Do social media companies have a social responsibility to self-regulate?

Ernesto Dal Bo, Jonathan Weigel, Guo Xu

The context

The growth of social media has reshaped human interactions and the dissemination of information. Social media has been praised for facilitating connections across distance and for tailoring news and serving customized content to users. Today, over 4.25 billion individuals use social media.[1]

The problem

The rise of social media has also provoked critique. Facebook in particular has come under pressure to increase moderation of its content.[2] A series of public allegations, most prominently led by former employee and whistleblower Frances Haugen, have highlighted the role of the social media platform to spread misinformation, amplify toxic content, and negatively affect mental health, especially among adolescents. As Frances Haugen puts it in her testimony, "the company [Meta] knows how to make Facebook and Instagram safer but won't make the necessary changes because they have put their astronomical profits before people."[3] Similar concerns exist about other social media platforms. Twitter, for instance, worried that Donald Trump's use of the platform after the 2020 presidential election created risks of inciting violence and banned his account altogether.

The discussion

Proponents of more stringent content moderation have pointed to the damage inflicted by exposure to toxic content—such as hate, harassment, and bullying—as well as the social comparisons that are the bread and butter of most social media algorithms. Indeed, there is growing evidence on the link between social media use and mental health. One study of mental health on college campuses found that students at universities where Facebook first entered reported higher rates of depression than students at universities where Facebook entered later.[4] Similarly, a recent experiment found that restricting access to social media increased self-reported well-being.[5] Critics also point out that algorithms are designed to entice users and capture their attention. Since attention is a key currency for advertisers, a business model based on capturing human attention raises concerns that social media platforms may become addictive by design.[6] Finally, although Facebook and other social media companies have improved their content moderation policies in recent years, critics argue that implementation of these policies remains highly imperfect. While high-profile accounts are carefully scrutinized by oversight boards, the sheer volume of interaction on the main platforms makes it difficult to monitor all accounts, even with the help of AI.[7] Age limits in accessing social media platforms are also rarely enforced.

Opponents of more stringent content moderation point out that social media platforms, like Facebook, are not liable for content posted by third parties. Internet platforms are protected by Section 230 of the Communications Decency Act, absolving providers of the need to monitor and approve all content posted on

---

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] Although Facebook has made few public statements about the extent of exact amount of content moderation on the platform, Facebook Files leaked a 3-5% rate: https://www.promarket.org/2022/11/10/the-economics-of-content-moderation-on-social-media/

[3] https://www.forbes.com/sites/abrambrown/2021/10/05/facebook-will-likely-resume-work-on-instagram-for-kids-whistleblower-tells-congress/?sh=75d1f0e64cda

[4] Braghieri, Levy, Makarin (2022): "Social Media and Mental Health", American Economic Review, 112(11): 3660-3693.

[5] Alcott, Braghieri, Eichmeyer, Gentzkow (2020): "The Welfare Effects of Social Media", American Economic Review, 110(3): 629-676.

[6] Allcott, Hunt, Matthew Gentzkow, and Lena Song. "Digital addiction." *American Economic Review* 112, no. 7 (2022): 2424-63.

[7] https://www.wired.com/story/facebook-and-the-folly-of-self-regulation/

their platform. As the Electronic Frontier Foundation points out, Section 230 has enabled the rapid growth of user-generated content and the social media industry in the United States. In countries without similar protections, such as Canada and Australia, there is less user speech, particularly on controversial topics. In non-democracies, governments often directly censor speech on the internet. "Without Section 230, they write, "The next great websites and apps won't even get started, because they'll face overwhelming legal risk to host users' speech."[8] Civil rights organizations such as the American Civil Liberties Union (ACLU) likewise emphasize the risk of arbitrarily enforced content moderation to limit access to information and free speech.[9] Some commentators go further in arguing that exposure to different views, even upsetting and offensive views, can better prepare young people for life in a divided world.[10]

The choice

Social media companies must choose from a continuum of content moderation policies. Most social media platforms engage in some form of content moderation. A platform that opens the door for scams, extreme violence, and criminal activities would not only bring public condemnation; it would also alienate users and advertisers, ultimately adversely affecting the platform provider's bottom line.

Differences in content moderation policies have, in recent years, also led to product differentiation. The ban of Donald Trump's accounts on Facebook and Twitter, for example, gave rise to the rise of Parler and Truth Social, platforms that market themselves as free-speech-focused alternatives to previous platforms. Similarly, the acquisition of Twitter by Elon Musk gave Mastodon, a decentralized platform, a significant boost in market share.

While the moderation of illegal content—scams, extreme violence, and other criminal activities—is less ambiguous, more controversy surrounds content that is "lawful but awful" – speech that, while legally permissible, may be considered harmful, offensive, or morally objectionable. Examples include hate speech, disinformation, conspiracy theories, and extreme political views. This type of often incendiary content is also known for provoking and stimulating user engagement.[11] Algorithms also often prioritize engagement, meaning that more controversial or provocative content may receive greater visibility. Such user engagement directly translates into advertising revenue. Restricting the amplification of "lawful but awful" speech can thus directly affect a company's profitability. This self-regulation beyond what is required by legal standards can create a direct tension between business interests and a content moderation policy that is socially responsible.

Should social media companies like Facebook further regulate "lawful but awful" content – content that is harmful but legal – even if it comes at the cost of user engagement and thus advertising revenue?

[8] https://www.eff.org/deeplinks/2023/03/bad-content-moderation-bad-and-government-interference-can-make-it-even-worse

[9] https://www.aclu.org/press-releases/aclu-and-partners-warn-supreme-court-about-dangers-suppressing-free-speech-online

[10] https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/

[11] Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment." *Available at SSRN* (2022). See also: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308