Platform Liability, Hate Speech, and the Failure of Section 230

Ryleigh Holthus
Old Dominion University
04.11.26

Introduction

With the rise of social media a lot of questions and debate have risen. A big topic of question is whether, and to what extent, do social media sites bear legal responsibility for the harmful, hateful, and discriminatory content that users generate on their platforms. Hate speech is a type of social deviance that refers to the use of offensive language focused on a specific person or a group with a common property (race, ethnicity, religion, sexual identity, etc) (Castaño-Pulgarín et al., 2021). This is where the Communication decency Act of 1996 Section 230 comes into play. Under 47 U.S.C. § 230 (2025) states that it protects online services from being legally liable for the content that is posted by their users. What was enacted as a push to grow the internet industry and to promote good intent content, Section 230 has become the subject of criticism and controversy across the political spectrum due to social media platform communication being amplified. This amplification is not just the news and exciting things that happen through people's day, it also includes the spread of hate speech, violence and facilitates harassment or targeting individuals and communities and the companies are not held liable. With the constant changing of the internet and social media, Section 230 has fallen behind and has become outdated; it needs a targeted reform.

Background

What does Section 230 actually state? The two most notable sections are (c)(1) “Treatment of Publisher or Speaker” and (c)(2) “Civil Liability” (47 U.S.C. § 230, 2025). Treatment of Publisher or Speaker protects the platforms when they do not remove content and protects them if they decide to remove, edit or post content by other users. Civil Liability or “Good Faith Moderation Immunity” protects users for being a “Good Samaritan” and trying to

make the internet safe and maintain that safety. Congress originally passed this law to allow users an interactive space and make their own content moderation decisions while still holding liability to certain limited contexts (Brannon & Holmes, 2021). It protected website creators from being sued for users' content, treating the users as a host of information rather than publishers of speech. It allowed the internet to become what it is today by enabling platforms to exist without intense moderation.

The decision to enact Section 230 was in response to a court decision, Stratton Oakmont Inc v. Prodigy Co (1995). In New York, an online service provider was held liable for their users' posts on their bulletin board as he was seen as the publisher for the cruel posts by users (Stratton Oakmont, Inc. v. Prodigy Services Co., 1995). Congress feared this ruling would deter others from using online platforms so they created Section 230 to eliminate fear and encourage platforms to engage in good faith content with the sites having liability and users having limited liability. Congress thought this new act could encourage more user friendly interactions and grow the internet, however; Kosseff (2019) believes that Section 230 is rooted in a “legal paradox” and benefits the unmediated platforms rather than the ones that try to create a healthy and positive environment. Now although this law grew the internet it also grew hate, and fostered an environment where people could post hateful content with no reprimand.

Legislation and Legal Framework

Stated above, Section 230 was enacted as part of the Communications Decency Act of 1996 and originally was created to regulate negative content on the internet. Section 230(c)(1) is focused on whether the users nor providers shall be held liable as a publisher or speaker for third-party content. This means that platforms like TikTok, Instagram and Youtube cannot be

sued for what their users post even if the platform knows the content is harmful. Section 230(c)(2) provides a second immunity for platforms. It states that platforms cannot be held liable for good-faith decisions to remove or restrict content they find offensive. This means that they have no obligation to mediate and edit or remove harmful content. Together these two provisions create a shield for platforms whether they moderate or choose not to. This statute does mention several exceptions where this immunity does not apply. It mentions that platforms can still face federal prosecution, can be held liable for facilitating sex trafficking (FOSTA-SESTA, 2018) , can still be sued for copyright and trademark as well as Electronic Communications Privacy Act violations (47 U.S.C. § 230, 2025). As you can see, there are no hate speech exceptions.

Section 230 also has an express preemption clause. It states that it can supersede any inconsistent state or local law (47 U.S.C. § 230, 2025). Courts have applied this preemption quite broadly. They've time and time again dismissed state claims against platforms for negligence, civil right violations, and products liability on the grounds of Section 230. This leads to victims of online hate speech not being able to find relief in the state-law claims as the federal immunity in Section 230 overrides (Kosseff, 2019). Some states have tried to work around but with limited success. Some states have tried framing claims around platform design defects like the algorithm itself is causing harm like in Lemmon v. Snap, Inc. or Force v. Facebook. Courts have been divided on whether design defect claims fall under Section 230 therefore causing inconsistent outcomes. The preemption issues mean reform must come from Congress rather than state legislatures or state courts.

In 2021, Florida (Florida SB 7072) alongside Texas (Texas HB 20), enacted laws attempting to restrict large social media platforms from moderating the speech of certain political figures and viewpoints. Florida prohibited platforms from defacing political candidates and

required consistent moderation policies while Texas went further with this and prohibited platforms from censoring users based solely on viewpoint (Hardig, 2022). Both laws were challenged as unconstitutional violations of the platform's first amendment right. They argued that their content moderation decisions were theirs to make and to protect their editorial content. This case blossomed into Moody v. NetChoise and NetChoice v. Paxton (2024). State level attempts to work around or alter Section 230 fell short showing that the federal legislative path could result in change.

Issues and Problems

Online hate speech is not just an offensive happening on the internet; research has documented direct links between platform hate speech and real world violence (Bradshaw & Howard, 2019). In 2018 the Pittsburgh synagogue shooting happened. It was carried out by someone who had been heavily active on a platform called *Gab* which is known for hosting extremists and hate-filled content with very minimal moderation. The perpetrators' post escalated in violence and hatred over the days before the attack which was visible to anyone on the platform. The case raises direct questions about whether the platform, that knowingly hosted this content and failed to act on the escalating violent hate speech, bears any responsibility for the harm done. In 2019 there was an attack in New Zealand at the Christ Church mosque which killed 51 people and was also carried out by a perpetrator that posted radicalized content on platforms like 8Chan and Facebook. The man livestreamed the attack on Facebook which attracted thousands of people before the removal of the livestream. New Zealand and other nations responded by creating the ChristChurch Call which is an international agreement to remove terrorist and violent extremist content online. Just these two cases show how online hate

speech is not kept in the virtual world but is carried over and causes physical harm to innocent people. Under Section 230 victims and families in these cases, especially American cases, have no legal claim against the platforms who kept the harm on their site.

The Facebook Papers are a great example of a company knowingly having hate speech on their platform yet not doing anything about it. In 2021 a former Facebook employee Frances Haugen leaked thousands of pages from the company's research documents to journalists and Congress and this is known as the Facebook Papers. The documents revealed that Facebook's personal researchers had its algorithms system amplifying hate speech and misinformation content at a scale the company's moderation system was unable to effectively address. The idea is that they didn't seem to have this as a priority to address because it created traction for their website and engaged users (Bradshaw & Howard, 2019). In this leak it came to light that leadership was repeatedly told that their design was making the platform a more hostile and dangerous environment for minority and marginalized communities. The company also had evidence that the Instagram platform was causing psychological harm to teen girls but chose not to make the design change that would reduce this engagement. The Facebook Papers play a significant role because it demonstrates that these platforms are not passive or unknowingly hosting harmful content; they have detailed information of the harm their system causes and choose profit over the user's safety (Citron & Wittes, 2017). This knowledge undermines the argument that Section 230's immunity is necessary to protect these platforms from liability from the harms they could not have anticipated or prevented.

One of the more serious consequences of Section 230's broad immunity is that individuals harmed by these platforms that host hate speech have almost no viable legal relief. Courts have consistently applied Section 230 to dismiss cases where: platforms notice that

harmful content was targeting the victim and refuse to move it, the platform's own algorithm recommendation system directed hate speech to the victim, and the platform previously promised in terms of service to remove the type of content at issue and failed to do so (Citron & Wittes, 2017). This resulted in victims who have been targeted by hate campaigns, receiving death threats, etc on social media platforms cannot get the justice they deserve and sue the platform for the harm they've caused regardless of what the platform knew or did or didn't do. The result is an absence of accountability to these platforms whose systems keep facilitating serious harm to identifiable individuals who speak up. The only resource available to victims is to report content to the platform and hope for the removal which is known to be inconsistent, slow and infrequently ineffective for hate speech targeting minority communities (Citron & Wittes, 2017).

Pros and Cons / Current State and Solutions

Some pros for Section 230 are that broad immunity prevents over censorship and protects expression and platforms have invested heavily in moderation infrastructures because of this protection. One of the strongest arguments in favor of keeping Section 230 would be its alternative, not having Section 230 and platforms face liability for their users' content. This would create an overwhelming incentive to over-censor its users' posts. The platform could be sued every time a user posts something harmful, which *is* subjective, but, the response would be to remove any content that could in any way shape or form create liability which would restrict the online environment. Goldman (2019) argues that section 230 effectively functions better than the first amendment itself and protects open expression because it removes litigation as a tool for silencing unfavorable speech. With this, Section 230's immunity was also designed to encourage platforms to engage in good faith content moderation without the fear that the moderation

decisions would expose them to publisher liability. Major platforms invest billions of dollars and their moderation infrastructure including AI and human moderators, partnerships with organizations and academic researchers and providing public data on content removal decisions (Goldman 2019). The idea is that without Section 230 companies would have no incentive to moderate their platform and would adopt a hands-off approach to avoid any conduct that could put liability on them.

Where there are pros there are cons. Section 230 is that the algorithmic amplification is direct conduct of the platform. There is a fundamental difference between a platform passively hosting user-generated content and a platform actively pushing its own algorithmic system to recommend an amplified content to users. Section 230(c)(1) Gives immunity to the company's acting as publishers for users' content but it does not by any means immunize the platform for their own creation and designing and enforcing their recommendation system. When the platform takes hate speech content from one user and recommends it to thousands of others who are not seeking this type of content they are no longer a passive host, they are an active participant in distributing this content to other people (Citron & Wittes, 2017). Courts have begun to examine whether algorithmic amplification claims can survive Section 230. Some judges are saying that recommendation systems are a platform creation and therefore are not held to Section 230. Reform along these lines would not require platforms to remove all hate speech but it would simply hold them accountable for their choices in engineering and designing these algorithmic systems.

A reform option could look like removable immunity for actively promoting hate speech, this would be targeting more of the background people like engineers in the company that create algorithms. This would leave Section 230's core immunity together but create a specific portion

to remove immunity with the platform's own algorithmic system. This approach targets the specific application that is pushing harm to viewers which is clearly the platform's own doing rather than the user. It would incentivize platforms to redesign recommendation systems to avoid promoting hate speech without requiring certain users that partake in hate speech to limit their viewpoint. Several Scholars have identified algorithmic applications as the most legally and practically promising target for Section 230 reform like Bradshaw and Howard (2019).

Conclusion

Section 230 was the right policy to enact in 1996 but it has severely fallen behind in the ever-changing technology world. These social media companies are abusing the immunity given to them by Section 230 and are invalidating victims who have been harmed and continue to be harmed by social media and other internet platforms. You can't change everything in one day but something that could last a couple years could help many people, something like targeting algorithmic recommendation systems. Without this change, big companies will continue to abuse the immunity and fellow users will continue to be affected by what has not been changed.

References

Caroline Hardig. (2022, October 6). *How should courts treat social media platforms under the First Amendment?*. University of Cincinnati Law Review Blog. https://uclawreview.org/2022/10/06/how-should-courts-treat-social-media-platforms-under-the-first-amendment/

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M., & López, H. M. (2021). Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Citron, D., & Wittes, B. (n.d.). *Fordham Law Review fordham law review volume 86 issue 2 article 3 2017*. Fordham Law Review Archive. https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5435&context=flr

Force v. Facebook, Inc., 934 F.3d 53 (2d Cir. 2019)

Goldman, E. (2019). Why section 230 is better than the First Amendment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3351323

Kosseff, J. (2019). *The twenty-six words that created the internet*. Cornell University Press.

Legal Information Institute. (n.d.). *47 U.S. Code § 230 - protection for private blocking and screening of offensive material*. Legal Information Institute. https://www.law.cornell.edu/uscode/text/47/230

Lemmon v. Snap, Inc., 995 F.3d 1085 (9th Cir. 2021)

Merrefield, C., & About The Author Clark Merrefield Clark Merrefield is the senior editor for economics and legal systems at The Journalist's Resource. He joined JR in 2019 after working as a reporter for Newsweek and The Daily Beast. (2023, November 4). *FBarchive, a searchable repository of Facebook whistleblower documents*. The Journalist's Resource. https://journalistsresource.org/home/fbarchive-new-reporting-tool/

Wikimedia Foundation. (2020, December 30). *Stratton Oakmont, inc. v. Prodigy Services Co*. Wikipedia. https://en.wikipedia.org/wiki/Stratton_Oakmont,_Inc._v._Prodigy_Services_Co