 25, 2018, https://www.theguardian.com/technology/2016/dec/05/facebook-twitter-google-microsoft-terrorist-extremist-content.

49 Selena Larson, "Tech Giants Bolster Collaborative Fight against Terrorism," June 26, 2017, accessed May 25, 2018, http://money.cnn.com/2017/06/26/technology/business/global-internet-forum-to-counter-terrorism/index.html.

50 "Extremist Propaganda and Social Media," C-SPAN, January 17, 2018, accessed May 25, 2018, https://www.c-span.org/video/?439849-1/facebook-twitter-youtube-officials-testify-combating-extremism&start=859.

51 "Extremist Propaganda and Social Media," C-SPAN, January 17, 2018, accessed May 25, 2018, https://www.c-span.org/video/?439849-1/facebook-twitter-youtube-officials-testify-combating-extremism&start=1125; see also "Extremist Propaganda and Social Media," C-SPAN, January 17, 2018, accessed May 25, 2018, https://www.c-span.org/video/?439849-1/facebook-twitter-youtube-officials-testify-combating-extremism&start=783; AFP, "Facebook, Twitter, YouTube Pressed over Terror Content," *New Vision*, January 17, 2018, accessed May 25, 2018, https://www.newvision.co.ug/new_vision/news/1469243/facebook-twitter-youtube-pressed-terror-content.

52 Alexis C. Madrigal, "Inside Facebook's Fast-Growing Content-Moderation Effort," *Atlantic,* February 7, 2018, accessed May 25, 2018, https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632.

53 Sam Levin, "Google to Hire Thousands of Moderators after Outcry over YouTube Abuse Videos," *Guardian,* December 5, 2017, accessed May 25, 2018, https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos.

54 Evan Engstrom and Nick Feamster, "The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools" (Engine, March 2017), accessed May 27, 2018, http://www.engine.is/the-limits-of-filtering.

55 Center for Democracy and Technology, "Mixed Messages?"

56 Monika Bickert, "Hard Questions: How We Counter Terrorism," Facebook Newsroom, June 15, 2017, accessed May 25, 2018, https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism.

57 Engstrom and Feamster, "The Limits of Filtering."

58 Malachy Browne, "YouTube Removes Videos Showing Atrocities in Syria," *New York Times,* August 22, 2017, accessed May 25, 2018, https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html; see also Scott Edwards, "When YouTube Removes Violent Videos, It Impedes Justice," *Wired*, October 7, 2017, accessed May 25, 2018, https://www.wired.com/story/when-youtube-removes-violent-videos-it-impedes-justice.

59 Craig Silverman and Jeremy Singer-Vine, "An Inside Look at the Accounts Twitter Has Censored in Countries around the World," BuzzFeed News, January 24, 2018, accessed May 25, 2018, https://www.buzzfeed.com/craigsilverman/country-withheld-twitter-accounts.

60 Both Facebook and Baidu have been sued, unsuccessfully, for removing US content allegedly at the behest of foreign governments. John Ribeiro, "Facebook Sued in US Court for Blocking Page in India," *PC World,* June 3, 2015, accessed May 25, 2018, https://www.pcworld.com/article/2930872/facebook-sued-in-us-court-for-blocking-page-in-india.html; Jonathan Stempel, "Baidu, China Sued in U.S. for Internet Censorship," Reuters, May 18, 2011, accessed May 25, 2018, https://www.reuters.com/article/us-baidu-censorship-lawsuit/baidu-china-sued-in-u-s-for-internet-censorship-idUSTRE74H7N120110518.

61 See Daphne Keller, "Law, Borders, and Speech Conference: Proceedings and Materials," December 15, 2017, Appendix 4, accessed May 25, 2018, https://cyberlaw.stanford.edu/publications/proceedings-volume (listing current cases and linking to further information).

62 https://www.gpo.gov/fdsys/pkg/PLAW-111publ223/html/PLAW-111publ223.htm.

63 T. C. Sottek, "Google Executive Arrested in Brazil as the Company Resists Orders to Freeze Political Speech on YouTube," *The Verge,* September 26, 2012, accessed May 25, 2018, https://www.theverge.com/2012/9/26/3413476/google-brazil-you-tube-arrest.

64 Simon English, "India Throws eBay Chief Into Prison," *The Telegraph,* December 21, 2004, accessed June 4, 2018, https://www.telegraph.co.uk/finance/2902203/India-throws-Ebay-chief-into-prison.html.

65 https://www.nytimes.com/2017/09/06/technology/facebook-china-shanghai-office.html.

66 "Russia Tells Facebook to Localize User Data or Be Blocked," Reuters, September 26, 2017, accessed May 25, 2018, https://www.reuters.com/article/us-russia-facebook/russia-tells-facebook-to-localize-user-data-or-be-blocked-idUSKCN1C11R5.

67 May Bulman, "Facebook, Twitter and Whatsapp Blocked in Turkey after Arrest of Opposition Leaders," *Independent,* November 4, 2016, accessed May 25, 2018, http://www.independent.co.uk/news/world/asia/facebook-twitter-whatsapp-turkey-erdogan-blocked-opposition-leaders-arrested-a7396831.html.

68 Engadget staff, "Why Has Malaysia Blocked Medium?" *Engadget,* January 28, 2016, accessed June 4 2018, https://www.engadget.com/2016/01/28/malaysia-medium-block-explainer/.

69 European Commission, "Your Rights in the EU," accessed May 27, 2018, http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf.

70 See Evelyn Aswad, "The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban," *Columbia Human Rights Law Review*, December 2016, accessed May 25, 2018, http://hrlr.law.columbia.edu/2017/11/15/the-role-of-u-s-technology-companies-as-enforcers-of-europes-new-internet-hate-speech-ban (arguing that the Code of Conduct is inconsistent with a human rights framework).

71 Gareth Corfield, "Another Day, Another Shot Fired at American e-Megacorps," Register, September 28, 2017, accessed May 25, 2018, https://www.theregister.co.uk/2017/09/28/eu_social_media_regulation_warning.

72 Christina Angelopoulos, Annabel Brody, Wouter Hins, Bernt Hugenholtz, Patrick Leerssen, Thomas Margoni, Tarlach McGonagle, Ot van Daalen, and Joris van Hoboken, "Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation" (Institute for Information Law, University of Amsterdam, December 2015), 61, accessed May 25, 2018, https://ivir.nl/publicaties/download/1796; see also Backpage v. Dart (7th Cir. 2015, Posner J.) (sheriff violated First Amendment by pressuring credit card companies to terminate services to prostitution advertising host).

73 "Thailand Warns Facebook to Block Content Critical of the Monarchy," BBC, May 12, 2017, accessed May 27, 2018, http://www.bbc.com/news/world-asia-39893073.

74 Glyn Moody, "Hollywood Helps China Set Up National Surveillance and Censorship System to Tackle Copyright Infringement," Techdirt, May 15, 2017, accessed May 27, 2018, https://www.techdirt.com/articles

/20170512/03344437346/hollywood-helps-china-set-up-national-surveillance-censorship-system-to-tackle-copyright-infringement.shtml.

75 47 USC § 230.

76 §230(e)(1) and (3).

77 §230(a)(4).

78 §230(e)(1), (2), and (3).

79 §230(c)(2)(A).

80 Stratton Oakmont v. Prodigy, 1995 WL 323710 (N.Y. Sup. Ct. 1995).

81 Cubby v. Compuserve, 776 F. Supp. 135 (S.D.N.Y. 1991).

82 Permanent Subcomm. on Investigations, "Backpage.com's Knowing Facilitation of Online Sex Trafficking," S. Staff Rep., accessed May 25, 2018, https://www.portman.senate.gov/public/index.cfm/files/serve?File_id=5D0C71AE-A090-4F30-A5F5-7CFFC08AFD48.

83 H.R. 1865, 115th Cong. (2018).

84 17 USC § 512.

85 Arnold P. Lutzker, "Primer on the Digital Millennium: What the Digital Millennium Copyright Act and the Copyright Term Extension Act Mean for the Library Community," ALA, March 8, 1999, accessed May 25, 2018, http://www.ala.org/advocacy/sites/ala.org.advocacy/files/content/copyright/dmca/pdfs/dmcaprimer.pdf.

86 Reply Comments of the Motion Picture Association of America, In the Matter of Request for Comments on United States Copyright Office Section 512 Study, February 1, 2017, accessed June 4, 2018, https://www.mpaa.org/wp-content/uploads/2017/02/Additional-Comments-of-the-MPAA.pdf.

87 US Copyright Office, "Section 512 Study: Notice and Request for Public Comment," *Federal Register* 80, no. 251 (December 31, 2015): 81862–68, accessed May 27, 2018, https://www.copyright.gov/fedreg/2015/80fr81862.pdf.

88 17 §512(g)(2)(B).

89 17 USC § 512(f).

90 Christian Ahlert, Chris Marsden, and Chester Yung, "How 'Liberty' Disappeared from Cyberspace: The Mystery Shopper Tests Internet Content Self-Regulation," accessed May 25, 2018, http://pcmlp.socleg.ox.ac.uk/wp-content/uploads/2014/12/liberty.pdf (UK and US study); John Leyden, "How to Kill a Website with One Email: Exploiting the European E-commerce Directive," Register, October 14, 2004, accessed May 25, 2018, http://www.theregister.co.uk/2004/10/14/isp_takedown_study (describing Netherlands study).

91 Perfect 10 v. CCBill, 488 F. 3d 1102 (9th Cir. 2007).

92 See Daphne Keller, "Counter-Notice Does Not Fix Over-Removal of Online Speech," Center for Internet and Society, October 5, 2017, accessed May 25, 2018, http://cyberlaw.stanford.edu/blog/2017/10/counter-notice-does-not-fix-over-removal-online-speech.

93 "Manila Principles on Intermediary Liability," accessed May 27, 2018, https://www.manilaprinciples.org.

94 18 USC § 2258A.

95 18 USC § 2258D.

96 18 USC § 2258C.

97 18 USC § 2258A(f).

98 18 USC § 2239B.

99 In a January 31, 2018, ruling, the Ninth Circuit in *Fields v. Twitter* rejected a civil material support claim, based on a causation requirement that appears only in the civil material support laws, and declined to review Twitter's CDA 230 defense. Other cases are pending. For additional analysis, see Benjamin Wittes and Zoe Bedell, "Tweeting Terrorists, Part III: How Would Twitter Defend Itself against a Material Support Prosecution?," *Lawfare* (blog), February 14, 2016, accessed May 25, 2018, https://www.lawfareblog.com/tweeting-terrorists-part-iii-how-would-twitter-defend-itself-against-material-support-prosecution.

100 This example is drawn from experience, and I have used it in discussing this topic for years. Developments involving the YPG in 2017–2018 are entirely coincidental.

101 Because these are presumably not American speakers, the First Amendment analysis could be different from that discussed below.

102 "YouTube Terminates YPG Account for Second Time," Rudaw, August 10, 2017, accessed May 27, 2018, http://www.rudaw.net/english/middleeast/syria/08102017.

103 Anne Barnard and Ben Hubbard, "Allies or Terrorist: Who Are the Kurdish Fighters in Syria?," *New York Times,* January 25, 2018, accessed May 27, 2018, https://www.nytimes.com/2018/01/25/world/middleeast/turkey-kurds-syria.html.

104 Tuvan Gumrukcu, "U.S. to End Weapons Support for Syrian Kurdish YPG, Turkey Says," Reuters, January 27, 2018, accessed May 27, 2018, https://www.reuters.com/article/us-mideast-crisis-syria-turkey-usa/u-s-to-end-weapons-support-for-syrian-kurdish-ypg-turkey-says-idUSKBN1FG08W.

105 See Daphne Keller, Intermediary Liability 101, Center for Internet and Society, February 13, 2018, accessed June 4, 2018, http://cyberlaw.stanford.edu/blog/2018/02/intermediary-liability-101

106 See Preamble Paragraph 46 and Article 14(1)(b) of the eCommerce Directive.

107 Preamble Paragraph 47 and Article 15(1).

108 Preamble Paragraph 45 and Articles 12(3), 13(2), 14(3), and 18(1).

109 L'Oreal v. eBay ¶¶ 139, 142 (emphasis added); SABAM v. Netlog ¶¶ 38, 52; SABAM v. Scarlet ¶ 40, 54; eBay v. L'Oreal; and McFadden v. Sony.

110 Laurel Wamsley, "Austrian Court Rules Facebook Must Delete Hate Speech," NPR, May 8, 2017, accessed May 25, 2018, https://www.npr.org/sections/thetwo-way/2017/05/08/527398995/austrian-court-rules-facebook-must-delete-hate-speech (translating "miese Volksverräterin" and "korrupten Trampel").

111 Jack M. Balkin, "Old-School/New-School Speech Regulation," *Harvard Law Review* June 20, 2014, 2309, accessed May 25, 2018, https://harvardlawreview.org/2014/06/old-schoolnew-school-speech-regulation.

112 MTE v. Hungary at 86. Note that the ECHR, which is a Council of Europe body, interprets a different human rights instrument than the CJEU, a European Union body. Two of the four CJEU cases rejecting monitoring obligations also cited individual rights to expression and privacy as reasons, among a longer list, for the decisions. Scarlet Extended v. SABAM, SABAM v. Netlog. In another case, the CJEU endorsed a form of procedural protection in intermediary liability: it held that when internet service providers block websites pursuant to court orders, potentially blocking lawful content, internet users should have an avenue for judicial review to enforce their rights to access the information. Telekabel Wien v. Constantin.

113 Delfi v. Estonia. For a discussion of the intersection of the two ECHR cases and their practical consequences, see Daphne Keller, "New Intermediary Liability Cases from the European Court of Human Rights: What Will They Mean in the Real World?," Center for Internet and Society, April 11, 2016, accessed May 25, 2018, http://cyberlaw.stanford.edu/blog/2016/04/new-intermediary-liability-cases-european-court-human-rights-what-will-they-mean-real.

114 Natasha Lomas, "UK to Hike Penalties on Viewing Terrorist Content Online," TechCrunch, October 3, 2017, accessed May 28, 2018, https://techcrunch.com/2017/10/03/uk-to-hike-penalties-on-viewing-terrorist-content-online.

115 Chloe Farand, "France Scraps Law Making 'Regular' Visits to Jihadi Websites an Offense," *Independent,* February 10, 2017, accessed May 28, 2018, http://www.independent.co.uk/news/world/europe/france-overturns-law-regular-visits-jihadi-websites-islamists-anti-terror-legislation-a7573281.html.

116 Robert Hutton, "Tech Firms Face Fines Unless Terrorist Material Removed in Hours," *Bloomberg,* September 19, 2017, accessed May 28, 2018, https://www.bloomberg.com/news/articles/2017-09-19/tech-firms-face-fines-unless-terrorist-material-removed-in-hours.

117 European Commission, "Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms," September 28, 2017, accessed May 28, 2018, https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms. The Commission followed up with similar proposals in its March 2018 "Measures to Further Improve the Effectiveness of the Fight against Illegal Content Online," accessed May 25, 2018, https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2018-1183598_en.

118 European Commission, "Communication."

119 European Commission, "Communication."

120 Julia Reda, "MEPs from across the Political Spectrum Resist the Plan to Make 'Censorship Machines' Mandatory in the EU," JuliaReda.eu, January 22, 2018, accessed May 28, 2018, https://juliareda.eu/2018/01/censorship-machines.

121 European Commission, "Proposal for a Directive on the European Parliament and of the Council on Copyright in the Digital Single Market," September 14, 2016, accessed May 25, 2018, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=17200; Christina Angelopoulos, "EU Copyright Reform: Outside the Safe Harbours, Intermediary Liability Capsizes into Incoherence," *Kluwer Copyright Blog*, October 6, 2016, accessed May 25, 2018, http://copyrightblog.kluweriplaw.com/2016/10/06/eu-copyright-reform-outside-safe-harbours-intermediary-liability-capsizes-incoherence.

122 Cynthia Kroet, "Terrorism Fears on the Rise: EU Study," *Politico,* July 29, 2016, accessed May 25, 2018, https://www.politico.eu/article/terrorism-fears-on-the-rise-eu-study-terrorism-immigration.

123 No author, "Should Digital Monopolies Be Broken Up?," *Economist,* November 27, 2014, https://www.economist.com/news/leaders/21635000-european-moves-against-google-are-about-protecting-companies-not-consumers-should-digital.

124 Xeni Jardin, "Vietnam Complained of 'Toxic' Anti-government Facebook Content, Now Says Facebook Has Committed to Help Censor," BoingBoing, April 26, 2017, accessed May 28, 2018, https://boingboing.net/2017/04/26/vietnam-complained-of-toxic.html.

125 Ma Nguyen, "Vietnam Leverages Google, YouTube Hate Speech Failings," *Asia Times,* March 27, 2017, May 28, 2018, http://www.atimes.com/article/vietnam-leverages-google-youtube-hate-speech-failings.

126 Olga Razumovskaya, "Russia Proposes Strict Online Right to Be Forgotten," *Digits* (blog), *Wall Street Journal,* June 17, 2015, accessed May 25, 2018, https://blogs.wsj.com/digits/2015/06/17/russia-proposes-strict-online-right-to-be-forgotten; "Russia's 'Right to Be Forgotten' Bill Comes into Effect," *RT,* January 1, 2016, accessed May 25, 2018, https://www.rt.com/politics/327681-russia-internet-delete-personal.

127 Global Freedom of Expression, Columbia University, "LLC SIBFM v. Roskomnadzor," accessed May 28, 2018, https://globalfreedomofexpression.columbia.edu/cases/llc-sibfm-v-roskomnadzor.

128 See, e.g., Associated Press, "Turkey Pulls Plug on YouTube over Ataturk 'Insults,'" *Guardian,* March 7, 2007, accessed May 25, 2018, https://www.theguardian.com/world/2007/mar/07/turkey.

129 Bulman, "Facebook, Twitter and Whatsapp Blocked in Turkey."

130 "G7 Taormina Statement on the Fight against Terrorism and Violent Extremism," accessed May 28, 2018, G7Italy.it, http://www.g7italy.it/sites/default/files/documents/G7 Taormina Statement on the Fight Against Terrorism and Violent Extremism_0.pdf.

131 "The Hamburg G20 Leaders' Statement on Countering Terrorism," G20 Information Centre, University of Toronto, July 7, 2017, accessed May 28, 2018, http://www.g20.utoronto.ca/2017/170707-counterterrorism.html.

132 521 U.S. § 844 (1997).

133 137 S. Ct. 1730 (2017).

134 Cass Sunstein, "Islamic State Challenges Free-Speech Laws," *Commercial Appeal,* November 28, 2015, accessed May 28, 2018, http://archive.commercialappeal.com/opinion/national/cass-sunstein-islamic-state-challenges-free-speech-laws-253cd8b9-679e-7394-e053-0100007fd402-356486171.html.

135 Pp. 1737, 1738.

136 *Packingham* at 1378, internal quotation omitted. The court also chided lawmakers' overreaction to new internet technologies: "For centuries now, inventions heralded as advances in human progress have been exploited by the criminal mind. New technologies, all too soon, can become instruments used to commit serious crimes. The railroad is one example . . . and the telephone another. So it will be with the Internet and social media" (internal citation omitted, p. 1376).

137 The analysis in this section pertains to social networks and similar user-facing "edge" online service providers. For infrastructure providers such as ISPs or content delivery networks, which function more like common carriers, analysis is both legally different and currently highly politicized because of net neutrality debates.

138 Smith v. California, 361 U.S. 147 (1959); Bantam Books, Inc. v. Sullivan, 372 U.S. 58 (1963).

139 *Smith* at 153.

140 *Smith* at 154.

141 *Bantam Books* at 64n6.

142 *Bantam Books* at 70.

143 337 F. Supp. 2d 606 (E.D. Penn. 2004).

144 376 U.S. 254 (1964) at 277.

145 *Sullivan* at 266.

146 776 F. Supp. 135 (S.D.N.Y. 1991).

147 In that case, the false claim would have been about the speech's audience and whether it included minors, as opposed to the speech's content. 521 U.S. 844 (1997) at 880.

148 129 F.3d 327 (4th Cir. 1997) at 333.

149 521 U.S. 844, 874 (1997).

150 See discussion in Daphne Keller, "SESTA and the Teachings of Intermediary Liability," Center for Internet and Society, November 2, 2017, accessed May 25, 2018, https://cyberlaw.stanford.edu/publications/sesta-and-teachings-intermediary-liability.

151 *Bantam Books* at 66; see also *CCBill* at 1109, *Telekabel* at §75–76 (CJEU), *Cartier* (UK).

152 Marco Civil (Brazil), "Marco Civil da Internet: 'Brazilian Civil Rights Framework for the Internet,'" Wilmap, April 23, 2014, accessed May 25, 2018, http://wilmap.law.stanford.edu/entries/marco-civil-da-internet-brazilian-civil-rights-framework-internet; Copyright Act (Chile), "Amending the Intellectual Property Law," Wilmap, May 4, 2010, accessed May 25, 2018, http://wilmap.law.stanford.edu/entries/law-no-20435-may-04-2010-amending-intellectual-property-law; Shreya Singhal (Indian Supreme Court), "*Shreya Singhal v. Union of India,* No. 167/2012," Wilmap, March 24, 2015, accessed May 25, 2018, http://wilmap.law.stanford.edu/entries/supreme-court-criminal-shreya-singhal-v-union-india-no-1672012-march-24-2015.

153 *Belen-Rodriguez* (Argentine Supreme Court); *Dr. Royo* (Spanish lower court); *Davison* (UK lower court).

154 See also Daphne Keller, "Problems with Filters in the European Commission's Platforms Proposal," Center for Internet and Society, October 5, 2017, accessed May 25, 2018, http://cyberlaw.stanford.edu/blog/2017/10/problems-filters-european-commissions-platforms-proposal.

155 See Section II.B.2 on law outside the United States.

156 17 USC § 512(m); 18 USC § 2258A(f).

157 See Jacobellis v. Ohio, 378 U.S. 184 at 197 (Justice Stewart concurring).

158 Courts, on the other hand, may strike down laws because of them. See, e.g., *Ashcroft v. Free Speech Coalition.*

159 This discussion focuses on Islamist terrorism, in line with much of today's public discussion. But much of the analysis could apply equally to other violent extremists, such as German or American neo-Nazis.

160 *Terrorism and Social Media: #IsBigTechDoingEnough? Hearing before the Senate Commerce Comm.* (January 17, 2018).

161 European Commission, "Code of Conduct on Countering Illegal Hate Speech Online: Results of the 3rd Monitoring Exercise," January 2018, accessed May 28, 2018, http://ec.europa.eu/newsroom/just/document.cfm?doc_id=49286.

162 Holder v. Humanitarian Law Project, 561 U.S. 1, 28 (2010) (internal quotations omitted). This case upheld the material support laws, which prohibit services including training and education, against a First Amendment challenge.

163 J. M. Berger and Jonathon Morgan, "The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter," Brookings Project on U.S. Relations with the Islamic World Analysis Paper No. 20 (March 2015), 3, accessed May 25, 2018, https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf. Research favoring more content removal includes Martyn Frampton, Ali Fisher, and Nico Prucha, "The New Net War: Countering Extremism Online," 2017), 70–77, accessed May 25, 2018, https://policyexchange.org.uk/wp-content/uploads/2017/09/The-New-Netwar-2.pdf; and materials published by Mark Wallace at counterextremism.com.

164 Berger and Morgan, "ISIS Twitter Census," 53. A peer-reviewed Rand study in 2013 assessed 150 articles about online radicalization and found that only eighteen were empirically derived. Ines von Behr, Anaïs Reding, Charlie Edwards, and Luke Gribbon, "Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism" (Rand Europe, 2013), 16, accessed May 25, 2018, https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf. The Rand report reviewed five common claims from the existing literature and concluded that while most tracked empirical correlations between internet extremist content and individual radicalization, none explained causation in a way that would help identify effective interventions.

165 Sageman 2014, quoted in *The Stagnation in Terrorism Research* (2014).

166 Alexander Meleagrou-Hitchens and Nick Kaderbhai, "Research Perspectives on Online Radicalisation: A Literature Review, 2006–2016" (VOX-Pol, 2017), 35.

167 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 19, 39 ("The majority of the literature takes a nuanced position that asserts the importance of online influences without negating the requirement of offline interactions"); von Behr et al, "Radicalisation in the Digital Era," xii (evidence "does not support the suggestion that the internet has contributed to the development of self-radicalisation" or "that the internet is replacing the need for individuals to meet in person during their radicalisation process. Instead, the evidence suggests that the internet is not a substitute for in-person meetings but, rather, complements in-person communication.").

168 Hamed el-Said and Richard Barrett, "Enhancing the Understanding of the Foreign Terrorist Fighters Phenomenon in Syria" (UN Office of Counter Terrorism, July 2017), 39, accessed May 25, 2018, https://www.academia.edu/34134270/Enhancing_the_Understanding_of_the_Foreign_Terrorist_Fighters_Phenomenon_in_Syria.

169 A "strong case for a causal connection between such materials, and violent acts perpetrated by those found to have been in possession of them, has yet to be made." Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 36.

170 Kim Cragin, Melissa A. Bradley, Eric Robinson, and Paul S. Steinberg, "What Factors Cause Youth to Reject Violent Extremism? Results of an Exploratory Analysis in the West Bank" (Rand, 2015), 16, accessed May 25, 2018, https://www.rand.org/pubs/research_reports/RR1118.html; von Behr et al., "Radicalisation in the Digital Era," iii ("Many of the policy documents and academic literature in this area focus on online content and messaging, rather than exploring how the internet is used by individuals in the process of their radicalisation.").

171 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 56.

172 Berger and Morgan, "ISIS Twitter Census," 3.

173 Peter R. Neumann, "Options and Strategies for Countering Online Radicalization in the United States," *Studies in Conflict and Terrorism* 36, no. 6 (January 2013): 437, accessed May 25, 2018, http://www.tandfonline.com/doi/pdf/10.1080/1057610X.2013.784568.

174 Gilles de Kerchove, review of *Radicalized: New Jihadists and the Threat to the West,* by Peter Neumann, Amazon, accessed May 25, 2018, https://www.amazon.com/Radicalized-New-Jihadists-Threat-West-ebook/dp/B01LX47YEN/ref=sr_1_1?s=books&ie=UTF8&qid=1516740790&sr=1-1.

175 See Malachy Browne, "YouTube Removes Videos Showing Atrocities in Syria"; Edwards, "When YouTube Removes Violent Videos."

176 Julia Carrie Wong, "Facebook Blocks Chechnya Activist Page in Latest Case of Wrongful Censorship," *Guardian,* June 6, 2017, accessed May 28, 2018, https://www.theguardian.com/technology/2017/jun/06/facebook-chechnya-political-activist-page-deleted.

177 Betsy Woodruff, "Facebook Silences Rohingya Reports of Ethnic Cleansing," Daily Beast, September 18, 2017, accessed May 28, 2018, https://www.thedailybeast.com/exclusive-rohingya-activists-say-facebook-silences-them.

178 "Facebook Bans 'Dangerous' Rohingya Militant Group," Hindu, September 21, 2017, accessed May 28, 2018, http://www.thehindu.com/news/international/facebook-bans-dangerous-rohingya-militant-group/article19727205.ece.

179 Maeghin Alarid, "Recruitment and Radicalization: The Role of Social Media and New Technology," in M. Hughes & M. Miklaucic, *Impunity: Countering Illicit Power in War and Transition,* Washington, DC: Peacekeeping and Stability Operations Institute (PKSOI), 2016, 313–330 (discussing ISIS recruiters "specifically targeting those who are young and computer savvy").

180 Alex Marthews and Catherine E. Tucker, "Government Surveillance and Internet Search Behavior," February 17, 2017, 40, accessed May 25, 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2412564; see also "Chilling Effects: NSA Surveillance Drives U.S. Writers to Self-Censor" (Pen America, November 12, 2013), 6–7, accessed May 25, 2018, https://pen.org/chilling-effects (journalists report avoiding writing about terrorism); Jon Penney, "Chilling Effects: Online Surveillance and Wikipedia Use," *Berkeley Technology Law Journal,* 31, no. 1 (2016):172, accessed May 25, 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645.

181 "PragerU Takes Legal Action."

182 Ariana Tobin, Madeleine Varner, and Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," ProPublica, December 28, 2017, May 28, 2018, https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

183 Some thinkers see greater platform transparency and appellate procedures as the highest priority in improving problems of this sort. They are right that these would be improvements. But the goal of moving to ever-more-robust and complex private speech governance systems warrants close examination.

184 Kevin Bankston and Liz Woolery, "We Need to Shine a Light on Private Online Censorship," Techdirt, January 31, 2018, accessed May 28, 2018, https://www.techdirt.com/articles/20180130/22212639127/we-need-to-shine-light-private-online-censorship.shtml; Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children," ProPublica, June 28, 2017, accessed May 28, 2018, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms; Content Moderation and Removal at Scale, conference, Santa Clara Law, February 2, 2018, Santa Clara, CA, accessed May 28, 2018, http://law.scu.edu/event/content-moderation-removal-at-scale; Sarah T. Roberts, "Research Proposal," Luskin Center for History and Policy, UCLA, accessed May 28, 2018; Online Censorship website, accessed May 28, 2018, https://onlinecensorship.org; Eric Goldman, "Facebook Wins Appeal over Allegedly Discriminatory Content Removal: Sikhs for Justice v. Facebook," *Technology & Marketing Law Blog,* September 13, 2017, accessed May 28, 2018, https://blog.ericgoldman.org/archives/2017/09/facebook-wins-appeal-over-allegedly-discriminatory-content-removal-sikhs-for-justice-v-facebook.htm.

185 See Mark Scott and Mike Isaac, "Facebook Restores Iconic Vietnam War Photo It Censored for Nudity," *New York Times,* September 10, 2016, accessed May 25, 2018, https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html.

186 See Mike Isaac and Daisuke Wakabayashi,, "Twitter's Panic after Trump's Account Is Deleted Caps a Rough Week," *New York Times,* November 3, 2017, https://www.nytimes.com/2017/11/03/technology/trump-twitter-deleted.html.

187 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 14; Emily Dreyfuss, "Blaming the Internet for Terrorism Misses the Point," *Wired,* June 6, 2017, accessed May 25, 2018, https://www.wired.com/2017/06/theresa-may-internet-terrorism; Alarid, "Recruitment and Radicalization," 314 ("Radicalization is more widespread where conditions of inequality and political frustration prevail").

188 Peter Neumann, "Options and Strategies for Countering Online Radicalization," 435 (citing Sageman); Alarid, "Recruitment and Radicalization."

189 Berger and Morgan, "ISIS Twitter Census," 56.

190 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 7.

191 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 63.

192 Berger and Morgan, "ISIS Twitter Census," 3. See also Neumann, "Options and Strategies for Countering Online Radicalization," 447 (discussing counterspeech); Alarid, "Recruitment and

Radicalization" (negative social media posts about ISIS can be "an effective tool in counterradicalization efforts"); Susan Benesch, "Countering Dangerous Speech to Prevent Mass Violence during Kenya's 2013 Elections" (Dangerous Speech Project, February 10, 2014), accessed May 25, 2018, https://dangerousspeech.org/countering-dangerous-speech-kenya-2013 (speech believed to be correlated to violence during Kenyan election overrepresented in closed Facebook discussion compared to Twitter).

193 Kevin Munger, "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment," *Political Behavior,* 39, no. 3 (September 2017): 629–49, accessed May 28, 2018, https://link.springer.com/article/10.1007/s11109-016-9373-5.

194 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 60–63.

195 Meleagrou-Hitchens and Kaderbhai, "Research Perspectives on Online Radicalisation," 62–63.

196 Elizabeth Cohen and Debra Goldschmidt, "Ex-terrorist Explains How to Fight ISIS Online," CNN, December 21, 2015, accessed May 25, 2018, https://www.cnn.com/2015/12/18/health/al-quaeda-recruiter-fight-isis-online/index.html.

197 Damian Whitworth, "How the War on Terror Is Raging Online," *London Times,* November 8, 2017, accessed May 25, 2018, https://www.thetimes.co.uk/article/how-the-waron-terror-is-ragingonline-fgx8zr58f.

198 "Reality Check: What Has Happened to Police Numbers," BBC, May 26, 2017, accessed May 28, 2018, http://www.bbc.com/news/election-2017-40060677.

199 Dreyfuss, "Blaming the Internet for Terrorism."

200 Peter Neumann and Shiraz Maher, "London Attack: How Are UK Extremists Radicalised?," BBC, June 5, 2017, accessed May 25, 2018, http://www.bbc.com/news/uk-40161333?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axiosam&stream=top-stories.

201 Neumann and Maher, "London Attack."

202 Berger and Morgan, "ISIS Twitter Census," 53. The third key question, the authors concluded, is whether it is "ethical to suppress political speech, even when such speech is repugnant."

203 Ellen Nakashima, "Dismantling of Saudi-CIA Web Site Illustrates Need for Clearer Cyberwar Policies," *Washington Post,* March 19, 2010, accessed May 25, 2018, http://www.washingtonpost.com/wp-dyn/content/article/2010/03/18/AR2010031805464.html.

204 Jeff Kosseff, personal website, accessed May 25, 2018, https://www.jeffkosseff.com.

205 47 USC § 230(b)(2).

206 Urban and Quilter, "Efficient Process or 'Chilling Effects'?," 651.

207 "Google Search Removals Due to Copyright Infringement FAQs," Google, accessed May 28, 2018, https://support.google.com/transparencyreport/answer/7347743?hl=en.

208 Ty McMahan, "Veoh Lives On: Behind the Acquisition of the Video Site," *Wall Street Journal,* April 7, 2010, accessed May 25, 2018, https://blogs.wsj.com/venturecapital/2010/04/07/veoh-lives-on-behind-the-acquisition-of-the-video-site.

209 Eliot van Buskirk, "Veoh Files for Bankruptcy after Fending off Infringement Charges," *Wired,* February 12, 2010, accessed May 28, 2018, https://www.wired.com/2010/02/veoh-files-for-bankruptcy-after-fending-off-infringement-charges.

210 David Kravets, "Google Wins Viacom Copyright Lawsuit," *Wired,* June 23, 2010, accessed May 25, 2018, https://www.wired.com/2010/06/dmca-protects-youtube.

211 Matthew Le Merle, Raju Sarma, Tashfeen Ahmed, and Christopher Pencavel, "The Impact of U.S. Internet Copyright Regulations on Early-Stage Investment: A Quantitative Study" (Booz & Company), accessed May 28, 2018, https://www.strategyand.pwc.com/media/uploads/Strategyand-Impact-US-Internet-Copyright-Regulations-Early-Stage-Investment.pdf.

212 Germany's law, interestingly, exempts social networks with fewer than two million registered users in the country. Act to Improve Enforcement of the Law in Social Networks, December 7, 2017, accessed May 24, 2018, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=099B20F1854667953970FB02EE7A0510 .1_cid334?__blob=publicationFile&v=2.

213 Peter Ray Allison, "Content Filtering a Potential Challenge in Digital Single Market," *Computer Weekly,* December 2017, accessed May 25, 2018, http://www.computerweekly.com/feature/Content-filtering-a-potential-challenge-in-Digital-Single-Market.

214 Urban, Karaganis, and Schofield, "Notice and Takedown," 2.

215 Engstrom and Feamster, "The Limits of Filtering."

216 Urban, Karaganis, and Schofield, "Notice and Takedown," 31.

21 20 19 18 5 4 3 2

The preferred citation for this publication is Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money,* Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1807 (June 13, 2018; revised June 28, 2018), available at https://lawfareblog.com/internet-platforms-observations-speech-danger-and-money.

About the Author

PHOTO BY AMANDA AVILA

DAPHNE KELLER

Daphne Keller is the Director of Intermediary Liability at the Stanford Law School Center for Internet and Society. Her work focuses on the ways that laws governing platforms' responsibility for online information affect the rights of Internet users. She was previously Associate General Counsel to Google, where she led the web search legal team.

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