Top Myths About Content Moderation

blog.ericgoldman.org/archives/2019/10/top-myths-about-content-moderation.htm

October 15, 2019

October 15, 2019 · by Eric Goldman · in Content Regulation, Derivative Liability Edit This
How Internet companies decide which user-submitted content to keep and which to remove—a process called "content moderation"—is getting lots of attention lately, for good reason. Under-moderation can lead to major social problems, like foreign agents manipulating our elections. Over-moderation can suppress socially beneficial content, like negative but true reviews by consumers.

Due to these high stakes, regulators across the globe increasingly seek to tell Internet companies how to moderate content. European regulators are requiring Internet services to remove extremist content within an hour and to install upload filters to prospectively block copyright infringement; and U.S. legislators have proposed to ban Internet services from moderating content at all.

Unfortunately, many of these regulatory efforts are predicated on myths about content moderation, such as:

***Myth*: Content moderation can be done perfectly.**

Reality: Regulators routinely assume Internet services can remove all bad content without suppressing any good content. Unfortunately, they can't. First, mistakes occur when the service lacks key contextual information about the content—such as details about the author's identity, other online and offline activities, and cultural references. Second, any line-drawing exercise creates mistake-prone border cases because users routinely submit "edgy" content. Third, a high-volume service will make many mistakes, even if it's highly accurate—1 billion submissions a day at 99.9% accuracy still yields a million mistakes a day.

***Myth*: Bad content is easy to find and remove.**

Reality: Regulators often assume every item of bad content has an impossible-to-miss flashing neon sign saying "REMOVE THIS CONTENT," but that's rare. Content is often obviously bad only in hindsight or with context unavailable to the service. Regulators' cherry-picked anecdotes don't prove otherwise.

***Myth*: Technologists just need to "nerd harder."**

Reality: Filtering and artificial intelligence play important roles in content moderation. However, technology alone cannot magically solve the problem. "Edgy" and contextless content vexes the machines, too.

***Myth*: Internet services should hire more humans to review content.**

Reality: Humans have biases and make mistakes too, so adding human reviewers won't lead to perfection. Furthermore, human reviewers sometimes experience an unrelenting onslaught of horrible content to protect the rest of us.

***Myth*: Internet companies have no incentive to moderate content.**

Reality: In 1996, Congress passed 47 U.S.C. 230, which says Internet services generally aren't liable for third-party content. Due to this legal protection, critics often assume Internet services won't invest in content moderation; and some companies have stoked that perception by publicly positioning themselves as "neutral" technology platforms. Yet, virtually every Internet service moderates content, and major services like Facebook and YouTube employ many thousands of content reviewers. Why? The services have their own reputation to manage, and they care about how content can affect their users (e.g., Pinterest combats content that promotes eating disorders). Furthermore, advertisers won't let their ads appear on bad content, which provides additional financial incentives to moderate.

***Myth*: Content moderation, if done right, will make everyone happy.**

Reality: By definition, content moderation is a zero-sum game. Someone gets their desired outcome, and someone else doesn't—and those folks won't be happy with the result.

***Myth*: There is a one-size-fits-all approach to content moderation.**

Reality: Internet services cater to diverse audiences that have different moderation needs. For example, an online crowdsourced encyclopedia like Wikipedia, an open-source software repository like GitHub, and a payment service for content publishers like Patreon all solve different problems for their communities. These services shouldn't have identical content moderation rules.

Reality: Google and Facebook have enough money to handle virtually any requirement imposed by regulators. Startup enterprises do not. Increased content moderation burdens are more likely to block new entrants than to punish Google and Facebook.

***Myth*: Poor content moderation causes anti-social behavior.**

Reality: Poorly executed content moderation can accelerate bad behavior, but often the Internet simply mirrors existing anti-social behavior or tendencies. Better content moderation can't fix problems that are endemic in the human condition.

Regulators are right to identify content moderation as a critically important topic. However, until regulators overcome these myths, regulatory interventions will cause more problems than they solve.