WIKIMEDIA
FOUNDATION

October 15, 2019

Re: Fostering a Healthier Internet to Protect Consumers

Dear Chairman Pallone, Ranking Member Walden, Chairman Doyle, Ranking Member Latta, Chairwoman Schakowsky, Ranking Member McMorris, and Members of the Committee:

The Wikimedia Foundation appreciates the opportunity to submit comments for the record on the hearing entitled "Fostering a Healthier Internet to Protect Consumers." As the non-profit organization that hosts and supports Wikipedia, among other projects, the Foundation relies upon the protections of Section 230 in order to exist and operate. Wikipedia could not exist were it not for Section 230. Therefore, proposed changes to the law should be carefully considered to ensure that they do not drastically increase liability for online platforms like Wikipedia, or disincentivize good-faith moderation by staff or volunteers.

Wikipedia has become one of the most visited websites in the world through the efforts of tens of thousands of volunteers who contribute articles, edit them, and reconcile differences between each others' edits. The contents of Wikipedia are therefore truly user-generated, covering a vast number of topics far outside the knowledge, let alone expertise, of any Foundation staff. These edits occur at a rate of 1.8 every second, and while most of them are adding to or refining the knowledge freely available on the site, some will consist of pranks, propaganda, or mistakes, many of which could amount to defamation or create other liabilities that can attach to speech.

The Need for Section 230 and the Necessary Interaction between 230(c)(1) and (c)(2)

Section 230(c)(1) is the reason that Wikipedia can operate with certainty that it will not face extraordinary liability exposure, and it is therefore essential to our mission of ensuring that the world's knowledge can be shared freely. Section 230(c)(2) also ensures that, despite a strong ethic of sharing all types of information, Wikipedia's content can be moderated by its users and administrators.

The community-developed rules for what is permitted in Wikipedia articles may differ in many ways from the types of rules on more conversation-oriented platforms. For example, content policies prohibit editors from including their original research in articles, even if that material is accurate and neutral. Without this rule, Wikipedia would be less of an encyclopedia and more like a forum for users to post the results of their research, with potentially detrimental effects on its reliability and verifiability.

This idiosyncratic rule--one that might appear arbitrary in the abstract--actually shapes Wikipedia as a platform. This is one example of what communication scholar Tarleton Gillespie means when he says that moderation makes platforms: that content moderation is not just a function of a platform, but that it defines what the platform is.[1]

The importance of this is not limited to the idea that section 230(c)(2) should continue to permit flexibility in platforms' content moderation. There is a tendency to separate the two parts of 230(c) and to treat them separately as though each has a completely separate purpose: (c)(1) to provide a limitation on liability, and (c)(2) to encourage content moderation. Regardless of what legislative historians and legal scholars might conclude about this, the practical reality for most online platforms is that the two necessarily work together. The idea that a platform does not stand in the shoes of its users as a speaker does not exist just because evaluating a large volume of content is hard; it also accounts for the fact that the content moderation choices that the platform makes under (c)(2) can always be accused of inconsistency, and cited as evidence that the platform is speaking, not merely moderating users' speech.

It is certainly possible for platforms to speak under the pretext of moderation: in an extreme example, a platform that "moderated" a user comment by removing the word "not" from a user's post would seem to be the platform creating a completely different message, and thus speaking for itself. However, the vast majority of the time, a platform is not trying to stand in the shoes of its users, but ensure that its community operates under a common set of rules or standards: a neighborhood discussion board might disallow national political debates, a sports blog devoted to one team might exclude its rival's fans, or an online encyclopedia might prohibit dictionary-style entries.

Should the many concerns about online content suggest changes to Section 230, the law must still permit this type of flexibility in content moderation, allowing websites, forums, and other content platforms to set their own boundaries for content that will be more restrictive and more focused than what the First Amendment requires the state to permit in the public square. Decisions about online content moderation are made far more often and must be made more quickly than constitutional

[1] Tarleton Gillespie, Custodians of the Internet 6 (2018).

litigation, and platforms developing their policies must have the space and ability to account for bad-faith users who attempt to game the platform's system as it develops.

Different Types of Moderation for Different Types of Harms

As we work to ensure a healthier online environment, the discussions include several different types of harmful content that can appear and spread online. It is crucial to recognize that the differences in these types of content, and the differences in how they cause harm, require that different content moderation strategies apply to each.

For instance, certain types of illegal and harmful content can be definitively identified, fingerprinted, and found if they appear again. Specific files that contain child sexual abuse material (CSAM), for instance, can be identified if copies reappear. However, other types of online harm, such as messages representing stalking or harassment, are not represented by any specific files or strings of text. While both represent serious problems, the type of moderation appropriate for each is substantially different.

Matters can become even more complicated when dealing with other harms, such as terrorist content. While certain specific graphic files can be actively identified automatically, as with CSAM, a general prohibition against "terrorist content" raises many questions about defining terrorism today. Many authoritarian governments are quick to label protest as promoting terror, sedition, or the overthrow of a government; activist groups can be labeled as terrorist organizations by their political opponents. Platforms seeking to reduce the presence of terrorist content on their systems must make difficult judgment calls that cannot be made automatically.

Machine learning and artificial intelligence systems cannot solve many of these inherent ambiguities. Furthermore, the Foundation's experience with machine learning systems confirms that such systems are only as good as their training, which is based on data gathered by humans and is implemented by human annotators, who bring their own biases to the training of the systems. For instance, one recent paper found that a system designed to automatically detect hate speech was biased against African-American speakers.[2]

[2] Maarten Sap, Dallas Card, et al., "The Risk of Racial Bias in Hate Speech Detection," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 1668 (2019).

Moderation of different types of harmful content also requires accounting for the balance of harms should the moderator make the wrong call. In some cases, a system that defaults to quick removal of suspected content is appropriate; in others, quick removal merely creates an easy way for bad faith actors to game the system and remove the speech of disfavored parties. For instance, consider a content moderation practice that removes a residential address that is suspected doxxing (the posting of a private individual's personal information in a way that threatens or creates harm). If the content is in fact harmful doxxing and is allowed to remain up, the harm to the doxxed user can be immense. If, however, the content is innocent but is removed by mistake or misidentification, the harm of that mistake is relatively small. This would suggest a policy that acts quickly on potential doxxing. In contrast, a policy regarding content that is reported for being defamatory could, if mistakenly left up, result in defamation being spread, but, if non-defamatory content is removed by mistake, could result in timely news reporting being suppressed. Policies incentivizing quick removal of suspected defamation could therefore cause more harm than good.

Recognizing the differences in how platforms can and should deal with different types of harms, however, cannot mean merely removing particular types of illegal, or even disfavored-but-legal content from the scope of Section 230. The ability for a platform to identify and deal with a particular piece of content has little to do with the severity of the harm it might cause and more to do with the type of information it is, and how it interacts with the platform. In the course of the 1.8 edits on Wikipedia made per second, any given edit could include a satirically false statement about a public figure, or an offer to sell drugs, but our volunteer moderators' or in-house staff's ability to locate these is the same, regardless of the scope of the potential harm.

Conclusion

Fostering a healthy online environment requires participation in good-faith content moderation on the part of platforms and their users, including moderation of content according to rules that are absent from the law--either because they represent difficult political judgment calls or because those laws would be unconstitutional. Section 230 currently allows that flexibility for rapidly evolving moderation and resists creating affirmative incentives for platforms to ignore problems. To the extent that policymakers and stakeholders want to create new incentives for increased moderation, we should take care that we do not incentivize platforms to take a path of least resistance that causes them to ignore mistakes in moderation or allow the gaming or subversion of moderation systems. As the Committee continues its exploration of these topics, the Wikimedia Foundation remains eager to answer any questions or assist in any way it can.