Insight

Content Moderation Using Notice and Takedown Systems: A Paradigm Shift in Internet Governance

JUAN LONDOÑO | NOVEMBER 8, 2021

Executive Summary

- "Content moderation" – the removal of undesirable, offensive or illegal user content from online platforms – has emerged as a potent political issue and important policy challenge.
- The current approach (Section 230) shields platforms from legal liability for user content; a variety of proposed bills would place higher legal liabilities on platforms, typically through "notice and takedown" systems.
- Notice and takedown systems require platforms to remove content in a timely manner after receiving a notification that the content is potentially unacceptable, and have been implemented in the context of copyright infringement and sex-trafficking content.
- While the Section 230 regime has fostered free speech and innovation online, experience with notice and takedown systems suggests that they are a threat to these objectives.

Introduction

Content moderation of online platforms – particularly those known as "Big Tech"– is a potent political issue and an important policy challenge. Republican members of the House Energy & Commerce Committee have introduced a broad package of draft bills seeking to reform Section 230 of the Communications Decency Act, a linchpin of internet content moderation, while Democratic Chairman of the House Judiciary Committee Jerrold Nadler introduced the SHOP SAFE Act, which aims to increase platforms' liability regarding the sale of counterfeit goods on their websites.

Section 230 provides platforms with a shield from liability so they can freely moderate content, placing the liability of the content on the users who create and post it. This has contributed to innovation in online content. An alternative approach to content moderation, used in the Digital Millennium Copyright Act (DMCA), the Fight Online Sex Trafficking Act (FOSTA), and the Stop Enabling Sex Traffickers Act (SESTA) is a "notice and takedown" system. Such a system places a higher responsibility on platforms, as they must remove offending content after a notification has been placed.

This insight examines these regimes and concludes that notice and takedown systems have resulted in over-moderation of content, thus suppressing innovation and speech online.

Content Moderation Under Section 230

Section 230 of the Communications Decency Act establishes that online platforms should not be treated as the publisher or speaker of any information provided by other content providers. In other words, online platforms will only be legally liable for the content they publish themselves, exempting them from legal liability for content posted by their users. This prevents platforms from receiving a potentially jeopardizing lawsuit from hosting users' content, therefore making online interaction less risky and more cost-effective.

This system benefits both platforms and users alike. The benefit for platforms, especially startups, is clear. This approach allows platforms to allocate more resources in the development of the core product, instead of needing to spend significant resources on a robust content-moderation strategy and a legal team to fight the consequences of any shortcomings that said strategy could present. Essentially, Section 230 provides a low barrier of entry for internet startups, as it eliminates the liability risk associated with hosting user-generated content.

For users, the main benefit of content moderation under Section 230 is that the content they wish to post online is more easily hosted. As platforms have less fear of legal liability, they will have more lenient review processes, and are less likely to subject content to a screening before it is posted. This approach allows users to be able to experiment and innovate with the content they generate, pushing platforms toward unintended content. This greatly benefits users trying to post more content considered counter-cultural or controversial, as platforms face lower legal ramifications for hosting it. As Democratic Senator Ron Wyden – one of the co-authors of the Communications Decency Act – has expressed, the bill took the principles of the First Amendment and applied them to internet governance.

Notice and Takedown Systems

Bills such as the DMCA, FOSTA, and SESTA have established what is known as a notice –and takedown system. As its name describes, it establishes that platforms must remove offending content in a timely matter after it has received a notice of infraction. The reasoning behind the system is that by notifying platforms, they will be aware of the existence of potentially problematic content, and they can quickly review and remove it if they find it violates either copyright law – in the case of the DMCA– or promotes sex trafficking – in the case of FOSTA and SESTA. Therefore, these bills create a completely different paradigm from Section 230: As platforms have been notified of potentially problematic content but have not removed it, they are now responsible for it, regardless of whether they or a third party published it.

Practical Challenges of Notice and Takedown Systems

On paper, notice and takedown systems seem to establish a clear path of action and a seemingly easy approach to combat offending content. In practice, however, its application has shown various unintended consequences. First, it pushes platforms to take an overly precautionary approach, preemptively removing content that may not be offending but has been flagged as so. This has been especially prominent in the case of copyright, in which the sheer number of complaints platforms receive makes it almost impossible to appropriately review content in a timely manner. As regulations establish that content must be removed in a timely matter or platforms will face onerous fines, platforms are incentivized to remove content first, and review it later.

This “remove first, review later” approach leads to the emergence of a second problem: the “guilty until proven innocent” regime. As platforms remove the content preemptively in order to ensure they are complying with the timeliness requirement, users are now the ones to bear the burden of proving their content is non-infringing. Thus, platforms will receive a sizeable number of appeals at a time, leading to an often-lengthy process of review before content may be allowed to be reposted. The burden of timely content review is shifted from the initial notice to the appeal process, harming users and benefiting notifiers. This shift in the burden of proof has become ripe for the use of notifications of copyright infringement as a method to extort or silence content creators, which see platforms de-monetize or outright shut down their profiles as a response to infringement notices.

The third issue is that notice and takedown has generated a need for further implementation of algorithms to speed up the content removal process. As platforms receive notifications for the millions of pieces of user-generated content that are posted every day, they have had to ramp up the automation process in order to remove potentially infringing content in a prompt and cost-effective manner. But the implementation of algorithms raises concerns. The main concern is the question of accuracy, as the artificial intelligence (AI) that powers these algorithms often lacks the ability to discern the context of the post. These algorithms also have to go through a lengthy trial and error process, which would theoretically allow the AI to develop accurate judgement criteria, a process industry insiders are skeptical about. Algorithms have also proven vulnerable to exploitation, as in the cases where law enforcement officers played copyrighted music in order to prevent civilians’ recordings from being uploaded to online platforms.

Looking at other countries that lack or have repealed analogous regulations, the impact of moving away from Section 230 becomes clearer. Australia’s high court ruled that media outlets are to be considered “publishers” of allegedly defamatory comments in their comments section on social media. Many media outlets and public figures reacted by disabling their comments section, alleging that due to the 24/7 nature of social media and the sheer volume of comments, maintaining constant, flawless moderation of comments would be impossible. Therefore, it was easiest to remove the comment section to shield themselves from legal liability.

Conclusion

Legislative efforts to improve online content moderation have focused on altering Section 230 of the Communications Decency Act, a key piece of legislation establishing that online platforms are only legally liable for the content they publish themselves. A review of the performance of the primary alternative to Section 230 –notice and takedown systems– indicates that the result is content over-moderation and further implementation of algorithms and automated systems. Policymakers should keep in mind that the principles behind Section 230 foster a dynamic online environment, as it allows platforms to host user-generated content at a lower cost and reduces the barriers to entry to new competitors.