The Modern Moderator’s Dilemma: Incremental Improvements to Address Unique Harms of Social Media

By Jeffrey Westling

Introduction

Humans have long explored ways to communicate more efficiently across distance. The Inca employed a relay system of runners along a 18,600 miles of paved roads, delivering messages across some of the most rugged terrain in the world and the Pony Express used 80 riders to carry mail from the Midwest to the California during the mid-1800s.[1]

Now, broadband Internet access has all but eliminated barriers to communication short of real-time physical connection. Social media apps allow users to post content about their lives or interests. Focused message boards like Reddit facilitate the development of communities about specific interest areas for people to congregate in one virtual location. Video conferencing services allow work to take place from home, though not without a few hiccups like a cameo from a pet or a child.[2] At a core level, the Internet breaks down barriers to connection, opening up the free flow of information at a scale those pre-1900 could only dream of.

In this idealized view of the Internet, the benefits stand front and center. However, breaking down the barriers of communication comes with some challenges and harmful effects. Disinformation, hate speech and obscene material can spread at a breakneck pace with few speed bumps to slow the way.[3] While regulators may seek to address these harms, narrowly targeted intervention is often difficult to achieve without diminishing or outright eliminating the benefits.[4] Platforms have invested significant resources in content moderation to stem the spread of undesirable content and users actively engage with or report content they disapprove of, which provides information that counters the speech, but there is still more work to be done.[5]

However, in an effort to stem problems like disinformation or hate speech many vocal leaders place the onus entirely on the companies to fix the problem because companies directly

[1] “Inca Roads and Chasquis,” *Discover Peru*, last visited May 14, 2021. http://www.discover-peru.org/inca-roads-chasqui; “Pony Express Route,” *National Geographic*, last visited May 14, 2021. https://www.nationalgeographic.org/maps/pony-express.

[2] David Moye, “Scottish Lawmaker Goes Viral After His Cat Photobombs Zoom Meeting,” *Huffington Post*, July 15, 2020. https://www.huffpost.com/entry/cat-photobombs-zoom-meeting-scotland_n_5f0f4de2c5b65426947a5189.

[3] Chris Meserole, “How misinformation spreads to social media – And what to do about it,” *Brookings*, May 9, 2018). https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it.

[4] “Section 230,” *R Street Institute*, March 2021. https://www.rstreet.org/wp-content/uploads/2021/03/explainer22-1.pdf.

[5] Jeffrey Westling, “Are Deep Fakes a Shallow Concern? A Critical Analysis of the Likely Societal Response to Deep Fakes,” *TPRC47*, 2019, p. 20. https://bit.ly/3wcdVqp.

encounter the issues, and thus are considered to be in the best position to stem the spread.[6] These expectations, paired with the profit that platforms derive from additional engagement positions companies squarely in the crosshairs of regulators.[7]

Misunderstanding of the Problem

Recently, the House Energy & Commerce Committee held a hearing exploring the spread of misinformation on social media platforms.[8] Unfortunately, rather than using the hearing as an opportunity to explore these complex questions, Congress used it to make political points and shift the blame to the CEOs of Twitter, Google and Facebook.[9]

The core of calls to regulate social media because of current content moderation decisions primarily stem from a misunderstanding of the specific challenges that social media presents. Contrary to what some critics may believe, social media does not present an entirely new paradigm. Instead, it breaks down the barriers of communication that may otherwise slow or prevent misinformation from spreading. In other words, social media magnifies our existing flaws and problems. Shifting the entire blame of the harms of misinformation and hate speech to social media conflates separate problems that content moderators face.

The Unique Challenges of Social Media

While social media is not an entirely new paradigm, it does nevertheless present new challenges and regulators should explore ways to work with industry and civil society to limit the harms derived from these challenges. While not all encompassing, this section details a few of the specific issues that apply to information spreading on social media, distinct from general challenges associated with bad content.

The first and most obvious challenge unique to social media is the vast amount of information that is available to users. Anyone can post to most social media platforms with no authorization other than making an account. This means that moderators must have a wide-ranging policy to

[6] Chris Mills Rodrigo, "Lawmakers vent frustration in first hearing with tech CEOs since Capitol riot," *The Hill*, March 25, 2021. https://thehill.com/policy/technology/545017-lawmakers-grill-big-tech-ceos-on-misinformation-leading-to-capitol-riot.

[7] Tim Wu, "Blind Spot: The Attention Economy and the Law," Antitrust Law Journal 82:771, 2017. https://scholarship.law.columbia.edu/cgi/viewcontent.cgi?article=3030&context=faculty_scholarship.

[8] Before the Subcommittee on Communications and Technology and the Subcommittee on Consumer Protection and Commerce of the Committee on Energy and Commerce, Hearing on "Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation," March 25, 2021. https://energycommerce.house.gov/committee-activity/hearings/hearing-on-disinformation-nation-social-medias-role-in-promoting.

[9] Shoshana Weissmann & Canyon Brimhall, "The Misinformation Congress Peddled at a Hearing to Combat Misinformation with Technology CEOs," *R Street Institute*, April 12, 2021. https://www.rstreet.org/2021/04/12/the-misinformation-congress-peddled-at-a-hearing-to-combat-misinformation-with-technology-ceos.

give them flexibility to make judgment calls and remove content that the platform would otherwise disapprove of on the service.

Further, psychological factors with information consumption online can present difficulty for moderation. For example, false information spreads quickly online at least in part because of how we attribute sources.[10] For instance, when a friend shares information from a separate source, the user who sees the information associates the information with the friend who shared it, not the original source.[11] This presents a novel problem, because people tend to trust those in their social networks, in part because evolutionarily people once benefited from relying on their social groups to identify potential dangers.[12]

At the same time, inflammatory or obscene material often drives emotional responses, increasing the chances that an individual will engage with that content.[13] Unfortunately, this spreads inflammatory content more rapidly and can make it challenging for "good speech" to stem its spread. This mechanism is similar to the speed at which the harmful content spreads, as a meme or viral story can diffuse over social networks at a breakneck pace that we traditionally have not seen in the past.[14]

The specific challenges that are unique to social media tend to stem from the speed and ease at which information is shared rather than any unique characteristics of the origination of that information. If poor information is part of the issue, online platforms will not be able to effect change without simultaneous work in other sectors, like education.

Deeply Rooted Problems Apart from Social Media

Given the unique challenges of the movement of harmful content over social media channels, it is understandable why calls for regulation seemingly point to social media as the problem and the solution. But this simplifies the complexity of the issue, which also encompasses a larger degradation of societal norms regarding truth and accuracy that positions content moderators in an untenable position.

Leading up to the presidential election, then-President Trump made bold claims, largely unsupported by any evidence, that democrats across the country actively engaged in efforts to

[10] "'Who shared it?' How American's decide what news to trust on social media,"
American Press Institute, Mar. 20, 2017. https://www.americanpressinstitute.org/publications/reports/survey-research/trust-social-media.

[11] Ibid.

[12] Jeffrey Westling, "Deception & Trust: A Deep Look at Deep Fakes," *Techdirt*, Feb. 28, 2019. https://www.techdirt.com/articles/20190215/10563541601/deception-trust-deep-look-deep-fakes.shtml.

[13] Westling, *supra* n. 6, p. 17.

[14] Testimony of R Street Institute, House Subcommittee on Communications and Technology and the Subcommittee on Consumer Protection and Commerce of the Energy and Commerce Committee, "Testimony for Fostering A Healthier Internet to Protect Consumers Section 230 of the Communications Decency Act," Oct. 16, 2019. https://www.rstreet.org/2019/10/16/testimony-for-fostering-a-healthier-internet-to-protect-consumers-section-230-of-the-communications-decency-act.

steal the election from Republicans and him in particular.[15] To many, these claims went beyond acceptable behavior, but his disregard for fact and his populist sentiments were the logical extension of many political talking points emerging, at times, from both parties.[16] Unsurprisingly, President Trump continued his claims after the election seemingly concluded, pressuring election officials across the country.

On January 6th, President Trump held a rally inviting supporters to the Capitol and urging Republican lawmakers not to certify the election results and to “stop the steal.”[17] After working the crowd up, the attendants stormed the Capitol building itself. In an absolute breakdown in a belief in the political process, the facts and truth of the election fell to interests in winning.

Many were quick to blame social media for the rise of Trump and the spread of the lies that he perpetuated.[18] Theoretically, Facebook or Twitter had ample opportunity to ban and limit President Trump on their services, but doing so would cut at the core values of free speech and open dialogue between the leader of the country and its citizens. The problem was not that the President could directly spread these lies to the citizenry, but rather the willingness to do so in the first place.

The fall of the Roman Republic provides startling comparisons to our current political climate. At the outset of the Republic, citizens felt the need to push back on monarchs, sole authority figures with absolute power. While disagreements and tensions always existed, there was at least a nominal principle that decisions should be made via consensus. Significant efforts were made to achieve this consensus and process mattered. Unfortunately, over time, these societal norms began to degrade and were replaced by an “ends justifies the means” mindset in which the only objective was to win. Unsurprisingly, two brothers pushed the norms beyond their limits, opportunistically establishing policies in ways that breached the boundaries of acceptable behavior.[19] As author Edward J. Watts put it, “[t]he quest for consensus that had made Rome’s republic so stable in previous centuries was quickly replaced by a winner-takes-all attitude toward political disputes.”[20] This ultimately led to the first act of lethal Roman political violence in three centuries.

Obviously, social media played no part in the downfall of Rome, and yet the similarities between 130 B.C. and the modern day are clear. Our system of political discourse seems sturdy

[15] “US election 2020: Fact-checking Trump team’s main fraud claims,” *BBC News*, Nov. 23, 2020. https://www.bbc.com/news/election-us-2020-55016029.

[16] Joe Kane, “Does the Internet Still Exist!?!?! Fact-checking net neutrality doomsday predictions,” *R Street Institute*, June 11, 2018. https://www.rstreet.org/2018/06/11/fact-checking-net-neutrality-doomsday-predictions.

[17] Julia Jacobo, “This is what Trump told supporters before many stormed Capitol Hill,” *ABC News*, Jan. 7, 2021. https://abcnews.go.com/Politics/trump-told-supporters-stormed-capitol-hill/story?id=75110558.

[18] Taylor Hatmaker, “Top tech CEOs will testify about social media’s role in the Capitol attack this week,” *Techcrunch* (Mar. 23, 2021). https://techcrunch.com/2021/03/23/tech-hearing-capitol-attack-facebook-twitter-google.

[19] Edward J. Watts, *Mortal Republic: How Rome Fell into Tyranny* (Basic Books, 2018).

[20] Ibid.

from the outside, but as the norms we rely on continually degrade, the foundations continue to weaken.

It is undoubtedly true that social media can be used as a tool to degrade these norms and institutions further, as information can be shared rapidly with relatively limited checks. Yet these issues are much more deeply rooted than a focus on technology would suggest. In the case of President Trump lying about an election, moderators must balance significant interests. On the one hand, spreading the lies and fear led to a direct attack on our democracy. On the other, President Trump was the people's elected leader and limiting his ability to reach the American people would have gone against core American values. Many individuals want to engage with political leaders, and limiting or removing their access to constituents is itself a harm to the political process. Theoretically, our elected officials are who we should turn to for leadership and guidance of times of crisis, but they themselves often cause the very problems we want social media to fix.

Further, despite theories that social media leads to information bubbles (i.e. individuals only associate with like-minded users and therefore only interact with self-affirming content), evidence suggests that it is not social media that drives these tendencies but in-fact traditional media outlets.[21] For example, while social media allows individuals to share content, news stories generally do not spread widely until a major news outlet like Fox News picks up the story and publishes it.[22] While social media moderators can and often do limit the distribution of stories from traditional news outlets when the story violates the terms and conditions, these determinations often raise difficult questions about balancing free, open reporting with the desire to stem the spread of harmful material on the service.

Calling on platforms and moderators to simply solve the problem ignores the fact that the disregard for societal norms about truth and process plague the entire information ecosystems, with many using exaggeration and lies to justify desired outcomes. Worse, with the threat of legislation looming in the background, eager regulators desiring to dictate content decisions can exert influence on the process to achieve outcomes for their own political success.[23]

Use of Multistakeholder Process to Develop Best Practices or Code of Ethics

[21] Yochai Benkler, et al., *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press 2018). https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624.

[22] Ibid.

[23] Charles Duan & Jeffrey Westling, "Will Trump's Executive Order Harm Online Speech? It Already Did," *Lawfare*, June 1, 2020. https://www.lawfareblog.com/will-trumps-executive-order-harm-online-speech-it-already-did.

While reforms to moderation cannot address these deeply rooted issues, there are still ways to mitigate the harms. We may not be able to solve every problem, but we can work to make incremental improvements that limit harm.

In 2019, Mike Godwin proposed that "tech companies develop an industry-wide code of ethics that they can unite behind in implementing their censorship and privacy policies – as well as any other information policies that may affect individuals."[24] Building on Yale law professor Jack Balkin's work on information fiduciaries, Godwin suggests that establishing a duty to the users can guide decision-making without forcefully dictating outcomes.[25]

Importantly, this would not result in a rigid system in which all content would be treated in the same way on every platform, and competing values between services could actually serve as a competitive advantage. For example, during the recent Facebook oversight board (itself a pseudo multistakeholder system for internal auditing) open comment period, R Street proposed to the board Framework Factors that should guide decision-making.[26] These include things like truth or falsity, harmfulness, imminence, incitement and the appropriateness of the sanctions. Like the copyright factors for fair use, no one factor guides the decision-making, and competing values can affect outcomes. Even if content is a blatant lie, it may not be harmful or incite violence. Each individual case will vary, but defining these factors can help guide the decision-making process to better achieve a desirable outcome.

Ideally, enacting a multi-stakeholder process to narrowly explore specific issues and responses will provide more guidance for content moderators. With this in mind, not every social media service would navigate these issues in the same way. Such an approach avoids overreaching legislation designed to direct content moderation decisions towards specific outcomes or simply ignores the practical effect that such legislation would have.[27]

Conclusion

The rise of social media has brought significant attention to our information ecosystem. While there are major, deeply rooted issues that society must address, this is not the case of a new technology coming in and turning the existing paradigm on its head. We cannot regulate ourselves out of the issues before us, but there are ways to limit the harms due to the unique challenges that the new technology presents. As long as we ignore these challenges and simply blame social media, the problems will only worsen.

[24] Mike Godwin, "A Facebook request: Write a code of tech ethics," *Los Angeles Times*, April 30, 2019. https://www.latimes.com/opinion/op-ed/la-oe-godwin-technology-ethics-20190430-story.html.

[25] Ibid.

[26] Chris Riley & Paul Rosenzweig, "R Street on Trump Ban to Oversight Board: 'Facebook is Justified,'" *R Street Institute*, Feb. 7, 2021. https://www.rstreet.org/2021/02/07/case-no-2021-001-fb-fbr-facebook-oversight-board.

[27] "Section 230," *R Street Institute*, March 2021. https://www.rstreet.org/wp-content/uploads/2021/03/explainer22-1.pdf.