ICLE International Center for Law & Economics

Who Moderates the Moderators?: A Law & Economics Approach to Holding Online Platforms Accountable Without Destroying the Internet
November 9, 2021
Geoffrey A. Manne, Kristian Stout and Ben Sperry

Executive Summary

A quarter-century since its enactment as part of the Communications Decency Act of 1996, a growing number of lawmakers have been seeking reforms to Section 230. In the 116th Congress alone, 26 bills were introduced to modify the law's scope or to repeal it altogether. Indeed, we have learned much in the last 25 years about where Section 230 has worked well and where it has not.

Although the current Section 230 reform debate popularly—and politically—revolves around when platforms should be forced to host certain content politically favored by one faction (i.e., conservative speech) or when they should be forced to remove certain content disfavored by another (i.e., alleged "misinformation" or hate speech), this paper does not discuss, nor even entertain, such reform proposals. Rather, such proposals are (and should be) legal non-starters under the First Amendment.

Indeed, such reforms are virtually certain to harm, not improve, social welfare: As frustrating as imperfect content moderation may be, state-directed speech codes are much worse. Moreover, the politicized focus on curbing legal and non-tortious speech undermines the promise of making any progress on legitimate issues: The real gains to social welfare will materialize from reforms that better align the incentives of online platforms with the social goal of deterring or mitigating illegal or tortious conduct.

Section 230 contains two major provisions: (1) that an online service provider will not be treated as the speaker or publisher of the content of a third party, and (2) that actions taken by an online service provider to moderate the content hosted by its services will not trigger liability. In essence, Section 230 has come to be seen as a broad immunity provision insulating online platforms from liability for virtually all harms caused by user-generated content hosted by their services, including when platforms might otherwise be deemed to be implicated because of the exercise of their editorial control over that content.

To the extent that the current legal regime permits social harms online that exceed concomitant benefits, it should be reformed to deter those harms if such reform can be accomplished at sufficiently low cost. The salient objection to Section 230 reform is not one of principle, but of practicality: are there effective reforms that would address the identified harms without destroying (or excessively damaging) the vibrant Internet ecosystem by

imposing punishing, open-ended legal liability? We believe there are.

First and foremost, we believe that Section 230(c)(1)'s intermediary-liability protections for illegal or tortious conduct by third parties can and should be conditioned on taking reasonable steps to curb such conduct, subject to procedural constraints that will prevent a tide of unmeritorious litigation.

This basic principle is not without its strenuous and thoughtful detractors, of course. A common set of objections to Section 230 reform has grown out of legitimate concerns that the economic and speech gains that have accompanied the rise of the Internet over the last three decades would be undermined or reversed if Section 230's liability shield were weakened. Our paper thus establishes a proper framework for evaluating online intermediary liability and evaluates the implications of the common objections to Section 230 reform within that context. Indeed, it is important to take those criticisms seriously, as they highlight many of the pitfalls that could attend imprudent reforms. We examine these criticisms both to find ways to incorporate them into an effective reform agenda, and to highlight where the criticisms themselves are flawed.

Our approach is rooted in the well-established law & economics analysis of liability rules and civil procedure, which we use to introduce a framework for understanding the tradeoffs faced by online platforms under differing legal standards with differing degrees of liability for the behavior and speech of third-party users. This analysis is bolstered by a discussion of common law and statutory antecedents that allow us to understand how courts and legislatures have been able to develop appropriate liability regimes for the behavior of third parties in different, but analogous, contexts. Ultimately, and drawing on this analysis, we describe the contours of our recommended duty-of-care standard, along with a set of necessary procedural reforms that would help to ensure that we retain as much of the value of user-generated content as possible, while encouraging platforms to better police illicit and tortious content on their services.

The Law & Economics of Online Intermediary Liability

An important goal of civil tort law is to align individual incentives with social welfare such that costly behavior is deterred and individuals are encouraged to take optimal levels of precaution against risks of injury. Not uncommonly, the law even holds intermediaries—persons or businesses that have a special relationship with offenders or victims—accountable when they are the least-cost avoider of harms, even when those harms result from the actions of third parties.

Against this background, the near-complete immunity granted to online platforms by Section 230 for harms caused by platform users is a *departure* from normal rules governing intermediary behavior. This immunity has certainly yielded benefits in the form of more user-generated online content and the ability of platforms to moderate without fear of liability. But it has also imposed costs to the extent that broad immunity fails to ensure that illegal and tortious conduct are optimally deterred online.

The crucial question for any proposed reform of Section 230 is whether it could pass a cost-benefit test—that is, whether it is likely to meaningfully reduce the incidence of unlawful or tortious online content while sufficiently addressing the objections to the modification of Section 230 immunity, such that its net benefits outweigh its net costs. In the context of both criminal and tort law generally, this balancing is sought through a mix of direct and collateral enforcement actions that, ideally, minimizes the total costs of misconduct and enforcement. Section 230, as it is currently construed, however, eschews entirely the possibility of collateral liability, foreclosing an important mechanism for properly adjusting the overall liability scheme.

But there is no sound reason to think this must be so. While many objections to Section 230 reform—that is, to the imposition of any amount of intermediary liability—are well-founded, they also frequently suffer from overstatement or unsupported suppositions about the magnitude of harm. At the same time, some of the expressed concerns are either simply misplaced or serve instead as arguments for broader civil-procedure reform (or decriminalization), rather than as defenses of the particularized immunity afforded by Section 230 itself.

Unfortunately, the usual course of discussion typically fails to acknowledge the tradeoffs that Section 230—and its reform—requires. These tradeoffs embody value judgments about the quantity and type of speech that should exist online, how individuals threatened by tortious and illegal conduct online should be protected, how injured parties should be made whole, and what role online platforms should have in helping to negotiate these tradeoffs. This paper's overarching goal, even more important than any particular recommendation, is to make explicit what these tradeoffs entail.

Of central importance to the approach taken in this paper, our proposals presuppose a condition frequently elided by defenders of the Section 230 status quo, although we believe nearly all of them would agree with the assertion: that there is actual harm—violations of civil law and civil rights, violations of criminal law, and tortious conduct—that occurs on online platforms and that imposes real costs on individuals and society at-large. Our proposal proceeds on the assumption, in other words, that there are very real, concrete benefits that would result from demanding greater accountability from online intermediaries, even if that also leads to "collateral censorship" of some lawful speech.

It is necessary to understand that the baseline standard for speech and conduct—both online and offline—is not "anything goes," but rather self-restraint enforced primarily by incentives for deterrence. Just as the law may deter some amount of speech, so too is speech deterred by fear of reprisal, threat of social sanction, and people's baseline sense of morality. Some of this "lost" speech will be over-deterred, but one hopes that most deterred speech will be of the harmful or, at least, low-value sort (or else, the underlying laws and norms should be changed). Moreover, not even the most valuable speech is of infinite value, such that any change in a legal regime that results in relatively less speech can be deemed *per se* negative.

A proper evaluation of the merits of an intermediary-liability regime must therefore consider whether user liability alone is insufficient to deter bad actors, either because it is too costly to pursue remedies against users directly, or because the actions of platforms serve to make it less likely that harmful speech or conduct is deterred. The latter concern, in other words, is that intermediaries may—intentionally or not—facilitate harmful speech *that would otherwise be deterred (self-censored) were it not for the operation of the platform*.

Arguably, the incentives offered by each of the forces for self-restraint are weakened in the context of online platforms. Certainly everyone is familiar with the significantly weaker operation of social norms in the more attenuated and/or pseudonymous environment of online social interaction. While this environment facilitates more legal speech and conduct than in the offline world, it also facilitates more illegal and tortious speech and conduct. Similarly, fear of reprisal (i.e., self-help) is often attenuated online, not least because online harms are often a function of the multiplier effect of online speech: it is frequently not the actions of the original malfeasant actor, but those of neutral actors amplifying that speech or conduct, that cause harm. In such an environment, the culpability of the original actor is surely mitigated and may be lost entirely. Likewise, in the normal course, victims of tortious or illegal conduct and law enforcers acting on their behalf are the primary line of defense against bad actors. But the relative anonymity/pseudonymity of online interactions may substantially weaken this defense.

Many argue, nonetheless, that holding online intermediaries responsible for failing to remove offensive content would lead to a flood of lawsuits that would ultimately overwhelm service providers, and sub-optimally diminish the value these firms provide to society—a so-called “death by ten thousand duck-bites.” Relatedly, firms that face potentially greater liability would be forced to internalize some increased—possibly exorbitant—degree of compliance costs even if litigation never materialized.

There is certainly some validity to these concerns. Given the sheer volume of content online and the complexity, imprecision, and uncertainty of moderation processes, even very effective content-moderation algorithms will fail to prevent all actionable conduct, which could result in many potential claims. At the same time, it can be difficult to weed out unlawful conduct without inadvertently over-limiting lawful activity.

But many of the unique features of online platforms also cut *against* the relaxation of legal standards online. Among other things—and in addition to the attenuated incentives for self-restraint mentioned above—where traditional (offline) media primarily host expressive content, online platforms facilitate a significant volume of behavior and commerce that isn’t purely expressive. Tortious and illegal content tends to be less susceptible to normal deterrence online than in other contexts, as individuals can hide behind varying degrees of anonymity. Even users who are neither anonymous nor pseudonymous can sometimes prove challenging to reach with legal process. And, perhaps most importantly, online content is disseminated both faster and more broadly than offline media.

At the same time, an increase in liability risk for online platforms may lead not to

insurmountable increases in litigation costs, but to other changes that may be less privately costly to a platform than litigation, and which may be *socially* desirable. Among these changes may be an increase in preemptive moderation; smaller, more specialized platforms and/or tighter screening of platform participants on the front end (both of which are likely to entail stronger reputational and normative constraints); the establishment of more effective user-reporting and harm-mitigation mechanisms; the development and adoption of specialized insurance offerings; or any number of other possible changes.

Thus the proper framework for evaluating potential reforms to Section 230 must include the following considerations: To what degree would shifting the legal rules governing platform liability increase litigation costs, increase moderation costs, constrain the provision of products and services, increase "collateral censorship," and impede startup formation and competition, all *relative to the status quo*, not to some imaginary ideal state? Assessing the marginal changes in all these aspects entails, first, determining how they are affected by the current regime. It then requires identifying both the direction and magnitude of change that would result from reform. Next, it requires evaluating the corresponding *benefits* that legal change would bring in increasing accountability for tortious or criminal conduct online. And, finally, it necessitates hazarding a best guess of the *net* effect. Virtually never is this requisite analysis undertaken with any real degree of rigor. Our paper aims to correct that.

A Proposal for Reform

What is called for is a properly scoped reform that applies the same political, legal, economic, and other social preferences offline as online, aimed at ensuring that we optimally deter illegal content without losing the benefits of widespread user-generated content. Properly considered, there is no novel conflict between promoting the flow of information and protecting against tortious or illegal conduct online. While the specific mechanisms employed to mediate between these two principles online and offline may differ—and, indeed, while technological differences can alter the distribution of costs and benefits in ways that must be accounted for—the fundamental principles that determine the dividing line between actionable and illegal or tortious content offline can and should be respected online, as well. Indeed, even Google has argued for exactly this sort of parity, recently calling on the Canadian government to "take care to ensure that their proposal does not risk creating different legal standards for online and offline environments."

Keeping in mind the tradeoffs embedded in Section 230, we believe that, in order to more optimally mitigate truly harmful conduct on Internet platforms, intermediary-liability law should develop a "duty-of-care" standard that obliges service providers to reasonably protect their users and others from the foreseeable illegal or tortious acts of third parties. As a guiding principle, we should not hold online platforms vicariously liable for the speech of third parties, both because of the sheer volume of user-generated content online and the generally attenuated relationship between online platforms and users, as well as because of the potentially large costs to overly chilling free expression online. But we should place at least the same burden to curb unlawful behavior on online platforms that we do on traditional media operating offline.

Nevertheless, we hasten to add that this alone would likely be deficient: adding an open-ended duty of care to the current legal system could generate a volume of litigation that few, if any, platform providers could survive. Instead, any new duty of care should be tempered by procedural reforms designed to ensure that only meritorious litigation survives beyond a pre-discovery motion to dismiss.

Procedurally, Section 230 immunity protects service providers not just from liability for harm caused by third-party content, but also from having to incur substantial litigation costs. Concern for judicial economy and operational efficiency are laudable, of course, but such concerns are properly addressed toward minimizing the costs of litigation in ways that do not undermine the deterrent and compensatory effects of meritorious causes of action. While litigation costs that exceed the minimum required to properly assign liability are deadweight losses to be avoided, the cost of liability itself—when properly found—ought to be borne by the party best positioned to prevent harm. Thus, a functional regime will attempt to accurately balance excessive litigation costs against legitimate and necessary liability costs.

In order to achieve this balance, we recommend that, while online platforms should be responsible for adopting reasonable practices to mitigate illegal or tortious conduct by their users, they should not face liability for *communication* torts (e.g., defamation) arising out of user-generated content unless they fail to remove content they knew or should have known was defamatory. Further, we propose that Section 230(c)(2)'s safe harbor should remain in force and that, unlike for traditional media operating offline, the act of reasonable content moderation by online platforms should not, by itself, create liability exposure.

In sum, we propose that Section 230 should be reformed to incorporate the following high-level elements, encompassing two major components: first, a proposal to alter the underlying intermediary-liability rules to establish a "duty of care" requiring adherence to certain standards of conduct with respect to user-generated content; and second, a set of procedural reforms that are meant to phase in the introduction of the duty of care and its refinement by courts and establish guardrails governing litigation of the duty.

Proposed Basic Liability Rules

Online intermediaries should operate under a duty of care to take appropriate measures to prevent or mitigate foreseeable harms caused by their users' conduct.

Section 230(c)(1) should not preclude intermediary liability when an online service provider fails to take reasonable care to prevent *non-speech-related* tortious or illegal conduct by its users

As an exception to the general reasonableness rule above, Section 230(c)(1) should preclude intermediary liability for *communication* torts arising out of user-generated content unless an online service provider fails to remove content it knew or should have known was defamatory.

Section 230(c)(2) should provide a safe harbor from liability when an online service provider does take reasonable steps to moderate unlawful conduct. In this way, an online service provider would not be held liable simply for having let harmful content slip through, despite its reasonable efforts.

The act of moderation should not give rise to a presumption of knowledge. Taking down content may indicate an online service provider knows it is unlawful, but it does not establish that the online service provider should necessarily be liable for a *failure* to remove it anywhere the same or similar content arises.

But Section 230 should contemplate "red-flag" knowledge, such that a failure to remove content will not be deemed reasonable if an online service provider knows or should have known that it is illegal or tortious. Because the Internet creates exceptional opportunities for the rapid spread of harmful content, a reasonableness obligation that applies only ex ante may be insufficient. Rather, it may be necessary to impose certain ex post requirements for harmful content that was reasonably permitted in the first instance, but that should nevertheless be removed given sufficient notice.

Proposed Procedural Reforms

In order to effect the safe harbor for reasonable moderation practices that nevertheless result in harmful content, we propose the establishment of "certified" moderation standards under the aegis of a multi-stakeholder body convened by an overseeing government agency. Compliance with these standards would operate to foreclose litigation at an early stage against online service providers in most circumstances. If followed, a defendant could provide its certified moderation practices as a "certified answer" to any complaint alleging a cause of action arising out of user-generated content. Compliant practices will merit dismissal of the case, effecting a safe harbor for such practices.

In litigation, after a defendant answers a complaint with its certified moderation practices, the burden would shift to the plaintiff to adduce sufficient evidence to show that the certified standards were not actually adhered to. Such evidence should be more than mere *res ipsa loquitur*; it must be sufficient to demonstrate that the online service provider should have been aware of a harm or potential harm, that it had the opportunity to cure or prevent it, and that it failed to do so. Such a claim would need to meet a heightened pleading requirement, as for fraud, requiring particularity.

Finally, we believe any executive or legislative oversight of this process should be explicitly scheduled to sunset. Once the basic system of intermediary liability has had some time to mature, it should be left to courts to further manage and develop the relevant common law.

Our proposal does not demand perfection from online service providers in their content-moderation decisions—only that they make reasonable efforts. What is appropriate for YouTube, Facebook, or Twitter will not be the same as what's appropriate for a startup social-media site, a web-infrastructure provider, or an e-commerce platform. A properly

designed duty-of-care standard should be flexible and account for the scale of a platform, the nature and size of its user base, and the costs of compliance, among other considerations. Indeed, this sort of flexibility is a benefit of adopting a “reasonableness” standard, such as is found in common law negligence. Allowing courts to apply the flexible common law duty of reasonable care would also enable the jurisprudence to evolve with the changing nature of online intermediaries, the problems they pose, and the moderating technologies that become available.

Read the full working paper here.

View Article