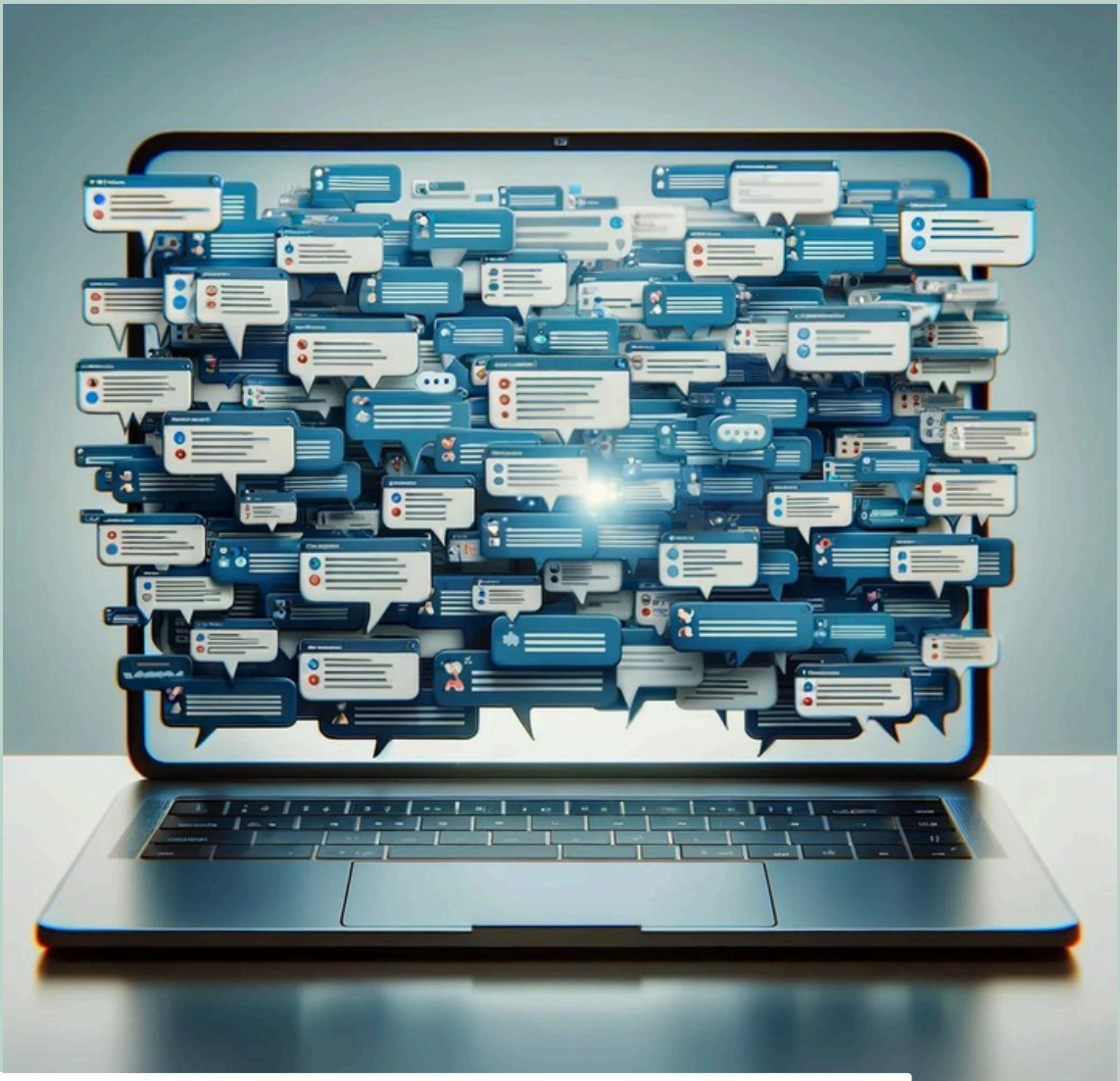


ENHANCING CONTENT MODERATION STRATEGIES:

EVOLVING SOCIAL MEDIA'S ROLE IN TIMES OF GLOBAL CRISIS AND CIVIL UNREST

KARSYN LEMMONS
COMMUNICATIONS, CULTURE, AND TECHNOLOGY
GEORGETOWN UNIVERSITY

MAY 2024

“The media's the most powerful entity on earth. They have the power to make the innocent guilty and to make the guilty innocent, and that's power. Because they control the minds of the masses.”

- Malcolm X

ABSTRACT:

This paper explores the crucial role of content moderation on social media during global crises and civil unrest. It assesses how platforms like Facebook and Twitter influence public behavior either positively by disseminating critical information or negatively through the spread of misinformation. The paper reflects on several historical instances in which social media was either used as a tool for positive development or destruction, illustrating the profound impact of social media in shaping public discourse and actions during emergencies.

Furthermore. the analysis underscores the necessity for stringent moderation policies that promote accurate information and prevent the spread of harmful content. It calls for enhanced transparency and collaboration between social media companies, fact-checkers, and regulatory bodies to address the challenges posed by misinformation effectively.

In today's digital era, the rapid dissemination of information during crises profoundly affects public safety and sentiment. Strong content moderation policies are vital to manage misinformation and disinformation, prevent panic, and protect against the manipulation of public opinion. These policies help maintain order and trust by promoting verified information and debunking rumors, especially during emergencies such as natural disasters, health crises, or political unrest. Effective content moderation supports credible digital platforms, aids in emergency responses by eliminating false information, and preserves social order by preventing content that only serves to incite violence or fear.

Social media's role in reshaping communication and its critical function during crises highlight the need for stringent moderation. Events like the Arab Spring, the Rohingya crisis, the COVID-19 pandemic, and the U.S. Capitol riots illustrate social media's dual potential to either fuel or mitigate conflicts through the spread of misinformation. These instances emphasize the urgency of robust content moderation. Because of the widespread nature of these platforms, social media companies have a significant duty to implement stricter policies, enhance transparency, and collaborate with fact-checkers and authorities to mitigate the risks of unchecked content during critical times. To quote an excerpt from *Reality and Other Stories* by researcher and author John Lancaster, “If you imagined some force or agency in the world that was leading us toward doom and destruction, towards the dark, and then you imagined what kind of tools and technologies it would use, you’d come up with something like social media” (Lancaster, 2022). Now is the time for social media platforms to implement enhanced policies for content moderation via real-time monitoring, transparent reporting, standardized labeling, and numerous other strategies before devolving into a tool for destruction.

Background and Current Landscape of Social Media Platforms

First, it's crucial to recognize social media's significant role in communication, especially during crises, and to review the policies and regulatory frameworks that already oversee these platforms. Historically, platforms like Facebook and Twitter have played key roles in events like the Arab Spring when they enabled activists to organize and share information,

bypassing state media. However, these platforms were also used for spreading misinformation and propaganda, sometimes endangering lives. During the 2020 U.S. elections, widespread misinformation led to public distrust and the Capitol riots, underscoring social media's significant impact on democracy and safety.

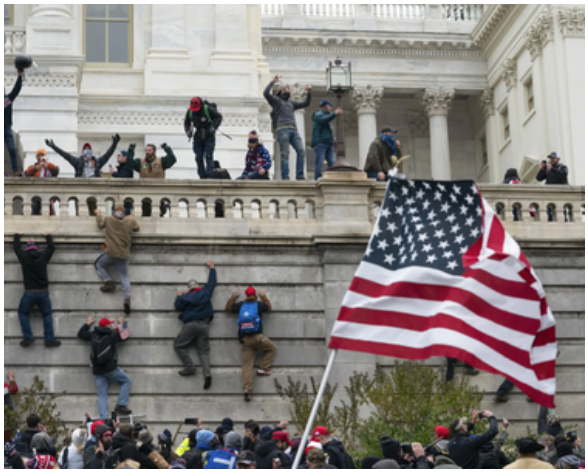
'NOW IS THE TIME FOR SOCIAL MEDIA PLATFORMS TO IMPLEMENT ENHANCED POLICIES FOR CONTENT MODERATION BEFORE DEVOLVING INTO A TOOL FOR DESTRUCTION.'

In recent years, social media companies, including Facebook, Twitter, and YouTube, have placed more emphasis on creating specialized content moderation policies to address issues from hate speech to disinformation, using both algorithms and human oversight. Yet effectiveness varies, particularly during pivotal events like elections or health crises when fast-spreading misinformation can have severe impacts. Under current policies, Facebook flags misleading content using AI and human fact-checkers, Twitter labels and restricts the sharing of such content, and YouTube reduces visibility of misinformation while partnering with health organizations for accurate public health information.

Regulations also differ globally. The EU's Digital Services Act pushes for stricter content accountability, while the U.S. relies on Section 230 of the Communications Decency Act, giving platforms considerable immunity. Conversely, China enforces strict censorship and control over social media. This regulatory diversity reflects varying global values and priorities regarding free speech and information control. **As crises evolve, social media strategies must also adapt, balancing the need for free expression with harm prevention.** And since the idea of completely overhauling content moderation on private social media platforms may seem overwhelmingly challenging, it is crucial to start first with the most pressing issues. Specifically, it's vital to address situations where people rely on these platforms for essential information but encounter a flood of misinformation instead.

Challenges and Issues in Addressing Misinformation/Disinformation

Defining and identifying misinformation and disinformation poses significant challenges for social media platforms, particularly due to the fine line between incorrect information, intentionally deceptive content, and legitimate disagreement. Complications arise from contextual subtleties, cultural differences, and the rapid evolution of misinformation tactics like deep fakes, which outpace current detection technologies.



Rioters scale a wall at the U.S. Capitol on Jan. 6, 2021, in Washington. (AP Photo/Jose Luis Magana)

Misinformation often mixes truth with misleading elements, requiring moderators and algorithms to have extensive background knowledge. Additionally, cultural variances complicate the establishment of consistent global standards for what qualifies as misinformation. Increased content moderation also brings about significant free speech concerns. There exists a critical balance to maintain between mitigating misinformation while preserving free expression – with strict moderation running the risk of potentially silencing marginalized voices and stifling debate. Any signs of biased enforcement could result in accusations of political censorship.

An additional issue is lack of knowledge regarding the exact quantity of misinformation on social platforms and the lack of companies revealing that data to researchers. In the book Regulating Digital Industries by Mark MacCarthy, he references Ethan Zuckerman's statement that "the prevalence of false information available on a platform is useless unless researchers also know how much material is available on the platform and how it is distributed. But social media companies affirmatively prevent outsiders from accessing this information." MacCarthy then highlights that "a regulatory requirement to disgorge this data to qualified researchers would begin to give regulators and policymakers the knowledge they need to begin to understand the scope of the problem and to craft effective remedies," thus providing the first step in breaking down the problem of the spread of misinformation/disinformation online during crisis times.

The last concern to discuss in this section is the fear of jawboning, defined by the Knight First Amendment Institute at Columbia University as informal government efforts to persuade, cajole, or strong-arm private platforms to change their content-moderation practices. Government enforced moderation mandates are prohibited by the First Amendment, however, "by working through intermediaries, government can suppress speech quickly, without broad support, and potentially without alerting anyone of its involvement" (Grossman & Shapiro, 2023), effectively skirting around the unconstitutionality of the situation.

According to a 2023 Pew Research Study, 30% of Americans report they regularly get their news on Facebook, and slightly less (26%) stated they regularly turn to YouTube as a news source.

Source

'THERE EXISTS A CRITICAL BALANCE TO MAINTAIN BETWEEN MITIGATING MISINFORMATION WHILE PRESERVING FREE EXPRESSION.'

Case Studies Concerning Successful and Failed Outcomes of Content Moderation

During the Christchurch mosque shootings in March of 2019, the attacker liver-streamed the event on Facebook Live. Although initially the video was only viewed by a handful, it was widely replayed before its removal from the platform. In response to scrutiny over Facebook's slow reaction time and already existing preventive measures, the platform implemented stricter live-streaming policies, especially for users with prior infractions. The rapid efforts to remove and limit the video's spread across platforms provides an example of effective crisis moderation, likely lessening further harm by curbing potential glorification of the attack and protecting users from graphic content. It is important to note, however, that in the aftermath, Facebook resisted efforts to require them to prescreen live videos.

On the other end of the spectrum is an example of failed content moderation in a time of civil unrest leading to extreme harm. In 2017, Facebook was widely used in Myanmar to disseminate hate speech and misinformation against Rohingya Muslims. The platform was used extensively by nationalist groups as well as some government officials to spread hate speech, incite violence, and organize rallies. Misinformation campaigns were rampant, painting the Rohingya as illegal immigrants and terrorists, which inflamed public sentiment and increased the justification for violent acts. Criticized for its slow response and lack of Burmese-speaking moderators, Facebook struggled to control the harmful content. The ineffective moderation exacerbated the violence and ethnic cleansing, emphasizing the critical impact of inadequate content management during crises.

During the COVID-19 pandemic, Facebook, Twitter, and YouTube each tackled the spread of misinformation about the virus. Facebook used AI for detection, guided users to sources like WHO and CDC, and worked with fact-checkers across the globe. Twitter used labels and warnings for misleading tweets and removed harmful content. YouTube adjusted its algorithms to reduce misinformation spread and promoted credible sources. While the effectiveness of these measures is varied, they likely helped prevent some misinformation from exacerbating public health issues.

These case studies can offer important lessons on the challenges and essentials of content moderation in today's digital ecosystem, highlighting the need for platform-specific content moderation strategies to effectively control misinformation and fit each platform's specific characteristics and audience. Effective content moderation can reduce harm significantly, while failing to moderate content properly can worsen crises and lead to real-world violence and the spread of false information.

Analysis of Current Policy

Before proposing new social media content moderation policies, it is imperative to analyze the policies already in place, including the impact these policies have on stakeholders as well as the legal and ethical implications.

The effectiveness of content moderation policies during crises varies widely across social media platforms. While many platforms have advanced systems for detecting and removing harmful content, their approaches often lack consistency and tend to be more reactive than proactive. Additionally, the accuracy of moderation tools, which often rely on AI, can misidentify content like satire as misinformation while missing other nuanced false claims.

Governments may benefit from assisting in effective moderation during crises to maintain public order and protect against misinformation, but risk rising tensions from insufficient moderation or perceived over-moderation as censorship. For users, strong moderation improves platform reliability and aids in safe, informed decision-making, though aggressive policies may raise privacy and free speech concerns.

RECENT CONTENT MODERATION TROUBLE AMONG MAJOR SOCIAL MEDIA COMPANIES

Meta announced that it will 'stop recommending political content from accounts that users don't follow' on Instagram and Threads, Similar political content limiting actions were applied to Facebook in 2021, following the January 6th Capitol riots. Find the official statement here.

X has been at the center of significant controversy in recent years following its acquisition by Elon Musk and the ensuing dismissal of many of the platform's content moderation staff. Learn more about the organizations policies at 'The X Rules.'

YouTube has consistently struggled with content moderation. Last year, Matt Halprin, Leader of Youtube's Trust and Safety team, addressed the challenges related to bias, enforcement consistency, and adapting community guidelines in a video with Creator Insider.

Non-users are also impacted by misinformation's effects as it spreads into mainstream media and public discourse. Globally, a unified strategy to combat misinformation, particularly during worldwide crises like pandemics, is beneficial; however, inconsistent policies can disrupt coordination and have varied regional impacts.

In legal terms, content moderation in the U.S. must balance freedom of speech with harm prevention, often leading to disputes over perceived free speech restrictions. Under Section 230 of the Communications Decency Act, platforms are not typically liable for illegal user-generated content, though there's debate over whether these protections need updating or should remain as they are.

In a recent piece commenting on the SCOTUS decision that Google, Facebook, and Twitter should not be held responsible for terroristic posts by the Islamic State, the Stanford Law School lecturer Daphne Keller called it the court's verdict an "everybody calm down moment", referencing those who call for Section 230's revision. According to her appraisal, these platforms "do not rise to the level of 'aiding and abetting' acts of terrorist" and instead run the risk of overbroad liability and limiting free speech.

Although Keller's reasoning may be valid that platforms like Facebook and Twitter are not directly involved in terrorism simply by hosting terrorist content, it's important to note that both Meta and Twitter have policies against such content. They have a duty to their users to more diligently shield them from exposure to these harmful materials and Section 230 should be adjusted to support that.

Furthermore, moderation requires transparency and accountability in decision-making, fairness in policy application without political or ideological bias, and a consideration of the platform's impact on public discourse, striving to maintain a healthy democratic exchange while preventing harm.

SOCIAL MEDIA CONTENT MODERATION POLICY PROPOSALS

These proposed regulations and implementation strategies create a stronger, more accountable, and transparent content moderation environment on social media platforms, especially during crises. The recommended monitoring and evaluation methods, supervised by an appropriate government agency, will ensure that these policies not only meet legal standards but also adaptively respond to the evolving dynamics of global information dissemination.

PROPOSED POLICIES

1. **Real-Time Monitoring During Crises:** Mandate that platforms establish real-time crisis response teams that can rapidly address emerging threats of misinformation. These teams should include experts in crisis management, public policy, and specific regional knowledge where applicable addressing harmful content before it spreads.
2. **Transparent Reporting:** Require platforms to publish detailed reports on content moderation actions taken during crises, including data on the types of misinformation identified, the sources of such misinformation, and the impact of the moderation efforts.
3. **Standardized Misinformation Labels:** Develop standardized labels that can be universally applied to misinformation content across all platforms, making it easier for users to recognize verified information.
4. **National Cooperation Framework:** Establish a national legal framework that facilitates cooperation in content moderation, ensuring consistent enforcement against global misinformation campaigns.
5. **Strengthened Legal Standards:** Amend laws such as Section 230 of the Communications Decency Act to condition the immunity of platforms on adopting best practices for crisis-related content moderation.

Implementation strategies for enhancing content moderation during crises include forming partnerships with independent fact-checking organizations to improve information verification accuracy. Investing in advanced AI technologies for initial content scans, supplemented by human moderators who provide essential context and cultural understanding, is also crucial. Relying entirely on algorithms or other technologies cannot be a solution at present as "we do not know enough to mandate algorithmic solutions or require specific technical or operational interventions" (MacCarthy, 2023). Although regular training for moderators on the latest misinformation trends and crisis-specific challenges, including cultural sensitivity, can help address global disparities. Conducting crisis simulation exercises can also prepare moderation teams for various scenarios, enhancing their readiness and response effectiveness.

'WHILE FREE SPEECH IS CRUCIAL, IT DOESN'T EXTEND TO SPREADING MISINFORMATION THAT CAN CAUSE REAL-WORLD HARM, THUS PROPOSED REGULATIONS AIM TO SAFEGUARD PUBLIC HEALTH AND SAFETY DURING CRISES WITH CLEAR GUIDELINES ENSURING LEGITIMATE FREE EXPRESSION ISN'T STIFLED.'

Methods for ensuring implementation might include regular audits of moderation actions by independent bodies to check for adherence to regulations and the effectiveness of moderation strategies. Develop clear performance metrics to assess moderation during crises, focusing on response speed, accuracy of misinformation identification, and user satisfaction with information quality. Lastly, utilize data from these evaluations to continuously refine and improve moderation policies and practices, adapting to new misinformation types and evolving societal norms.

Counterarguments and Rebuttals

Critics of increased content moderation raise concerns about infringing on free speech rights, as they believe individuals should freely express diverse opinions, including controversial ones.

Additionally, opponents argue that the proposed regulations, such as real-time monitoring and international cooperation, could be too costly and complex for effective implementation, especially for smaller platforms. Moreover, there are concerns about privacy risks, with advanced AI technologies potentially leading to the misuse of data and surveillance overreach.

In terms of rebuttals, while free speech is crucial, it doesn't extend to spreading misinformation that can cause real-world harm, thus proposed regulations aim to safeguard public health and safety during crises with clear guidelines ensuring legitimate free expression isn't stifled. To mitigate bias and censorship fears, proposed frameworks include transparent practices and independent oversight with regular audits and public reports to maintain fairness across all content.

Addressing cost concerns, phased support for smaller platforms, possibly through shared resources or governmental subsidies to facilitate technology and training adoption, might resolve this issue. Regarding privacy worries, robust data protection is recommended in AI systems for content moderation, to comply with strict privacy laws and implementing top security practices to prevent data misuse.

The above counterarguments and rebuttals tackle key concerns about increased content moderation, but it is important to stress that with well-designed policies and safeguards, it's feasible to improve information integrity on social media during crises while upholding user rights and freedoms.

Conclusion

As social media's impact on society grows, it's increasingly urgent that these platforms are used safely and ethically. Policymakers, social media companies, and civil society groups need to work together to bring about changes that reduce harm and strengthen how we handle misinformation and disinformation. There is a desperate need to act quickly to update our regulatory frameworks for the digital age, focusing on stricter, clearer, and fairer moderation policies that protect human rights and freedom of speech.

Our world today relies on social media for spreading information, especially in urgent situations, but managing content on these platforms still poses a major challenge. **The continual challenge is to find the optimal balance between curbing misinformation and safeguarding free speech, alongside addressing technological challenges.** This is a demanding yet essential task to ensure our public discourse remains free yet safeguarded in this digital era.

About the Author

Karsyn Lemmons is currently pursuing a Master's degree in the Communications, Culture, and Technology program at Georgetown University, where she focuses on strategic intercultural communications for social impact. Before her studies at Georgetown, she earned her Bachelor of Journalism with university honors from The University of Texas at Austin, complemented by a minor in History. Her current academic pursuits encompass intercultural and international crisis communications, as well as developing people-centered outreach and engagement strategies aimed at promoting public good.

Additional References

French, D. (2020, January 24). The Growing Threat to Free Speech Online. TIME. https://time.com/5770755/threat-free-speech-online/

Grossman, A. M., & Shapiro, K. A. (2023, October 3). Shining a Light on Censorship: How Transparency Can Curtail Government Social Media Censorship and More. Cato.org. https://www.cato.org/briefing-paper/shining-light-censorship-how-transparency-can-curtail-government-social-media#conclusion

Keller, D. (2023, May 19). Stanford's Daphne Keller on Scotus decision that Google, Twitter, and Facebook not responsible for Islamic State Deadly Posts. Stanford Law School. https://law.stanford.edu/2023/05/19/stanfords-daphne-keller-on-scotus-decision-that-google-twitter-and-facebook-not-responsible-for-islamic-state-deadly-posts/

LANCHESTER, J. (2022). Reality and other stories. W W NORTON. 75.

MacCarthy, M. (2023). Regulating Digital Industries. Regulation of Social Media Algorithms.

**ChatGPT was implemented for assistance with grammar and sentence structure as well as developing report cover image