CHAPTER V: COMPETITION IN THE MARKETPLACE OF IDEAS — DIVERSITY, CENSORSHIP, AND THE PROBLEM OF CONTENT MODERATION.

Ever since the first comprehensive regulation of broadcast radio under the Federal Radio Act of 1927, we have struggled to protect free expression while protecting the public from the harms of amplifying misinformation and hate speech. A 1939 *Columbia Law Review* note on "Radio Censorship and the Federal Communications Commission"[100] observes the concern with limiting hate speech and fraud versus the concern that the FCC effectively censored controversial speech through its policy of invoking renewal hearings based on broadcast content (including commercials). It also surveys concerns about favoritism to the administration in power and the need to enhance minority representation. A reader familiar with efforts by members of Congress to push Google, Facebook, and Twitter to moderate content — from terrorist recruitment to supposed bias against conservative viewpoints — would find the article's concerns equally applicable to today's social media platforms. Likewise, many of the proposed solutions would have a familiar ring, such as government ownership of radio stations, promotion of programming about controversial issues, and treating radio stations as state actors for First Amendment purposes, subject to strict scrutiny for content-based censorship.

Indeed, the caution that government censorship of "bad" content can be used to suppress information that we now would consider protected by the First Amendment and even beneficial to the public, predates electronic media. In 1873 Congress passed the Comstock Act to criminalize the dissemination of "obscene literature and immoral use." This included birth control and materials relating to safe abortion. The Comstock Act thus became a tool to threaten advocates of women's reproductive rights.

At the same time, the evidence has become overwhelming that malign actors are using platforms to disseminate content designed to interfere with the function of democracy, intimidate or harass individuals or targeted groups, or otherwise disrupt society. Existing law has proven incapable of addressing these harms. Nor have platforms found effective ways to police themselves. To the contrary, when platforms undertake efforts to address harmful content, they have invariably provoked criticism that they are over-inclusive, under-inclusive, biased against progressive causes in favor of white supremacist hate speech, biased against conservatives in favor of liberals, or simply do not care because they profit from user engagement (Klonick 2018; Van Zuylen-Wood 2019). Processes vary from platform to platform. Users have often complained that platforms are unresponsive to complaints, that the guidelines for each platform's "community standards" are confusing, and that platforms may have either no process to appeal a takedown decision or

[100] Note, "Radio Censorship and the Federal Communications Commission," 39 Col. L. Rev. 447 (1939).

confusing processes filled with lengthy delays (Heins 2019). This is true not merely for social media platforms such as Facebook and Twitter. On Amazon, it has become common for commercial rivals using the platform to manipulate the rules to shut out rivals through such means as planting fake reviews on a rival's product page, then complaining to Amazon about the fake reviews and triggering a takedown of the rival merchant (Dzieza 2018).

Without a set of laws clearly delineating the rights and responsibilities of platform operators and users, as well as criminalizing genuinely harmful conduct such as fraud and harassment, it is difficult to see how platforms can function effectively going forward. Worse, as observed in the controversy over the 2016 presidential election, failure to address this problem threatens the function of our underlying democracy. Whether or not Russian interference tipped the balance in that election, it cast a shadow over the election's legitimacy and reinforced hyper-partisan divisions and racial tensions.

It is not my purpose here to thoroughly examine the proper balance between the spirit of the First Amendment and the very real harms we are seeing in the form of ethnic cleansing, harassment, misinformation campaigns designed to promote distrust and influence elections, and other serious problems. To do the matter justice requires a great deal more discussion and debate, particularly with individuals and communities that have been the targets of overbroad censorship and the targets of harassment campaigns. Nevertheless, some significant discussion is necessary because any comprehensive sector-specific regulation cannot ignore these issues. Addressing issues in the age of electronic media that we have recognized for almost a century as central to our democracy must be embedded in the DNA of platform regulation.

A. Defining the Problem: Discouraging "Bad" Content While Promoting "Good" Content.

I therefore propose to set out in this section a basic framework for addressing the general category of "content moderation." First, we must consider which problems we believe require a policy solution. These divide into two almost opposite categories. In the first instance, we wish to restrict (and possibly punish) the dissemination of "bad" content. "Bad" content can include content long punished under the common law or falling outside the First Amendment, such as false advertising, fraud, and threats of physical violence or harassment. But it can include other forms of content hurtful to individuals but not necessarily illegal (such as "revenge porn"), ideas broadly considered harmful to society (such as racist or misogynist ideologies, "fake news," or conspiracy theories), or activities harmful to individuals, such as "cyberbullying." Some of these ideas and expressions enjoy various levels of First Amendment protection, setting them beyond direct government prohibition. Others fall into a gray area, where the conduct arguably could be

criminalized or subject to civil penalty. Some content may be considered unsuitable for children but acceptable for adults. For purposes of policy consideration, however, we can characterize this generally as the problem of limiting (or providing incentives for platforms to limit) the creation and/or dissemination of "bad" content.

The second general category is the creation and dissemination of "good" content. Although this is much less of a public concern, the question of how to promote various forms of "good" content has been a central policy question in electronic media for nearly a century. Examples of "good" content include educational content (particularly for children), news about local, national, and international affairs necessary for self-government, and representation in entertainment of a broad and diverse population mirroring the diversity of our society as a whole. This "diversity rationale" has at times produced policies designed to create greater opportunities for the public to be exposed to diverse and competing viewpoints, and at other times has included mandates to create certain types of content.

Although the last century of regulation of electronic media offers valuable lessons both as to what our goals should be in moderating content and which methods might be useful or harmful, there are important differences between digital platforms, telecommunications networks, and media of mass communication. In the world of electronic communications, we could easily distinguish between direct communication between one individual and another (telecommunications), and mass communication from one to many (broadcasting and cable). Platforms blend these concepts seamlessly. In addition, platforms create an entirely new form of communication, self-organizing many-to-many communities.[101] While the gatekeeper function of the platform mirrors the gatekeeper function of the telephone network or the cable system, it is no longer straightforward to develop a set of rules governing telecommunications on the one hand and mass media on the other.

This raises important questions that must inform any proposed regulatory regime. Is the goal of moderating "bad" content to shield people from content they find hurtful or objectionable, or to ban the production of harmful content entirely? Do we seek to punish those who produce "bad" content or who engage in certain types of behavior? For example, should it be illegal to use any digital platform to advance racist concepts such as "the need for a white ethnostate" or "how Jews control the world?" Or should it be difficult to find such content, so that only those who actively seek it can participate in it? Should platforms monitor messaging functions to prevent the spread of potentially violent racist polemics? At what point do group conversations move beyond a typical

[101] The concept of self-organizing many-to-many communities, while technically possible in the world of telecommunications, is highly limited. By contrast, the ability to form self-organized communities is a core function of many platforms.

small-scale conversation and become something more global and potentially more harmful? What if the problem is not violence, but indecency?

Similarly, when considering how to promote "good" content, we must decide where the public interest in promoting diversity of content ends and becomes counterproductive, or even an exercise in state propaganda. Studies of media literacy demonstrate the value of exposing people to diverse content and avoiding "silos" that reinforce stereotypes and hyper-partisan political views (Sunstein 2018). Yet there are limits. People have a right to decide that they enjoy particular types of entertainment or favor a particular political perspective and cannot be forced to view content they do not like (or find offensive) as the price of getting their desired content. In addition, recent studies show that forcing people to read or listen to diametrically opposed views can backfire and actually reinforce a person's pre-existing views (Benkler, Farris, and Roberts 2018; Klein 2018).

Finally, factors that encourage production of "good" content may also encourage production of "bad" content. For decades, advocates have debated whether online anonymity is an important protection for speakers or a shield for wrongdoers. The reality, of course, is both. The question of whether to favor those who require anonymity to speak honestly, explore new possibilities, and create, versus suppressing the ability of those engaged in hurtful or illegal activity to hide behind anonymity, is not amenable to some sort of logical proof or mathematical balancing. It is invariably a judgment by whomever the law empowers to make that decision.

There is no single right answer to any of these questions. Resolving them inevitably requires line-drawing, resulting in some arbitrary outcomes. But the value of any framework is found in its overall contribution, not in the most difficult edge cases. I propose the following framework to guide further debate, understanding that ultimately, drawing lines will require societal and legislative consensus.

B. A Basic Framework for Moderating Harmful Content.

The Communications Act contains multiple provisions for policing conduct deemed harmful for a variety of reasons. In the realm of telecommunications, we criminalize fraud,[102] harassment,[103] and unwanted robocalls, texts, or faxes.[104] But efforts to criminalize indecent content (or to force telephone networks or cable operators to take steps to prevent exposing minors to indecent content) are found to violate the First Amendment. In the realm of electronic communications, the policy and First Amendment analysis become even more complicated. Depending on the medium

[102] 18 U.S.C. §1343.
[103] 47 U.S.C. §223.
[104] 47 U.S.C. §227.

and the nature of the regulation, the Supreme Court has used a variety of tests with regard to prohibiting content on broadcast or cable. The Court has also addressed certain types of limitations on internet content generally and on access to social media.

Because laws addressing moderation of expressive conduct raise concerns under the First Amendment, I begin with a short overview of the doctrines relevant to regulation of "bad speech" in electronic media. This includes the commercial speech standard, the various levels of First Amendment "scrutiny," and doctrines that permit regulation of speech despite a First Amendment interest on the part of either the speaker or the platform. I will next discuss the advantages and disadvantages of relying on the platforms themselves to police content, either through extra-legal social pressure or through imposition of legal incentives and penalties such as civil liability. Finally, I will make several broad recommendations for approaches consistent with the Constitution and policy, recognizing that the details of specific language require careful drafting when translating these into statutory language.

1. Direct Government Regulation of Content and the Confusing Question of First Amendment Standards: Commercial Speech Doctrine, Strict versus Intermediate Scrutiny, Reasonable Time and Space Restrictions, "Intrusiveness," and Other Mitigating Doctrines.

Unsurprisingly, some forms of "bad" content have proven easier to regulate than others. In general, content that violates common-law torts, furthers criminal activity, or is otherwise unfair and deceptive does not qualify for protection under the First Amendment. For example, content that is deceptive or clearly causes consumer harm, such as false advertising, generally enjoys no constitutional protection. Similarly, content designed to intimidate through the threat of violence or other unlawful actions, false statements about individuals that damage their reputation, or clear steps to engage in a criminal enterprise have long been prohibited by common law and statute. The difficulty in policing these kinds of "bad" content lies primarily in distinguishing between prohibited content and content which, while socially harmful or distasteful, enjoys First Amendment protection.

Even where speech is protected, however, not all speech is protected equally under the First Amendment. Much commercial activity relates in some way to speech. If the First Amendment prevented regulation of all commercial activity that somehow related to speech, virtually no commercial regulation would be possible. This is especially true with regard to online businesses, where information is constantly being collected, manipulated, and distributed as part of the normal course of business. The law distinguishes between mere commercial activity subject to regulation and commercial speech. Furthermore, while commercial speech enjoys some protection under the

First Amendment, it does not enjoy the same level of protection as non-commercial speech.[105] Under the commercial speech/*Central Hudson* test, a regulation that affects commercial speech (assuming the speech is neither misleading nor related to unlawful activity) survives if (a) the regulation serves a "substantial" government interest; (b) the regulation directly advances the government interest; and, (c) the regulation "is not more extensive than necessary" to serve that interest.

The inquiry does not end with the distinction between commercial and non-commercial speech. Often the Supreme Court employs an additional layer of analysis, particularly when it comes to electronic media. Traditionally, broadcast media receive little direct First Amendment protection. As long as there is a rational (and content-neutral) reason for limiting or requiring speech by a broadcaster, the regulation will survive First Amendment scrutiny. The next highest level of scrutiny is "intermediate scrutiny." This applies to platforms such as cable providers that do not generally "speak" to their customers but play a role in selecting the content available on the platform. Intermediate scrutiny requires that the regulation advances an important government interest unrelated to the suppression of speech and does not burden more speech than is necessary.

Laws regulating newspapers or otherwise directly regulating speech or the press must survive "strict scrutiny" under the First Amendment. Strict scrutiny is an extremely difficult hurdle to overcome. To survive strict scrutiny, a regulation must serve a "compelling" government interest, and the regulation must be "narrowly tailored" to serve the compelling interest. This standard is sufficiently hard to meet that it is often described as "strict in theory, fatal in fact."

Additionally, all laws that in some way regulate speech must be "content neutral." Defining "content neutral" is difficult, since the purpose of a law that addresses speech in some way is to affect "content" and is therefore not "content neutral" in the conventional sense. "Content neutral" in this sense generally means not favoring or disfavoring speech because of its point of view. For example, in traditional media regulation, government efforts to promote local news production by limiting the number of media outlets a broadcaster may own in a local market, or by requiring cable operators to carry local broadcast stations even when the cable operator does not wish to carry them, are content neutral because the government is not favoring a particular viewpoint (e.g., Republican or Democratic, liberal or conservative). While it can be argued that the government is favoring one perspective over another in the conventional sense, i.e., local over non-local, the government is not providing or prohibiting specific content (Sunstein 1994). Similarly, when the government bans "harassing" or "threatening" speech, it is not making a judgment as to whether specific ideas are good or bad in themselves. To survive as content neutral, regulation of harassing

[105] *See Central Hudson Gas and Electric v. Public Svc. Comm'n* 447 U.S. 557 (1980) ("*Central Hudson*").

or threatening speech judges whether the intent is to cause fear or emotional distress, rather than to expose people to "bad ideas."[106]

Finally, in drafting potential laws relating to content moderation, we must consider two somewhat related doctrines that inform the First Amendment analysis. The Supreme Court has found that the government can impose "reasonable time and space restrictions" on expression and can protect members of the public from unwanted or unusually intrusive speech. For example, the First Amendment does not prohibit the government from imposing noise limits in residential neighborhoods or banning trucks with loudspeakers playing during residential sleeping hours.[107] The First Amendment permits the government to regulate "robocalls" and "robofaxes."[108] The First Amendment permits restricting "adult" entertainment to particular areas (or excluding them from particular areas) as a consequence of their secondary effects on property values and concerns about crime.[109]

Perhaps most relevant, the Supreme Court has found the prohibition on broadcasting "indecent" content constitutional, in *FCC v. Pacifica Foundation*.[110] There, the Supreme Court found constitutional a warning issued by the FCC against a radio station for playing a comedy monologue by comedian George Carlin called "Filthy Words." The Court found the content "indecent" rather than obscene, and therefore subject to protection under the First Amendment. Although the *Pacifica* decision noted in passing that broadcasting enjoys a lower measure of First Amendment protection than any other medium, the decision did not rely on this distinction. Rather, the Court found that the uniquely "pervasive" nature of broadcasting (at least at the time of the decision) made it effectively impossible to keep unwanted indecent speech out of the home. This uniquely "pervasive" and "intrusive" quality of broadcasting posed a particular challenge to parents trying to shield their children from exposure to indecent content. Analogizing congressional and FCC regulation of indecent broadcast content to nuisance law, the Court found in *Pacifica* that punishing broadcast of indecent content at times when children were likely to be in the audience did not violate the First Amendment.

It is therefore possible, at least in theory, that restrictions on certain types of content on digital platforms would pass muster either as a means of blocking unwanted, intrusive speech or as "reasonable time and place" restrictions. As demonstrated by Congress's difficulties prohibiting or

---

[106] *Compare Ashcroft v. Free Speech Coalition*, 535 U.S. 234 (2002) (Congress cannot criminalize "virtual" child pornography on grounds that doing so normalizes actual child pornography) *with Virginia v. Black*, 538 U.S. 343 (2003) (cross-burning can be prohibited if done with 'intent to intimidate,' but cannot itself be *prima facie* evidence of intent to intimidate).
[107] *Kovacs v. Cooper*, 336 U.S. 77 (1949).
[108] *Moser v. FCC*, 46 F.3d 970 (Ninth Cir. 1995).
[109] *City of Los Angeles v. Alameda Books, Inc.*, 535 U.S. 425 (2002); *Renton v. Playtime Theaters*, 475 U.S. 41 (1986).
[110] 438 U.S. 726 (1978).

otherwise limiting indecent content in other areas of electronic communications, and the FCC's subsequent trouble enforcing indecency regulation in broadcast, employing these doctrines effectively is extremely difficult in practice. The Supreme Court has generally found that where a party takes "affirmative steps" to bring content into the home, then the indecent content cannot be considered "intrusive" and should receive First Amendment protection.

United States v. Playboy Entertainment Group[111] provides an instructive example of how these doctrines interact, and therefore a useful roadmap for drafting content-moderation regulations applicable to digital platforms. *Playboy* dealt with a provision of the Communications Decency Act (itself part of the Telecommunications Act of 1996) requiring cable operators to take steps to prevent the intrusion of unwanted sexually oriented programming that was considered merely indecent, rather than obscene.[112] In addition to limiting access by children to indecent programming directly, Congress sought to address a problem with analog cable systems known as "signal bleed," where a blocked channel's content was nevertheless partially available in some comprehensible form due to the provider's inability to screen it out completely. Section 504 required a cable operator offering adult-oriented indecent programming channels to "fully block" such channels at the request of the cable subscriber. Section 505 additionally required that cable operators "fully block" channels dedicated to sexually oriented indecent programming, "so that one not a subscriber to the channel does not receive it." The definition of "fully block" was designed to include blocking any signal bleed. Where cable operators could not technically comply with the blocking requirement, the statute required them to limit transmission of indecent material to the same "safe harbor" hours as for broadcasting, 10 p.m. to 6 a.m., when young children are presumed not to be watching.

In analyzing the constitutionality of these provisions, the Court agreed with the government that cable programming, like broadcasting, "comes into the home uninvited" and could therefore be regulated in a manner similar to broadcast indecency. However, because Sections 504 and 505 directly targeted speech based on its content, it nevertheless required strict scrutiny. The Court distinguished its zoning cases because the statute was not designed to address secondary effects of constitutionally protected speech. The Court then found that the requirement to limit indecent programming to the "safe harbor" hours imposed an impermissible burden on adults wishing to receive the indecent protected content. The Court then concluded that the government failed to show that Section 505 was the "least restrictive" means of achieving the goal of keeping unwanted indecent material out of the home. By contrast, Section 504, which required cable operators to "fully

[111] 529 U.S. 803 (2000).
[112] The distinction is important for First Amendment purposes. Obscene material enjoys no protection under the First Amendment and may be criminalized (and is prohibited by law from transmission on cable systems). Indecent content is protected by the First Amendment. Discussion of the distinctions between obscene content and indecent content are beyond the scope of this paper.

block" such channels at the request of the cable subscriber, did survive scrutiny as the "least restrictive" means of allowing subscribers to block the unwanted content.

Several other elements in the analysis are worthy of note. First, the Court faulted Congress for its failure to compile a substantial record as to the nature of the problem and to explain why the universal blocking requirement of Section 505 was necessary, rather than the less restrictive means of requiring subscribers to request blocking of specific channels. The Court dismissed the argument that because few subscribers availed themselves of this remedy, the remedy was ineffective. As the Court stressed, the general rule for handling unwanted content is to require people who wish to avoid unwanted content to do so, rather than to silence speakers. Even when the unwanted speech is intrusive and therefore the government may take steps to assist individuals in blocking unwanted content, it is appropriate to require some action on the part of the individual wishing to block the speech rather than placing the entire burden on the controversial speaker.

Finally, when Congress or a regulator does regulate on the basis of the specific content of the speech, strict scrutiny will always apply even if the speech interest would be considered "weak" under intermediate scrutiny. As the Court explained in its *Playboy* decision, when content is directly targeted based on its controversial expression, the analysis revolves around the direct speaker and the adult listener desiring to receive the content. As a result, the Court will not consider the strength or weakness of the speech interest. By contrast, when regulating to address negative secondary effects (such as zoning adult theaters or bookstores) or when regulating a platform under a content neutral regime (such as cable must-carry), the Court does look to the relative strength of the speech interest and/or the importance of the speech.

2. A Checklist for Congress on Content Moderation Laws.

As we have seen, the First Amendment is not the complete barrier to regulating content some have claimed, but neither can it be ignored. No one can doubt that the rise of hate speech and the use of social media and other communications platforms to encourage and even organize physical assaults on vulnerable individuals and communities raises deep concerns for the safety of life and property, and for the ability of targeted individuals to participate fully in society. The First Amendment does not require the government to sit idly by while people are harassed and threatened. Likewise, the First Amendment does not protect fraudulent content in commerce or in non-commercial speech. Nor is the First Amendment blind to the way in which technology amplifies the power of bad actors. Just as the government can reasonably limit the use of voice-amplifying equipment in the real world to protect sleeping residents at 2 a.m., the government may take action to address digital platforms' amplification of bad actors' harmful conduct. In doing so, Congress may

impose regulation on the platforms themselves, as well as on those who use the platforms for illegal or harmful purposes.

But the First Amendment does impose limits. Because of the importance we have attached in our democracy to allowing controversial speech even when hateful or offensive, it is not enough for Congress simply to find that particular speech-related activities are harmful. Nor can the rights of individuals engaged in bad conduct be casually thrust aside. Courts have found that bad actors do not lose the entirety of their First Amendment rights as a consequence of their bad acts. As the Supreme Court found in *Packingham v. North Carolina*, even a convicted child sex offender cannot be perpetually barred from access to all social media.[113]

To address the First Amendment concerns around legislation aimed at harmful content, I propose the following checklist for consideration.

1. ***Does the regulation raise a First Amendment question at all?*** There is a tendency to assume that all online conduct is somehow speech, and therefore is eligible for First Amendment protection. But a good deal of regulation has nothing to do with speech. A requirement to collect sales tax on video subscription services, for example, is simply a mechanism for raising revenue despite the fact that it increases the cost of distributing expressive content. However, if the tax is structured in a way that is clearly designed to impose a burden on speech, or on a particular type of disfavored speaker, then it does raise First Amendment concerns. Alternatively, the conduct may be speech-related but fall into one of the recognized categories of speech that does not receive First Amendment protection, such as speech relating to illegal activities, threats, or other forms of harassment.

   Sometimes this question may be mixed. For example, laws requiring truthful disclosure generally do not raise First Amendment concerns, but laws that require a specific type of notice or image might.[114]

2. ***Is the regulation a regulation of commercial speech or a general regulation of speech?*** Speech proposing a commercial transaction is subject to the *Central Hudson* test. It is important to recognize, however, that simply because the speech occurs in a commercial context does not make it "commercial speech."

3. ***Is the regulation content neutral or directed at a particular speaker or viewpoint?*** If the speech enjoys First Amendment protection, then a regulation of the speech or speaker will

[113] 582 U.S. ___ (2017).
[114] *R.J. Reynolds Tobacco Co. v. FDA*, 696 F.3d 1205 (D.C. Cir. 2012).

generally need to meet strict scrutiny. A content-neutral restriction on speech may be subject to the intermediate scrutiny standard.

4. ***What is the nature of the of the government's interest? What is the nature of the speech interest?*** Unlike strict scrutiny, both intermediate scrutiny and the Commercial Speech Test look at both the nature of the government interest and the importance of the speech interest.

5. ***Does the regulation fall into a category where speech concerns are otherwise relaxed?*** For example, is the regulation designed to protect minors? Is the medium of speech unusually intrusive, or otherwise impossible to block when unwanted? Does the regulation primarily relate to curtailing secondary effects rather than the speech itself? Is the regulation a "reasonable time and space" restriction?

6. ***Is the legislative or regulatory record substantial enough to support the regulation?*** Any time the court determines that a First Amendment interest is at stake, it will require a substantial record to demonstrate the importance of the government's interest and why the regulation is either the least restrictive means, or burdens no more speech than necessary (depending on the standard of scrutiny applied).

7. ***To whom does the First Amendment interest belong?*** It is often argued that any regulation of online services inherently impinges on the First Amendment interest of the online service provider. While this is often stated with great vehemence, the law does not support this view. Rather, where a platform takes no role in selecting the content, it has no speech interest in the content.[115] Alternatively, the platform may have a weak speech interest, but those directly affected might have a stronger speech interest to consider. For example, a prohibition on selling Nazi books or Nazi memorabilia online is not a First Amendment limitation on eBay, which merely serves as a meeting point between a willing buyer and a willing seller. It has no interest in the nature of the goods sold on its platform. At best, eBay might be considered to have a weak interest in maximizing the available content for sale. But such a restriction would clearly affect the First Amendment rights of those seeking to sell such books or merchandise and those affirmatively seeking to buy them.

[115] *Sable Communications v. FCC*, 492 U.S. 115 (1989).

Table of Relevant Constitutional Doctrines

Type of 1st A Test	Primary Case(s)	General Description	Additional Notes
Commercial Speech	*Central Hudson Gas and Electric Corp. v. Public Service Commission*, 447 U.S. 557 (1980)	Limitation on commercial speech must directly advance a "substantial" government interest; burden on speech must be "no more extensive than necessary."	Applies only to commercial speech. Intermediate scrutiny uses similar standard for speech restrictions generally.
Strict Scrutiny	*Miami Herald Publishing Co. v. Tornillo,* 418 U.S. 241 (1974). Buckley v. Valeo, 424 U.S. 1 (1976).	Must serve "compelling government interest" and be "narrowly tailored" to advance the compelling government interest.	"Strict in theory, fatal in fact."
Intermediate Scrutiny	*Turner Broadcasting Systems, Inc. v. FCC*, 512 U.S. 622 (1994) (*Turner I*). *U.S. v. O'Brien*, 391 U.S. 367 (1968).	Regulation must advance an "important" government interest unrelated to suppression of speech, and must not burden more speech than necessary.	Regulation must be "content neutral." If speech interest protected is relatively "weak," then lesser showing required.
"Intrusive"	*FCC v. Pacifica Foundation*, 438 U.S. 726 (1978).	If speech is unwelcome and intrusive into the home, the government may act to protect unwilling listeners.	If an affirmative act is required to access the unwelcome speech, such as subscribing to a service that has some welcome and some unwanted speech, the unwanted speech is not deemed intrusive.
Secondary effects	*City of Renton v. Playtime Theaters*, 475 U.S. 41 (	Where the record shows that certain types of speech have negative "secondary effects" unrelated to the content of the message (e.g., increase of crime in neighborhoods with adult theaters), then regulation on basis of content permissible.	Requires extensive record proving correlation between the regulated First Amendment activity and the harmful secondary effects.
Reasonable Time and Place	*Grayned v. City of Rockford*, 408 U.S. 104 (1972). *Ward v. Rock Against Racism*, 491 U.S. 781 (1989).	Reasonable content-neutral regulations governing the use of venue, narrowly tailored to avoid burdening more speech than necessary.	

i. An Example: Addressing the Problem of Hate Speech on Platforms.

Taking all of these together, let us consider the question of legislating policies designed to address online harassment and hate speech. The term "hate speech" is very broad, encompassing a range of conduct from direct harassment of individuals based on a protected characteristic (e.g., race, gender, sexual orientation, religion) to broad political speech directed against a target group. It can include many tactics and tools, such as the use of bots and massive numbers of accounts controlled by a single individual to increase the potency of the harassment. Concerns about hate speech range from the impact on individuals directly threatened, to the broader impact on the ability of members of targeted groups (or those not wishing to be associated with the hate speech) to use the platform even when not the direct target, to the use of platforms by extremist hate groups to recruit and radicalize individuals to commit acts of violence. At the same time, it is very clear that at least some hate speech we as a society would generally find offensive is protected by the First Amendment.

I will address some potential approaches in Chapter VI. For purposes of this example, I want to focus on how Congress or a regulatory agency would conduct its First Amendment analysis. First, we should be clear on the specific problem we are trying to address. Is the law (or provision of the law) designed to protect individuals from harassment, to enable those wishing to avoid hate speech (or prevent their minor children from accessing hate speech), or to guard against recruitment to criminal activity linked to "radicalization?" Each harm would require a different approach. If the goal is to protect individuals from harassment or allow those not wishing to be associated with the speech (such as advertisers) to avoid that association, then we would want to focus on building a record of the vast secondary harms, and why existing voluntary solutions (such as social media users' ability to block individuals) are not sufficient. We would need to build an extensive record detailing the compelling government purposes served by any solution that burdened speech, including burdening the ability of the platform to offer the service in the manner it wishes and to present the content it wishes. Because the restriction is clearly content-related and not viewpoint-neutral, the record will need to be far more extensive and detailed than usually required for legislation.

Part of the inquiry would be whether the legislation concerns specific platforms that are dominant and therefore unusually intrusive or difficult to avoid. It would also examine whether it bans more speech than necessary by interfering with the ability of speakers to reach willing listeners. For example, legislation that bans anyone from creating specific online forums dedicated to constitutionally protected hate speech would be virtually impossible to justify on a theory of

protecting individuals. Even legislation that bans hate speech on digital platforms over a certain size would create grave First Amendment concerns. By contrast, legislation designed to segregate hate speech into clearly designated online communities, or to require platforms to exclude hate speech from their search results or recommendations unless specifically requested, could be analyzed as a means of addressing the intrusiveness of digital platforms and the existing difficulty for individuals trying to block such speech as unwanted. The proposed remedies could also be supported by demonstrating hate speech's harmful secondary effects on the platform economy. These include a reduced willingness by advertisers or merchants to use platforms because of the consequences of being associated with hate speech, the enhanced burden on platforms trying to police such conduct, and the inability of individuals to participate in online civic discourse because of the hostile environment created by online hate and harassment.

If we are trying to prevent the harms of radicalization and organizing for violent illegal activity, it is not enough to shield individuals. It is extremely difficult to imagine a law that could successfully ban content that might radicalize individuals enough that they commit violent acts without suppressing constitutionally protected speech. But this would not leave the government entirely helpless. Remedies would be limited to outlawing behavior that falls outside the First Amendment (such as speech directly related to planning a crime), or requiring mechanisms designed to assist law enforcement in surveillance that are narrowly tailored to address the unique properties of digital platforms.[116]

On the other hand, certain approaches might not trigger strict scrutiny. Congress could directly prohibit (or impose civil penalties for) certain types of conduct that fall outside the realm of First Amendment protection, such as harassment. For example, Congress could outlaw harassment via a digital platform in the same way it has outlawed harassment over the telephone. Congress could focus on regulating the tools that make harassment easier online, such as prohibiting the use of bots and fake accounts for deceptive purposes or to harass individuals or without the express consent of the platform provider.[117] To the extent these trigger First Amendment concerns, they are far easier to draft in a content-neutral manner that would trigger intermediate scrutiny rather than strict scrutiny, or that could be addressed as commercial speech under *Central Hudson*.

---

[116] An example of this kind of legislation is the Communications Assistance to Law Enforcement Act (CALEA), which requires operators of telephone networks to build into their networks an ability for law enforcement to surveil their systems. This legislation does not change the due process standard necessary for law enforcement to secure a warrant to conduct the surveillance, but it ensures that ongoing surveillance is possible.

To be clear, I am not proposing such an approach for digital platforms. I merely cite CALEA here as an example of legislation that can assist law enforcement in countering illegal activities without suppressing speech or violating other constitutional protections.

[117] The right to speak anonymously is constitutionally protected under some circumstances and is valuable in many circumstances — including in cases that involve controversial speakers. Similarly, not all uses of bots are harmful. Many uses of bots, including anonymous bots, are beneficial. Any legislation would clearly need to take these issues into consideration. My sole purpose here is to provide an example of how Congress or a regulator should analyze any proposed remedy under the First Amendment.

To conclude, while the First Amendment imposes significant restrictions on how federal or state regulation may address the problem of moderating unwanted or harmful content, it does not render the government helpless in the face of real harms. With careful drafting and a substantial record, Congress and other regulators can address problems of harassment and disinformation.

3. Platforms as Gatekeepers: Advantages and Problems of Private Censorship; Potential First Amendment Issues.

Generally speaking, the First Amendment ban on direct prohibitions applies only to government action. As we have seen over the last several years, existing laws banning deceptive advertising or other types of traditionally prohibited content have proven ineffective against the flood of harmful content that undermines our ability to engage in positive civic discourse online and diminishes the value of digital platforms for either commercial or social use. This does not mean that more effective laws cannot be drafted. But this explains why platforms have faced increasing social and political pressure to police themselves (Keller 2019).

Whether to rely on informal private censorship of platforms for content moderation has given rise to an extensive literature and intense discussion over the last few years (Keller 2019; Klonick 2018; Hassen 2018; Balkin 2018b). For purposes of brevity, I do not explore these arguments here in any great detail. Suffice it to say that proponents of leaving content moderation to private companies, relying on incentives and social pressure (and criminal penalties for individuals who use digital platforms for illegal purposes), highlight the dangers of government censorship. They worry especially that governments will use laws designed to protect people from harmful content to suppress dissent and control information. Both history and present practice support these arguments and illustrate the cost to freedom and innovation when these laws are abused.

Those who argue against turning digital platforms into gatekeepers/censors, either through informal social pressure or through legal liability for permitting “bad” third-party content (however defined) to appear on their platforms, rightly point out that corporate censorship can be just as inimical to free speech as government censorship, and lacks the legal protections of due process and the First Amendment. For-profit businesses (especially large businesses) tend to try to avoid controversy and to limit litigation risk. This historically has led to over-censoring permissible speech, particularly political speech or speech from traditionally marginalized communities that the mainstream might find uncomfortable. Worse, large platforms are subject to political pressure — so-called “regulation by raised eyebrow” — to block speech that governments or law enforcement officials find embarrassing or upsetting to the status quo. Yet this is precisely the kind of speech that deserves the highest levels of First Amendment protection.

Instead of choosing a side, I wish to draw attention to several uncomfortable truths that argue against complete immunity for platforms or making platforms the primary police of content.

No scheme of content moderation can be effective unless it combines both public and private rulemaking. We have seen the effects of delegating content moderation entirely to the private sector by immunizing providers from any consequences for their failure to successfully censor their platforms.[118] The result, all would agree, has proven highly unsatisfactory. Often, companies have no incentive to address harmful content, or have incentives to ignore or even encourage harmful content. Even where companies have strong incentive to develop effective mechanisms to police content, such as eliminating fraudulent product reviews that undermine confidence in reviews generally, these processes often punish innocent conduct, fail to identify disallowed content, are difficult to use for ordinary consumers or businesses, and are still subject to manipulation and gaming by "bad actors." (Emerson 2018; Hazony 2018; Dzieza 2018)

This is not to say that systems developed by statute or regulatory agencies are perfect either. Rather, it is clear that the issues and potential policy tradeoffs are far too complicated for private companies to handle alone. Digital platform providers and users are enormously frustrated by the lack of clear rules of the road or guidance. Nor is this a situation in which competition among platforms is likely to produce platforms with optimal policies that draw users "voting with their feet." Setting aside the numerous objections to whether "voting with one's feet" is even possible when it comes to digital platforms (given problems such as information asymmetry, switching cost, lack of alternatives, and potential collective action problems), there are many situations in which people are injured even if they do not subscribe to the platform where the harmful speech occurs. For example, if someone creates a "deep fake" pornographic video and distributes it through a platform dedicated to "deep fakes," the harm occurs regardless of whether the victim subscribes to the platform (Cole 2017). Given that there is demand for the technology and product, no rational person can expect a market mechanism to emerge to address this issue.

It will therefore take a combination of private action and public regulation even to begin to address the problems that have emerged.

Regulators and the public must have realistic expectations for the ability of private platforms to administer content moderation policies. This requires ongoing oversight by regulators capable of familiarizing themselves with the technology and retaining institutional knowledge as the technology evolves. Anyone using modern digital platforms

[118] I will discuss Section 230 of the Communications Act, 47 U.S.C. § 230, in greater detail below.

understands that the technology seems both capable of anything and frustratingly incompetent. On the one hand, platforms and advertising firms promise incredible precision for targeted advertising, based on creating incredibly detailed digital profiles. At the same time, these companies cannot seem to reliably identify political advertising or distinguish between real people and bot armies controlled by a handful of individuals in the service of foreign governments (Hazony 2018; Emmerson 2018; Confessore *et al.* 2018). The companies always explain why they can't possibly do something they don't want to do but seem always capable of building in new capabilities to do something profitable. Because the technology itself is an enormously complex "black box," regulators, advocates, and the general public often lack a way to check independently whether the companies' claims about feasibility are true.

As a result, regulators in the last two decades have treated technology as some fragile Rube Goldberg-like contraption that could shatter at the lightest touch of regulation. Today, as scandals mount and the good will these companies long enjoyed has been replaced with suspicion, officials' perception has swung to the opposite extreme: If only companies have incentive to "nerd harder" they will come up with filters that can make highly individualized, context-dependent decisions perfectly, to everyone's satisfaction, every nanosecond, on things that human beings frequently disagree about. Neither view permits development of healthy and sustainable policy.

One important reason Congress traditionally uses legislation to define broad policy goals and then delegates to agencies the power to achieve those goals, is the difficulty of monitoring dynamic and changing sectors of the economy. Especially in the case of a complex and rapidly evolving industry, public policy works best when there exists a regulator that is capable of defining the parameters of a problem, is responsible for hearing input from all interested parties, and whose decisions are reviewable by a court. Additionally, designation of an agency allows for development of expertise, preserved over time, so that regulators and the public do not constantly need to re-educate themselves every time a new policy question arises.

Development of an effective system will only emerge over time. No system can operate perfectly on Day 1, or even Day 100. Enforcement systems will need to evolve over time as they are applied. This means designing systems that are tolerant of error and capable of adapting over time.

Aligning platforms' incentives with those of the public interest requires mechanisms to lower the cost of good behavior and raise the cost of bad behavior while not mandating censorship of permissible speech. Policy is not about getting people to do the right thing for the right reason. Policy is about getting people to do the right thing for their own reasons. As we have seen, the existing incentives of digital platforms (including the desire to avoid bad publicity and the desire of advertisers to avoid association with harmful content) are insufficient to address the

numerous problems associated with harassment, fraudulent content, and the impacts of racist and sexist content. We must therefore keep in mind that public policy works most effectively by lowering the cost of desired behavior and raising the cost of undesired behavior.

This requires a combination of clear instructions and obligations on platforms to make compliance possible; safe harbors to protect platforms that are genuinely working to comply with the law; and penalties of sufficient magnitude that platforms do not consider the cost of violation an affordable cost of doing business. Enforcement powers and private rights of action are longstanding mechanisms with a substantial track record for success. Private rights of action are a necessary supplement to government enforcement for two reasons. First, it would be difficult, if not impossible, for a single agency to police an entire industry sector. Second, enforcement is a matter of political will. When the prevailing winds of policy shift away from enforcement, private rights of action remain available to ensure that corporate incentives remain properly aligned with the goals of public policy.

Because we deal here with speech, we must be particularly careful in how we balance the incentives. If penalties are too harsh or private rights of action are too liberal, then platforms will take the conservative route and prohibit even clearly permissible speech. This is why it is particularly important that duties are clear and that safe harbors are available for platforms operating in good faith to the best of their ability.

The debate on this balance of liability and safe harbors generally revolves around Section 230 of the Communications Act (47 U.S.C. § 230) and, to a lesser extent, Section 512 of the Copyright Act (17 U.S.C. § 512). Section 230 limits liability of interactive services for content generally, and Section 512 limits liability for copyright infringement by users of interactive services. Because these provisions have been the subject of considerable debate in recent years, I discuss them at greater length below.

Systems must be transparent to both complainants and their targets, and must incorporate reasonable safeguards to prevent bad actors (either complainants or objects of complaints) from gaming content moderation systems. Any process for moderating content, especially one mandated by law, will succeed only if the public and reviewing courts see it as fair. This requires a process that is straightforward to use for all parties (the complainant, the object of the complaint, and the administrator of the complaint process), transparent as to the decision-making process, and affording remedies commensurate with the nature of the harm and the size of the platform. Additionally, given that speech is often time sensitive, the system needs to reflect the likelihood that users will seek to manipulate the content moderation system — to gain advantage over a competitor, to gain political advantage, or simply as a new form of harassment. At the same

time, platforms must have freedom to respond when there is good reason to believe that there is an immediate risk to life or safety.

While we tend to think of the problem of content moderation as arising primarily in the realm of social networks, political speech, or other forms of controversial speech, moderation of reviews is also an important part of content moderation. As the importance of reviews and review sites has grown, competitors have found ways to manipulate the process to have rivals removed from Amazon or other important commercial platforms (Maynes 2019; Dzieza 2018). Any policies adopted must sharply distinguish between the broad protections afforded to political speech, criticism, and other non-commercial uses, versus regulation of industry practices that, while they may involve speech, are more properly analyzed as commercial speech, or even simply as commercial activity raising no First Amendment concerns.

It is the responsibility of government, not the private sector, to find the appropriate balance between protecting freedom of speech, protecting individuals from harm, and ensuring that the digital world does not become so polluted with false, fraudulent, or harassing content that it cannot reach its potential for serving the public interest. The temptation for Congress and federal regulators to rely heavily on the private sector to set content moderation rules and policies is immense. Reliance on finger-wagging, threats of regulation, and pressure from public shaming avoids the thorny problem of balancing First Amendment interests. It also allows decision-makers to avoid the hard process of drafting laws that will, inevitably, be over-inclusive, under-inclusive, or both. This temptation to "pass the buck" is not simply a refusal to draft. It also includes use of vague terms and standards. Germany's NetzDG law, for example, requires digital platforms to remove "obviously illegal" content within 24 hours. In a no-doubt unintended irony that highlights the problems with such standards, NetzDG extends this deadline to seven days when it is complicated and non-obvious to determine whether the content is "obviously illegal" (Feld 2018b; Kinstler 2018).

Particularly in the sensitive area of balancing what we as a society find intolerably threatening and vile, Congress cannot outsource the decision to private companies. The political process is the process by which we as a society try to reach a rough, workable consensus on how to manage the right of individuals to live freely in our digital society while maintaining the right of individuals to live without fear of harassment as the cost of participation. It is not merely our elected representatives' legal responsibility, but their moral imperative, to find the appropriate balance and to embody that balance in sound policy. We may never reach complete agreement on how to strike that balance or how to craft effective policy. But to refuse to act, thus delegating fundamental judgments on the governance of digital content to a handful of private actors, would be a stunning failure and an act of moral cowardice.

C. "Publisher Liability," Section 230 and the Digital Millennium Copyright Act (DMCA).

As noted above, one important aspect of aligning the incentives of platforms to protect users from unwanted, harmful content is through private rights of action. Imposing fear of liability for negligence is the traditional means of encouraging businesses to observe a basic duty of care. Fear of liability for defective products is the traditional means of encouraging businesses to build products that function as advertised. Fear of liability for the acts of employees and subordinates is a traditional means of encouraging businesses to exercise reasonable oversight over their employees. Jack Balkin and others have proposed the idea of an "information fiduciary," imposing an obligation to protect information that users disclose to platforms (Balkin 2018b). As a consequence of decisions made as part of the Telecommunications Act of 1996, providers of "interactive computer services" enjoy special protection from liability for third-party content. The Telecommunications Act of 1996 amended the Communications Act to include a new Section 230, designed to limit the liability of newly emerging "interactive computer services" for third-party content or for failing to perfectly protect users from third-party content they promised to block.

Few sections of the Communications Act have enjoyed such a storied history of hasty drafting, radically broad interpretation, and subsequent misunderstanding. As a consequence, Section 230 has been interpreted by the judiciary as conferring broad civil immunity for third-party content to broadband providers and digital platforms, and now sits at the center of the debate over content moderation and liability. Worse, it has confused the entire issue by focusing on the red herring of "publisher liability," an extremely limited form of liability that would not address the question of liability for third-party content in the manner envisioned by proponents of eliminating or substantially modifying Section 230.

To untangle this debate, I will review the history of Section 230 and why "publisher liability" (or even "speaker liability") would do little to affect platforms' content moderation policies. Worse, because Section 230 has been in place for so long, simply to remove it (or substantially modify it without clear guidance on how the new liability regime should operate) would create enormous uncertainty and chaos as courts examine what "publisher" or "speaker" liability should actually mean in this context and whether they apply even in the absence of Section 230. I will also argue that the notice and takedown schemes that replicate the flawed model used in Section 512 of the Copyright Act, a model often proposed for requiring platforms to moderate content deemed harmful by the regulating authority, would be disastrous if applied to digital platforms generally.

1. History of Section 230 and How It Has Confused the Current Content Moderation Debate.

Online services became commercially available to consumers in the 1980s. These services, however, were quite different from modern internet service providers, or even the dial-up ISPs that would flourish after 1994. These early online services were primarily focused on "walled gardens" where users interacted with one another. Even sending external emails from one online provider to another might incur a separate charge. These services were actually called "electronic publishing" by the FCC in its *Computer* proceedings, and "electronic publishing" was the term used to describe these and similar activities in the breakup of AT&T.[119]

For some years, these online services remained the province of early adopters and technophiles. The invention and popularization in the early and mid-1990s of hypertext, the World Wide Web, and web browsers such as Mosaic (and later Netscape) changed that. Suddenly, "the internet" became a cultural phenomenon. New content proliferated, and ISPs increasingly shifted from trying to keep subscribers inside their "walled gardens" to permitting greater exchange of content and messages between subscribers and the rest of the online world.

This context is important in understanding the origin and evolution of Section 230. As internet access proliferated, so did "harmful" content, however defined. This triggered lawsuits against the existing online communities under various theories. In one of the first cases in 1991, *Cubby, Inc. v. CompuServe, Inc.*,[120] the court found that the online provider CompuServe could not be held liable for content uploaded to one of its 'forums' because it had no opportunity to review or alter the content. Because CompuServe had no specific knowledge of the content it distributed, it could not be held liable as the publisher of supposedly libelous statements against Cubby by a business rival.

In 1995, a different court found the exact opposite. In *Stratton Oakmont, Inc. v. Prodigy Services Co.*[121] a New York state court found that Prodigy could be liable as a publisher for purportedly libelous statements against Stratton Oakmont. The *Stratton Oakmont* court relied on early statements by Prodigy that it marketed itself as a "family-friendly" service that actively screened content, that Prodigy had an acceptable-content policy, and that Prodigy did make efforts to screen content that violated this policy. The court distinguished *Cubby* on the grounds that

[119] The newly created ILECs, aka the "Baby Bells," were initially prohibited from offering electronic publishing services for fear that they would favor their own content and affiliates over those of rivals. The Court found this prohibition served the interests not only of competition, but the interests of the First Amendment. *United States v. Western Electric Co.*
[120] 776 F. Supp. 135 (SDNY 1991).
[121] 23 Media L. Rep. 1794 (N.Y. Sup. Ct. 1995).

CompuServe made no promises to moderate content and had no history of efforts to police content. The court concluded that explicitly holding itself out as moderating content made Prodigy a publisher, subject to potential liability for the defamatory third-party statements.

At the same time, the rise of the internet brought with it the rise of online pornography and indecent content, as well as new opportunities for harassment. Even more alarming, the anonymity of online communications, particularly "chat rooms" and "bulletin boards" used by subscribers to exchange text information, provided new opportunities for sexual predators. Lurid accounts of the supposed smorgasbord of smut available to minors through a new technology their parents barely understood filled news reports, while accounts of pedophiles using chat rooms to target and recruit victims led to predictable moral panic (Wu 2016). In 1995, Sen. J. James Exon (D-NE) introduced S. 314, the Communications Decency Act (CDA), to limit access to indecent material online.

The proposed Communications Decency Act gained broad support in the Senate and was eventually added by amendment to the Telecommunications Act of 1996. In the process of negotiation, Sen. Ron Wyden (D-OR)and others raised concerns about imposing liability on these new "interactive computer services" (defined in the CDA to include what we would now think of as both online access providers and digital platforms) for indecent content. One argument raised in opposition was the *Stratton Oakmont* decision. Opponents argued that interactive computer services that attempted to filter out obscene content and offer "family-friendly" services would, under the logic of *Stratton Oakmont*, be liable for any indecent material that slipped through. Opponents of "intermediary liability" (holding internet services liable for third-party content) argued that, absent legal protection, online services would take the safe course under *Cubby* and try to shield themselves from liability by explicitly adopting a "no content moderation" policy. Opponents of intermediary liability warned this would prevent anyone from offering "family-friendly" or otherwise curated services, since a single failure to catch a post in violation of the family-friendly policy could result in liability, as in *Stratton Oakmont*.

As a compromise, the CDA included an amendment from Sen. Wyden containing the "Good Samaritan Provision for Blocking and Screening of Offensive Material." This amendment added Section 230(c), which provided that "no provider or user of an interactive computer service shall be treated as the publisher or the speaker of any information provided" by a third party.[122] Although the Supreme Court would later strike down the indecency provisions of the CDA in *Reno v. ACLU*,[123] the rest of Section 230 — including Section 230(c) — remained good law. In a series of subsequent decisions, courts found that Congress intended to confer broad immunity on interactive computer

[122] 47 U.S.C. § 230(c)(1).
[123] 521 U.S. 824 (1997).

services for third-party content.[124] For those concerned about the impact of potential liability on digital platforms, especially during the start-up phase, Section 230 is regarded as a foundational law that protects innovation. For those seeking to force digital platforms to police various sorts of harmful content, Section 230 is regarded as a major obstacle that needs to be repealed or significantly modified.

In point of fact, elimination of Section 230 would do little to get at the kinds of harmful speech increasingly targeted by advocates. Liability as a "publisher" or "speaker" was at issue in the CompuServe and Prodigy cases because the complaint involved defamation and libel, causes of action for which publishers and speakers are traditionally liable. But publishers are not generally responsible for bad acts committed by people inspired by their published works, or if the works they publish offend people.[125] Even liability for harassment as a "speaker" typically requires some kind of intent to harass,[126] which would clearly be absent in efforts to hold digital platforms liable for harassing hate speech. Following the precedent in *Cubby v. CompuServe*, digital platforms are likely to respond to repeal of Section 230 by eliminating their existing content moderation policies rather than by enhancing them to address problems of harassment or hate speech.[127]

Nor is modification of Section 230 necessary to prosecute cases of actual criminal law. Section 230 exempts several categories of third-party content from protection. One such exception is for federal or state criminal statutes. This is important, as discussion of modifying or repealing Section 230 liability protection often involves arguments that these changes are necessary to punish criminals and deter crimes such as sex trafficking or illegal drugs. It is important to recognize that this argument is not literally accurate. In circumstances where an actual publisher would be liable for assisting criminal activity, so would any digital platform even under current law. Advocates of intermediary liability should note that both the 2018 anti-sex-trafficking act known as SESTA

[124] *See, e.g. Klayman v. Zuckerberg*, 753 F.3d 1354 (D.C. Cir. 2014); *Zeran v. America Online*, 129 F.3d 327 (4th Cir. 1997). *But see Doe v. GTE Corp.*, 347 F.3d 655 (7th Cir. 2003) (questioning broad interpretation of scope of Section 230).

[125] For example, in *Doe v. GTE Corp.* the court found that it did not need to address the scope of Section 230 because, even absent Section 230, simply providing internet access and content storage to a third party does not trigger liability for third party's tortious actions. For example, an effort to sue Oliver Stone for purportedly "inspiring" a violent shooting with the movie "Natural Born Killers" was ultimately dismissed because plaintiffs could not show any intent by Stone or Warner Brothers to actually advocate for violent killing.

[126] *Elonis v. United States*, 135 S. Ct. 2001 (2015). In this case, the Supreme Court found that an individual posting "rap lyrics" describing violent fantasies about his ex-wife on his public Facebook page did not violate the criminal statute against harassment by wire without a showing of specific intent to harass. While *Elonis* was a matter of statutory interpretation of a criminal statute, it is consistent with the First Amendment ruling in *Virginia v. Black*.

[127] It is, of course, impossible to say with any certainty how the common law on liability for third-party content would have developed without the passage of the CDA. Nor can we predict how it would evolve today if Section 230 were repealed. It does seem likely that, absent any other guide to behavior, platforms would reflect the distinctions made between *Cubby* and *Stratton Oakmont* and decline to adopt any moderation policy as the safest course.

[128]and the notice and takedown provisions of the DMCA did not simply exempt the targeted content from the protection of Section 230. Both needed to take additional steps to impose liability on the platforms for the third-party content deemed harmful.

To say that eliminating Section 230 would neither create liability for many kinds of harmful content nor aid criminal prosecutions, is not to say that Section 230's protections have been meaningless, or that its removal would be harmless to the internet ecosystem. To the contrary, Section 230 has protected digital platforms (and ISPs) from particular kinds of lawsuits. Amazon, TripAdvisor, and Yelp, for example, have not needed to worry about being sued out of existence over every bad review. The recent lawsuit by Rep. Devin Nunes (R-CA) against Twitter for $250 million (Coaston 2019) is precisely the kind of lawsuit brought for political reasons, or to intimidate critics, that digital platforms would face regularly — at least until the law settled. The legal foundation provided by Section 230 is now settled law. Eliminating Section 230 without providing a meaningful replacement would create legal uncertainty for websites, ISPs, and the entire internet ecosystem potentially classifiable as "interactive computer services."[129] Years of litigation against every sector of the internet economy under every possible theory of liability would ensue until a new legal equilibrium was reached. But worse than the possible cost would be the likelihood that eliminating Section 230 would do absolutely nothing to address the problems of harassment, hate speech, or other harmful content that advocates of eliminating Section 230 believe they could reach with civil suits in the absence of Section 230.

The best policy is therefore to leave Section 230 alone as irrelevant to resolving the issues of content moderation. Instead, Congress should focus on developing the details of the appropriate regime for content moderation along the lines of the recommendations below.

[128] SESTA began life as a version in the House called the Fight Online Sex Trafficking Act (FOSTA). Some sources therefore refer to the law as FOSTA-SESTA. Additionally, some source material refers either specifically to the House bill or to the law as passed as FOSTA rather than SESTA.

[129] If broadband providers were classified as Title II common carriers, they would have no need for protections such as those provided by Section 230. Common carriers are automatically immune to liability for the traffic they carry, since they are powerless to prohibit it. The FCC's decision in December 2017 to reclassify broadband as an "information service," based in part on a novel interpretation of Section 230(a), eliminated the default liability protection for common carriers and restored broadband access providers to the "interactive computer services" definition under Section 230.

A discussion of the classification of ISPs is not germane to this paper. I merely point out here that the collateral damage of elimination or radical restructuring of Section 230 would go well beyond social media or even the entire digital platform sector. Eliminating Section 230 would transform ISPs back into "electronic publishers," with accompanying liability for content under the same circumstances as digital platforms. Legislation imposing liability must specifically exempt ISPs or otherwise modify the definition of "interactive services" in Section 230 (unless, of course, legislators intend to impose similar liability).

2. Lessons from Existing Content Moderation Regimes: SESTA, DMCA, and NetzDG.

Since Section 230 became law, we have seen two major exceptions added in the United States. In 1998, Congress passed the Digital Millennium Copyright Act (DMCA), which contains a "notice and takedown" requirement for allegedly infringing material posted by third parties on digital platforms. Variations on the DMCA "notice and takedown" regime can now be found in the laws of many nations, creating the best-known approach to civil liability for platforms for third-party content. More recently, Congress passed the Stop Enabling Sex Trafficking Act (SESTA) in April, 2018 (Romano 2018a). In 2017, Germany passed its NetzDG law, a "notice and takedown" requirement for "obviously illegal" third-party content. All of these regimes provide insight into the difficulties in using civil liability as a means of requiring platforms to police their content (Feld 2018b).

Predictably, platforms have been extremely aggressive in moderating content in the face of potential liability. As we shall examine below, this leads to chilling effects for legal content and creates significant opportunities for political or commercial entities to use these systems to target rivals. This not only imposes significant costs on those improperly blocked and those denied access to legal content. It imposes significant costs on digital platforms that are not paralleled in the offline world.

In the case of SESTA, whatever the long-term effectiveness, it appears to have had immediate unintended consequences that may aggravate the problem of human trafficking rather than alleviating it. Sex workers have stated that a law arguably designed to protect them has, in fact, placed them in life-threatening danger by requiring them to return to streetwalking and the use of pimps. San Francisco experienced a 130 percent surge in the number of human trafficking complaints in the last year, as well as associated complaints from neighborhoods where streetwalking has increased, since SESTA triggered the takedown of online personal sites used by sex workers to screen potential clients (Steimle 2019). Other police departments have likewise complained that SESTA has made it harder to catch pimps and that sex trafficking has actually increased as a result of the law (Masnick 2018). While there is less direct evidence with regard to the impact of Germany's NetzDG law, it is noteworthy that nearly all of Germany's opposition parties have called for its repeal, and that free speech advocates contend that platforms have been overaggressive with regard to policing speech (Pearson 2018).

A common weakness in all three regimes is the lack of reporting metrics that allow lawmakers to track the impact of these laws over time. (NetzDG has metrics and a reporting requirement, but as I explain below, these metrics are not terribly useful in gauging whether critics are right that platforms are overaggressively censoring content.) This is one of the difficulties in acting directly rather than through an enforcement agency. An agency with permanent oversight

jurisdiction can monitor the impact of a law over time, and can mitigate impacts from a law that turns out to be too harsh in practice, or creates uncertainty, or otherwise has negative unintended consequences. If nothing else, the agency has the capacity to report to Congress on the need to amend legislation in light of unfolding developments.

i. Impact of SESTA — Simple Civil Liability.

SESTA is an example of a direct imposition of civil liability to require platforms to screen and prohibit content deemed harmful. SESTA imposes criminal and civil liability for anyone who "recruits, entices, harbors, transports, provides, obtains, advertises, maintains, patronizes, or solicits by any means," for a business where either a person under age or a person subject to threats or coercion (defined by the statute) engages in a "commercial sex act" (also defined), or who "benefits" from any such "venture."[130] But whereas criminal penalties generally require a fairly high standard of either actual knowledge of the forbidden conduct or reckless disregard for clear signs that the intent is criminal, civil liability requires a lower standard. SESTA therefore also imposes civil liability on anyone who "knew or should have known" that the thing they were advertising or "soliciting by any means" violated the law.[131] Civil liability also applies to anyone who "benefits" from any of these activities in support of such a "venture." A victim or a state attorney general can bring a civil suit for up to ten years following the conduct at issue.[132]

The language "knew or should have known" is highly ambiguous, and often implies a responsibility to engage in some sort of investigation to ensure that the conduct in question does not violate the law (Clough 2018a). This was arguably intentional on the part of the drafters. Imposing potentially broad liability maximizes the incentive for platforms to police themselves and ban content well beyond what could be subject to the due process and other constitutional protections of criminal law. Even if a platform were ultimately found not liable, the expense of litigation is quite high — and can be crushing for small and medium-size platforms. Platforms — especially smaller platforms — therefore have particular incentive to refuse advertisements that a physical publication would routinely accept.

Whether intended or not, SESTA triggered a widespread takedown of interactive websites and advertising venues that could conceivably incur liability under SESTA. Platforms began to suspend broad swaths of content and activities. Craigslist, for example, removed its

[130] 18 U.S.C. §1591(a).
[131] 18 U.S.C. §1595.
[132] SESTA also permits states to pass their own criminal and civil liability statutes consistent with the federal statute.

entire personals section — a mainstay of traditional classified advertising that contained predominantly legal content (Romano 2018a). In a lawsuit brought in the Federal District Court for the District of Columbia Circuit, advocates for sex workers and free speech generally argued that SESTA creates a chilling effect on clearly protected speech, such as advocacy to make sex work legal (Gullo and Greene 2019).

As a consequence of the mass takedown of websites and services traditionally used by sex workers, as well as the refusal of remaining websites to take any advertising or permit content that could arguably trigger liability under SESTA, sex workers who voluntarily engage in sex work[133] have complained that SESTA has placed them in far greater physical danger, the opposite of SESTA's purported intent (McCombs 2018). Prior to SESTA, sex workers could use direct advertising to avoid traditional "streetwalking," to provide phone information so they could pre-screen clients, and to avoid pimps and other potentially abusive middle-men. It also led to the shutdown of "bad date lists," online resources maintained by sex workers to avoid dangerous clients. As one sex worker told a reporter, "The bill will, and already has been, responsible for the murder, rape and arrest of sex workers" (McCombs 2018).

The law has also severely affected those engaged in legal indecent and pornographic expression. Its reach goes beyond advertising to all digital platform activities. Electronic payment processors are now declining to process payments for smaller websites associated with pornography (or indecent content) that in any way may relate to any kind of sex work. While these sorts of activities are often disfavored by law and society at large, legal erotic and indecent content is protected by the First Amendment. To the extent shutting down such protected speech is an intended rather than unintended consequence, it represents an end run around constitutionally protected rights.

Opponents of the law also point out that the law falls particularly hard on traditionally marginalized communities such as people of color, LGBTQ groups, low-income people, and the disabled. Those able to "class pass," as sex workers call it, are able to evade the law by advertising in more expensive venues and using language that evades detection. Additionally, the stereotypes associated with traditionally marginalized communities make it far more likely that their activities will be perceived as sexual and/or illegal even when they are not (Elliot and Gillula 2017).

A year, of course, is far too short a time in which to assess the law's effectiveness at reducing sex trafficking, its primary purpose. But determining the appropriate cost/benefit

[133] The term "sex worker" and "sex work" can include a wide range of activities and is not limited to traditional prostitution, although it certainly includes traditional prostitution.

tradeoff is further complicated by the lack of any kind of reporting mechanism or indicator of what metrics constitute success or failure. Anecdotal evidence, however, underscores the hazards of imposing direct liability on platforms for third-party content without any consideration for the difficulties platforms will encounter when trying to pre-screen content that may incur liability.

ii. DMCA and NetzDG: Notice and Takedown and Safe Harbors.

Section 230 exempted violations of intellectual property law from its broad protection from third-party liability. Laws governing liability for copyright infringement, and in particular laws relating to liability by third parties or providers of new communications technologies, are more complicated and contentious than those governing defamation or libel. It would far exceed the scope of this paper to explain the complicated nature of copyright and the politics surrounding the passage of the Digital Millennium Copyright Act. Suffice it to say that each evolution of communications technology, such as the invention of movies, the development of broadcasting, and the advent of digital media, have all prompted a robust and contentious debate over how to balance the right of copyright holders to profit from their creations, the rights of readers or listeners or other "consumers" of copyrighted material, and the strong public interest in promoting new technologies and competition. The advent of the internet was no different. Congress passed the DMCA in 1998, creating (among other things) a new regime governing the liability of digital platforms and "transient networks" (ISPs and other providers of communications services that do not store copies).

The DMCA added a new section to the Copyright Act entitled "Limitations on Liability Relating to Material Online."[134] Section 512 distinguishes between "transitory digital networks," essentially communications networks, and digital networks that store content of any kind. For convenience, I'll refer to these as digital platforms.

Section 512 provided that digital platforms could be liable for financial or injunctive relief if third parties used their services to store or exchange infringing content, unless the digital platform complied with the safe harbor. The provision requires the digital platform to have no knowledge of the infringing activity, to take steps to remove infringing material when discovered, and to remove allegedly infringing material if a rights holder provides notice containing the information dictated by the statute. The digital platform is then obligated to inform the alleged infringer of the takedown. The alleged infringer may then send a "counter-

[134] 17 U.S.C. § 512.

notice" to the digital platform challenging the allegation of infringement. The digital platform will then forward the counter-notice to the accuser, informing the accuser that it will restore the content in 10 days unless the accuser files a copyright-infringement action in federal court. Assuming no such action is filed, the digital platform is obligated to restore the challenged content in 10-14 days (Urban, Karaganis, and Schofield 2017).

I have deliberately elided numerous details in the statute that are highly relevant to DMCA practitioners, and thus potentially missed details relevant in assessing the DMCA's overall costs and effectiveness. To reiterate, my purpose here is not to evaluate the DMCA or suggest any alteration, but to provide a basic understanding of the "notice and takedown" regime, which has been replicated in other countries and is being considered for other content moderation purposes in the EU. Although some stakeholders have criticized the DMCA as insufficient to prevent widespread infringement, and others believe it has imposed significant costs on individuals, businesses, and free expression (Urban and Quilter 2006), it has managed as a reasonably workable regime for 20 years. Section 512 was the first genuine effort to strike a balance recognizing on the one hand the limitations of technology for monitoring and making judgments in a universe where at any moment millions of users are uploading and downloading billions of bits of user-generated content, and on the other the urgency to injured parties of gaining quick relief. It also made some effort to protect the process from abuse by imposing potential penalties for false allegations of infringement and by providing for a counter-notice by which parties may have content restored that was wrongly alleged to be infringing. For this reason, the DMCA "notice and takedown" regime has become an attractive model for lawmakers considering other types of content that requires moderation.

In considering this balance, however, lawmakers should take several cautionary lessons. First and foremost, the statute contains no mandatory reporting or other metrics to ascertain whether the "notice and takedown" regime is effective, or whether it imposes significantly higher costs than anticipated (or necessary). There is no assessment of whether the DMCA has disproportionate impact on particular communities, whether DMCA is actively used to suppress speech, or whether it results in significant loss of opportunity for fair uses such as criticism or education. A number of reports and studies have suggested that platforms are often lax in complying with obligations to restore content subject to counter-notice, and that DMCA takedown notices are often employed as a weapon against critics and rivals (Urban, Karaganis, and Schofield 2017).

More importantly, lawmakers need to recognize that no matter how technically difficult it may be to use technological means to identify and/or filter infringing uses, or to assess

whether a particular use constitutes fair use, and no matter how complicated, difficult, and expensive it is to apply these standards globally, the problem of identifying and moderating hate speech or other harmful content is far worse. Sometimes harassing or deceptive speech is blatantly obvious, such as posting "revenge porn" or releasing someone's personal information publicly without their consent and encouraging their mass targeting (a practice known as doxing). But it is far more common for harassment or deception to be context-dependent. To take a simple example, calling someone a Nazi may be a hurtful insult (especially if directed at a Holocaust survivor), overblown rhetoric, political commentary, or literal truth.

Germany's NetzDG law attempts to address these concerns. The law imposes a "notice and takedown" provision for any "obviously illegal" content and refers to specific German laws under which the relevant content might be deemed "obviously illegal." The statute only applies to platforms with 2 million or more users, so as to limit cost to smaller services. It requires takedown within one day (seven days if the "obviously illegal" content is sufficiently non-obvious to require consultation with legal experts) and mandates a right of appeal for anyone taken down. Finally, NetzDG requires platforms that receive complaints to publish a "transparency report" twice a year that must include quantitative metrics such as the number of complaints received, the average processing time, the number of takedowns in response to complaints, the number of appeals, and the ultimate resolutions of the complaints (Library of Congress 2017).

Despite this comprehensive effort to address the legitimate concerns of overreach while still providing meaningful relief, and despite the requirement for transparency reports, there is no consensus within Germany as to whether the law is effective and whether it is suppressing protected speech. Reporters Without Borders, for example, has argued that the first transparency reports issued in August of 2018 show that lawful content is being blocked (Reporters Without Borders 2018). Some controversial speakers have argued that their rights to free expression are being violated (Kintsler 2018). Others dispute this characterization. No one doubts, however, that hate speech and other types of harmful content remain a serious issue on digital platforms in Germany.

Taking all this together, lawmakers should be wary of the argument that the DMCA "notice and takedown" regime is easily exportable outside the realm of copyright infringement. While Section 512 provides many important lessons, positive and negative, it is by no means a comprehensive solution.

D. Specific Recommendations for Content Moderation Policies Designed to Maximize Effectiveness and Minimize Unintended Consequences.

1. Recommendation 1: A Mixed Model of Government Prohibition, Reporting Requirements, and Private Policing with Government Oversight.

No single model or set of rules can resolve the many problems associated with content moderation across the wide range of digital platforms. Manipulating reviews on Amazon or TripAdvisor for commercial advantage is very different from maintaining swarms of fake accounts on social media, which is very different from recruiting by terrorist organizations. This complexity argues for a multipronged approach that triages the nature of the problem and divides the type of content moderation into different categories.

First, there is conduct that falls into the long history of clearly criminal or harmful conduct. We should criminalize such conduct and subject it to civil penalty, just as we have done with harassment by telephone[135] and use of wire, radio, or television to commit fraud.[136] We have seen examples of such conduct that are unique to digital platforms and should be directly criminalized and made subject to civil penalty as well as private rights of action. These include doxing (publishing a person's personal information for the purpose of harassment), revenge porn (publishing sexually explicit photographs or video without consent), and manipulation of a digital platform through false reviews or false complaints.

Although these activities are clearly harmful in most cases, they do raise some potential First Amendment concerns. For example, exposing the name and home address of a public figure to organize political protests is a politically protected activity, very different from exposing a person's information for the purpose of encouraging harassment and death threats. Indeed, even the question of when a person becomes a public figure can be difficult to determine. As noted above, there are ways to address these concerns. The real world presents similar issues when addressing laws around harassment, hate speech and libel. The First Amendment must be respected, but it is not an excuse to do nothing.

Nevertheless, we must be conscious of the limitations of this first step. Because it criminalizes (or subjects to civil penalties) the conduct of users rather than the platform itself, it raises problems of enforcement, as we have already seen. Discovering the real identity of the party committing the alleged bad acts can be difficult. The proliferation of incidents makes enforcement

[135] 47 U.S.C. §223.
[136] 18 U.S.C. §1343.

by state or federal authorities challenging, and it is relatively easy for parties to return with new identities.

Means of addressing these problems have their own limitations. For example, civil penalties can provide an ability and incentive for individuals to punish bad actors. They also shift the enforcement burden to the injured individual. If penalties are too large, they may be used to deter legitimate speech or harass innocent speakers. If penalties are not large enough, there is no value in pursuing them. Platforms must be compelled to cooperate with investigations. This imposes expense on the platforms, and if poorly designed can become a tool of privacy violation and abuse.

One important means of mitigating these limitations is to provide an enforcement agency with power and resources to handle complaints. Properly designed agency processes can protect against abuses and are more likely to be responsive than state or federal law enforcement agencies. Additionally, administrative agencies are better suited to address bad conduct in the commercial sphere, including the ability to set rules governing commercial conduct and to enforce those rules. Such an agency can help relieve platforms of the responsibility for policing conduct by providing clear guidelines for appropriate policies, which will also facilitate enforcement by creating standard practices across the industry.

But what of more complicated conduct that has no clear analogy in the digital space or presents difficult First Amendment concerns? What about complicated schemes that may not be readily apparent to individual businesses but may be recognized by platforms? For example, the use of social media by Russia to influence elections was not readily apparent. It was only after the federal government began its investigation that social media platforms and researchers found patterns indicative of manipulation. Similarly, the use of bot armies or swarms of fake accounts is more readily detectable by the platform than by anyone else. Still, suspicious activity is not necessarily criminal or fraudulent activity.

The “know your customer” rules imposed on the financial industry to track money laundering by criminal or terrorist organizations provide a potentially useful model. Banks do not have an obligation to ferret out such criminals and deny them service — something they are wholly unsuited to doing. Instead, law enforcement and regulatory agencies have worked with financial institutions to develop a list of suspicious signs that trigger reporting requirements. It then falls to the relevant agency, which is subject to due process and is specifically designed to make such determinations, to investigate and take appropriate action.

Rather than impose a responsibility on platforms to make judgments about criminal or civil issues for which they are unsuited, the law can require digital platforms to work with law

enforcement on appropriate tools to detect suspicious activity and to report such activity to the appropriate agency. This might be in addition to a digital platform's own acceptable-use policies. Indeed, a hybrid model may be most effective. The platform is obligated to adopt certain best practices with regard to detecting suspicious activity. The platform is obligated to either report to the relevant agency (which then has responsibility to act) or take action directly. Where the platform takes action directly, the party protesting its innocence may appeal to the relevant agency to reverse the platform's actions. This would permit platforms to protect their users and the overall integrity of their business without putting them in the no-win position of arbitrating over-inclusiveness or under-inclusiveness. It would also provide to those cut off from the platforms a right to appeal to a government authority, mitigating concerns over private censorship of time-sensitive political speech.

This approach would also impose costs, but these could be scaled to the size and nature of the platform, and be made dependent on the nature of the concern, the conduct, and the context. It is certainly appropriate for a platform to decide that it will prohibit hate speech or erotic content. But the First Amendment requires that we tolerate such speech between willing participants. At the same time, it does not violate the First Amendment to obligate forums to cooperate with law enforcement to prevent real crimes when there is probable cause.

As an example, consider Gab, a social media platform designed for users seeking content deemed by Facebook, Twitter, and other social media platforms to violate its content policies on hate speech and harassment (Coaston 2018). It is the archetypal example of conflict between competing First Amendment concerns. Because of Gab's user base, its user-generated content is often vitriolic and racist, comparing Jews, immigrants, and people of color to animals or vermin and warning that white people need to defend themselves. The content often uses terms that may or may not cross the line from protected speech to steps in preparation for actual violence. Indeed, Robert Gregory Bowers, the man who killed 11 people at the Tree of Life Synagogue in Pittsburgh on October 27, 2018, posted on Gab just prior to the shooting his theory that Jews were assisting South and Central American immigrants to cross into the country illegally. He went on to say that he was "going in" to stop this "invasion" (Roose 2018b).

Some argue that the presence of forums such as Gab allows hate groups to recruit and radicalize individuals. Under this reasoning, Gab itself becomes a uniquely present danger and should fall outside First Amendment protections. Even if none of the speech on Gab or similar platforms constitutes an "imminent danger" in the abstract, the law should recognize that the nature of the technology and of the speech in question facilitates and incites violence in ways that traditional print and broadcasting do not (Sunstein 2018). The proper analogy is therefore not to a set of individual speakers making individual statements in isolation, but to an angry mob that needs

to be dispersed before it riots. Just as an angry mob threatening a targeted individual or group may be dispersed without offending the First Amendment's right of freedom of assembly, the law can impose limits on digital platforms to prevent the evolution of similar threats of violence.

Others argue that unpopular speech such as this is precisely when the First Amendment rights of freedom of speech and freedom of assembly are most critical. As we have seen throughout our history, organizations and causes we now take for granted as mainstream have been, and sometimes still are, prosecuted under laws designed to protect the public from harmful speech. Those engaged in organizing for labor rights and unions were prosecuted as anarchists or communists. Anti-war protesters in World War I and opponents of the draft were subject to criminal prosecution. Birth control literature was criminalized as immoral under the Comstock Act. Under this reasoning, even encouraging private censorship via the critical infrastructure necessary to operate a digital platform — such as domain hosting and electronic payment processing — creates a serious danger to robust debate and controversial content which may ultimately prevail in the marketplace of ideas.

How would a proposed mixed regime address Gab and similar sites? Certainly, they would be allowed to continue operation. However unwelcome and revolting many of us find such content, and even if we find there is an increased correlation between the availability of such forums and radicalization, the First Amendment does not permit us to ban "bad" ideas and "bad" speech on the grounds that it merely increases the likelihood that someone will commit a crime of violence. This is the risk that a free society requires.[137] Gab might not even be required to respond to complaints of harassment on its platform, as it is a small platform known for attracting users who engage in such behavior. Individual users therefore "assume the risk" of being targeted and attacked in ways that would be considered outrageous on other platforms. By contrast, a larger platform such as Facebook or YouTube might be required to have some anti-harassment policy in place.

This difference in obligation honors the Supreme Court's distinction between "intrusive" content and content that a person affirmatively seeks out. It also respects the increased harm to individuals who can't use a more important or dominant platform without subjecting themselves to harassing behavior. Where a platform is sufficiently large that exclusion from it would impose a persistent and not insignificant cost, the government has a legitimate interest in preserving access to that platform and in individuals not being required to subject themselves to harassment or even physical danger as the price of participation. In the same way that the federal government has a

[137] It is worth noting in this regard that other democracies that ban various sorts of "hate speech" do not ban an amorphous general category. Bans are quite specific, such as specific Nazi symbols. Speech that merely evokes such symbols to transmit the same ideology, such as the "quenelle" or "reverse Nazi salute" popularized by French comedian Dieudonné M'Bala M'Bala, are not subject to prosecution despite the deliberate invocation of the Nazi salute.

sufficient interest in preventing harassment by phone to require that telephone operators protect the telephone number and other personal information of subscribers,[138] the government can protect users of digital platforms from harassment.

But even recognizing that a platform like Gab may have greater leeway in light of its smaller size and clear warnings about the nature of the user base, it and platforms like it could still be required to prevent themselves from being used to engage in prohibited activity. For example, the law could require all platforms, even platforms like Gab, to create specific mechanisms for responding to complaints by non-users that users had posted personal information in order to organize harassment campaigns (i.e., doxing). Steps might include taking down such content and cooperating with any civil suit or investigation. All platforms could be required to report any illegal activity of which they have actual knowledge and could be required to monitor for specific types of activity associated with terrorist recruitment or organization of violent activities.[139]

2 Recommendation 2: Distinguish Between the Broadcast/Many-to-Many Functions and Common Carrier/One-to-One; Distinguish Between Passive Listening and Active Participation; and Limit Penalties Imposed for Off-Platform Conduct.

Digital platforms combine, or potentially combine, functions that we have traditionally thought of as telecommunications, i.e., enabling transmission of information from one point to another at the direction of the user, typically in a one-to-one or one-to-few configuration. At other times, platforms replicate what we think of as more media-like functions, making content generated by users available to large numbers (sometimes millions) of people. When we consider the ways in which people use digital platforms, they include absorbing information as passive listeners, participating in online communities, or purchasing goods and services. Many of these activities are quite valuable, but do not involve any sort of content creation.

While digital platforms may not themselves create content, they do influence how easy or hard it is for users to find relevant content. Users themselves might limit access to the content or seek to promote it to a broader audience. At other times, users might seek out content through search engines or by following particular content creators or communities. Additionally, the platform

[138] *National Cable Television Association v. FCC*, 567 F.3d 659 (D.C. Cir. 2009).
[139] The use of artificial intelligence and data processing to predict behavior and determine the risk of illegal activity is highly controversial. It is sometimes called "*Minority Report* policing," in reference to the movie *Minority Report*, in which a combination of technology and human "precogs" arrest people who are predicted to commit crimes before they happen. Numerous studies have shown that efforts to create such predictive-behavior programs typically incorporate the racial and social biases of the developers and of the existing data set (Clough 2018b). Any use of such programs must be subject to considerable scrutiny to ensure that racial profiling or other suspect analysis are not embedded into digital platforms — despite the fact that they are too often reflected in society generally.

itself may recommend particular content, either through its recommendation algorithms or because it was paid to advertise or promote the content. But despite these myriad ways in which users find, absorb, or respond to content, we still think of content moderation and penalties for violating content policies in simplistic terms. Generally, content is either banned or permitted. Similarly, users are either permitted on the platform or banned from the platform. While Facebook and YouTube have begun to experiment with ways to address "borderline" content (YouTube 2019; Zuckerberg 2018), the public debate still generally revolves around banning content or content creators, and whether to make such bans temporary or permanent.

As platforms have become increasingly central to our economic and social lives, we need to recognize that such draconian penalties are both increasingly difficult to enforce and increasingly harmful when misapplied. As we have acknowledged in the realm of economic regulation, we should likewise acknowledge in the realm of content moderation that participation on digital platforms — or, at least, on those platforms that we can classify as dominant in some way — is important for a wide variety of reasons that have nothing to do with speaking. In a world where officials hold debates and make announcements through platforms such as YouTube or WhatsApp, or where important messages about services or public safety are issued in real time through Facebook or Twitter, cutting off the ability of people to follow these developments or receive necessary information is far harsher a penalty than we would permit for comparable activities. A court may punish someone for harassing someone by phone and may issue an injunction to require an individual cease any communication with another specific person (or specific group). But courts do not issue injunctions preventing offenders from ever using a telephone again for any purpose. We recognize that the telephone has too many important uses to treat telephone use as a privilege rather than as a right.

We should therefore distinguish between the "broadcast" functions of digital platforms, the "communications functions" of digital platforms, and their passive information collection or marketplace functions unrelated to communications. We should require that platforms adjust their penalties according to the nature of the offense, as well as establish clear criteria by which a person can regain full privileges. Times and people change, and the greater the penalty the more reluctant we should be generally to impose it permanently.

The same is true for content. Content inappropriate for mass audiences may be permitted, or even protected, between willing adults. Erotic content may be offensive to many, including advertisers, but it is still protected by the First Amendment and welcome between willing speakers. Recently, Facebook announced that it would no longer treat content as simply prohibited or permissible. The more "borderline" the content, *i.e.*, the closer to violating Facebook's content guidelines, the more Facebook will degrade the content in its search and recommendation

algorithms, making it harder to find for anyone who is not already aware of the content and actively seeking it (Zuckerberg 2018). YouTube has announced it will screen its recommendations to avoid recommending videos that promote conspiracy theories (YouTube 2019). This approach balances protecting users from unwanted content, and society generally from the promotion of harmful content, while still allowing speakers to speak and willing listeners to hear.

While these examples involve social media, we can apply them to other platforms, such as review sites. Parties can lose their privileges to post reviews for violating review guidelines (such as failure to disclose a financial interest) without losing the ability to read reviews or purchase products. Social media sites might also make exceptions for specific purposes, such as interactions with official accounts for government departments or officials, or to speak directly to public safety. Of course, these exceptions would also be subject to revocation in cases of abuse. False alarms spread through Twitter are no different from false fire alarms or false 911 calls. But we should be as reluctant to ban people from reaching out to public safety through social media as we are to prohibit people from using 911.

Finally, it is important to distinguish between actions on the digital platform and actions off the digital platform. As use of digital platforms becomes increasingly necessary to engage in society generally, it becomes increasingly inappropriate to regard participation as a reward for good behavior. It is appropriate to address the behavior of someone advocating violence and hate through their Instagram account. But if someone is a neo-Nazi in their offline time but uses their Instagram account purely to post pictures of puppies, there is no reason to treat their use of Instagram as a privilege to be revoked because they are a bad person. We do not ask mobile carriers to revoke the subscriptions of neo-Nazis simply because they are bad people and don't "deserve" to talk on the phone. We should similarly not require (or, in the case of dominant platforms, permit) platforms to make moral judgments about who is or is not intrinsically worthy to participate in digital society generally. To paraphrase Gilbert and Sullivan, our object should be to make the punishment fit the crime — no more, no less.

Taking this together, we may establish a simple hierarchy of rules, subject to rights of appeal within the platform and/or to an oversight agency. If the offensive speech is directed against individuals, such as repeated harassment of speakers, then the platform should revoke the ability of the harasser to reply to individuals, comment on content, or participate in public forums. The more egregious the conduct the broader the ban, until a bad actor may be reduced to a purely passive listener. Additionally, platforms should be required to give individuals the ability to block specific other individuals from commenting, as well as completely block them from seeing or being seen by the individual in question. Individuals should have the option to permit a racist relative or militant ideologue to continue to follow them but not but not be able to respond. But platforms themselves

(or an enforcement agency) should have the same ability to mute someone whose conduct is so toxic that it undermines the utility of the platform for others.

Similarly, the closer content comes to violating the platform's standards of conduct, the harder it should become to promote such conduct. As discussed above, some content falls outside the scope of societal norms and protections and should be taken down, such as libelous or fraudulent content (including "fake news"). But we may anticipate many "borderline cases" involving protected content where violation of content standards is difficult to judge. Accusing someone of behaving like a jackal or a pig could simply be an insult to an individual for their specific behavior, or dehumanizing a group based on racist stereotypes. Understanding the nature and intent of speech requires context. Bad decisions have punished victims of dehumanizing content for responding while leaving the initial harassing content untouched (Jeong 2018; van Zuylen-Wood 2019). This is particularly true in the context of social media, where exchanges may be technically open to the general public but have the feel and quality of personal conversations.

It has proven unworkable and unsatisfying to ask platforms to decide whether to take down such content or leave it alone. Graduated response to both the content and the content creator moderates the danger of banning controversial but acceptable speech, while still permitting platforms to ban content or speakers who ultimately prove toxic.

Finally, we may require platforms to take steps to prevent conversion of applications intended for communications to more broadcast-like functions. To address the spread of false material designed to incite racial violence, WhatsApp limits the ability to forward a message to no more than five times. This does not ban any specific content, but introduces friction into the spread of content that may be deliberately calculated to foment violence or manipulate markets. This does impose a potential problem for emergency speech or other content that should spread as quickly as possible. But all moderation has tradeoffs. Providing platforms an incentive to differentiate clearly between their point-to-point communications services and broadcast-like services may help to prevent rapid proliferation of harmful content while minimizing the burden on non-harmful speech and innovation.

3 Recommendation 3: Determining the Goal of Regulation of Bad Content and Measuring Its Effectiveness.

As discussed above in the context of DMCA and NetzDG, society retains a strong interest in monitoring whether rules requiring content moderation achieve their goals, and at what cost. Often, however, we fail to articulate clearly which of many possible goals we are trying to achieve. Unsurprisingly, without an understanding of what we are trying to do, we cannot measure whether

we are, in fact, achieving it. To make matters worse, the metrics we select to measure the effects will drive behavior. If we prioritize speed of complaint-resolution, for example, we will prioritize resolving complaints quickly rather than correctly.

Content moderation regulations can have many different goals. If our primary concern is to avoid radicalization, that is different from detecting potential violent actors or protecting individuals from unwanted content or harassment. Of course, we often pass laws with more than one goal in mind. But whether we have a specific primary goal or multiple goals, we need to articulate them clearly and adopt proper metrics measuring success or collateral harm. For example, whether SESTA is intended to prevent sex trafficking, sex work generally, or both, is important for determining whether proponents or critics are correct about the law's effectiveness. If we intend to follow a specific model, we ought to have some notion of how it works.

So far, the evidence is mixed about the effectiveness of "deplatforming," or depriving a user of a given platform, and about how to set realistic goals and expectations. Research shows that in the case of high-profile individuals like Alex Jones, banning them from popular platforms does deprive them of audience and reduces their impact (Koebler 2018). There is also, however, considerable evidence that speech bans by platforms are easily evaded and manipulated, with victims targeted as retaliation for effective reporting or whistleblowing — which suggests that existing policies may have significant costs even to people these policies are designed to protect (Jeong 2018; Maynes 2019). Deplatforming can also be disruptive to communities generally (Feld 2018d), though there is evidence that over time community members will seek new platforms on which to re-form their communities (Schwedel 2018). This is not necessarily a bad thing. When Facebook decided to ban nudity regardless of context, nudists migrated to Twitter to share non-sexual nude content related to nudism, nudist lifestyle, and nudist philosophy (Lorenz 2018). On the other hand, evidence also shows that the same is true for creators of racist content, violent content and other disturbing content (Lord and Murray 2019).

Designating an enforcement agency to track the effectiveness and unintended consequences of content moderation regulations is an important safety mechanism. Information collection can alert policymakers to the need to modify an initial policy or validate the success of a specific approach. While it may not be possible to consider all possible effects and therefore provide for metrics to answer all questions, empowering an agency to collect data and provide constant oversight is an important mechanism for such efforts.