COVID-19 CONTENT MODERATION AND PLATFORM ACCOUNTABILITY

Tuomas Heikkilä, Salla-Maaria Laaksonen and Matti Pohjonen

MAPS - Media platforms and social accountability

ECREA postconference Digital media and information disorders, Oct 24, 2022

UNIVERSITY OF HELSINKI

PLATFORMS AND RESPONSIBILITY

YouTube and Facebook allowed another COVID-19 conspiracy theory video to go viral

MIT Technology Review

The coronavirus is the first true social-media "infodemic"

MICROSOFT GOOGLE AMAZON

What Techlash? Virus Could Remake Industry Giants' Image

By Cory Weinberg | March 23, 2020 9:52 AM PDT

Photo: Photo: AP

Fact-Checking

On Facebook, health misinformation is king. And it's a global problem.

THE INTERFACE | GOOGLE | POLICY

How COVID-19 is changing public perception of big tech companies

The backlash against tech giants may not be over — but at the very least it's on pause

Spotify finally responds to Joe Rogan controversy with a plan to label podcasts that discuss COVID-19

The company is also publishing its content rules for the first time

LEGITIMACY AND SOCIAL ACCOUNTABILITY

Legitimacy

General perception of the appropriateness of the actions of an entity [1]

How the exercise of power is morally justified [2]

Why should we do this? [3]

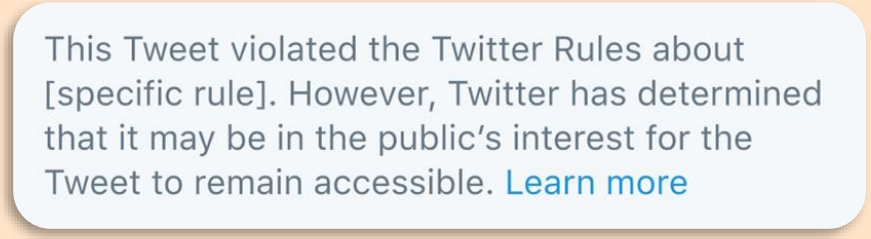
Social accountability

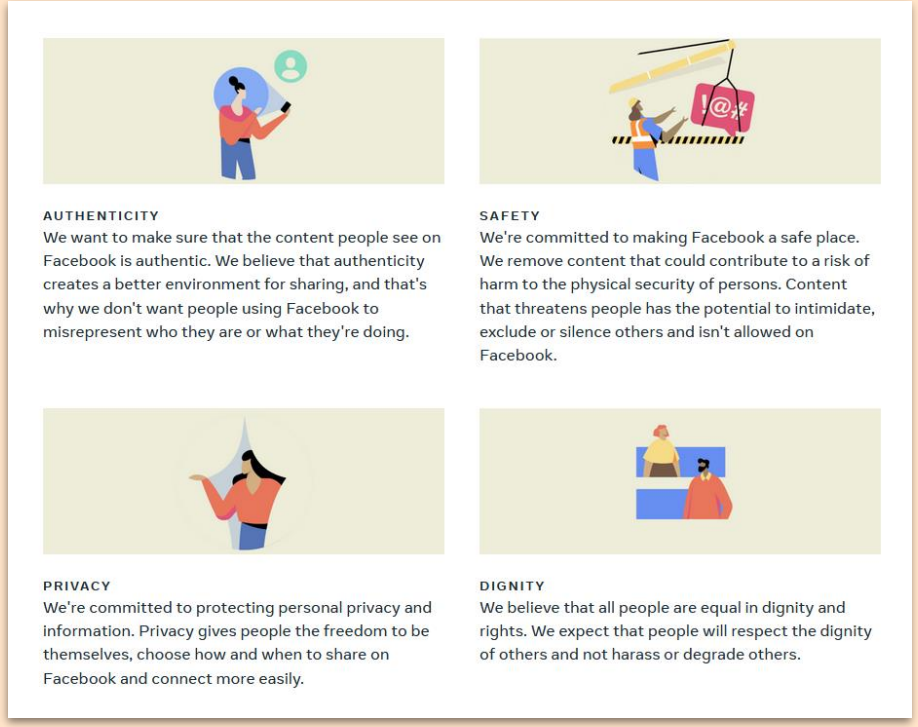
Informal and voluntary accountability between private and public organizations and their societal stakeholders [4]

Self-regulation as platforms' central accountability mechanism [5]

ALL PLATFORMS MODERATE

- **Detection**, **review** and **enforcement** of unacceptable content or behaviour [6]
- Legal liability: CDA Section 230 [7]
- Commercial, at-scale content moderation [8]
- Tradition of discursive performances that legitimize content governance decisions [9]





THE CASE FOR MISINFORMATION REGULATION…

- Long-standing accountability demands for factually accurate platforms
- Ample evidence about the negative effects of (health) misinformation [10]
- Numerous experimentally tested potential interventions [11,12]
- Dominant platforms have demonstrated to posess capabilities, systems and human talent to tackle similar issues (e.g. terrorism) [13]
- Journalistic fact-checking can be incredibly labour-intensive task [14]
- Misinformation routinely outperforms its debunks [15,16]

...AND THE CASE AGAINST IT

- Practical reasons – how?
- Reasons of principle – why? [17,18]
- Lack of regulatory pressure [19]
- Misgivings by most valuable user segments [20]

META'S MISINFORMATION POLICY IN COMMUNITY GUIDELINES, DEC 2021

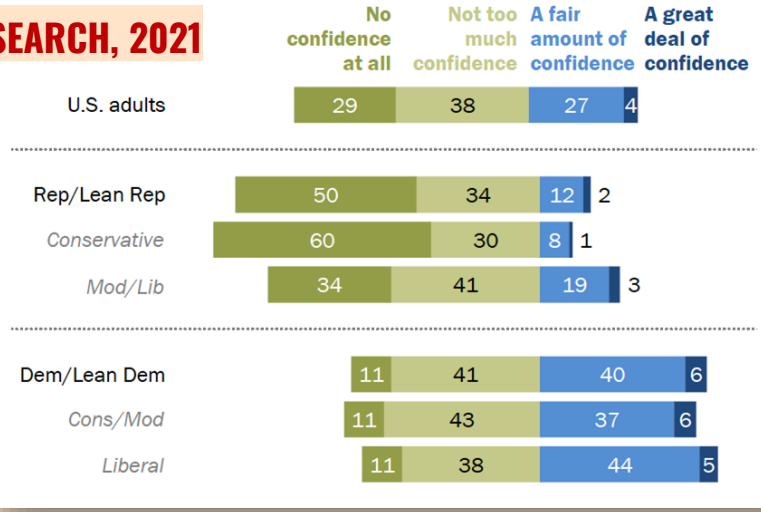
Policy rationale

Misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited. With graphic violence or hate speech, for instance, our policies specify the speech that we prohibit, and even persons who disagree with those policies can follow them. With misinformation, however, we cannot provide such a line. The world is changing constantly, and what is true one minute may not be true the next minute. People also

PEW RESEARCH, 2021

A majority of conservative Republicans have no confidence in social media companies to determine which posts should be labeled as inaccurate

% of U.S. adults who say they have __ in social media companies to determine which posts on their platforms should be labeled as inaccurate or misleading



METHODS AND MATERIALS

MATERIALS

- **Corporate blogs** scraped from Meta (n=1,508), Twitter (n=463) and YouTube (n=2,695)
- **Filtered set:** Posts that mentioned keywords related to Covid-19 (n=125).
- **Analysis:** tracing COVID-19 related corporate actions and discursive strategies to legitimize them.

RQ

What discursive strategies platforms employ in their public statements when presenting their roles, actions, rationales, and successes in correcting and disrupting misleading claims regarding the COVID-19 pandemic?

PRELIMINARY RESULTS

PROTECTION FROM HARM

- Transforming difficult questions about facts and truths to matters of safety and well-being

Our goal is to help messages about the **safety** and efficacy of vaccines reach a broad group of people, while prohibiting ads with misinformation that could **harm public health efforts**. (Meta, 13 Oct 2020)

We are committed to our responsibility to protect the YouTube community, and expanding our fact check information panels is one of the many steps we are taking to raise up authoritative sources (YouTube, 28 April 2020)

We are focused on mitigating misleading information that presents **the biggest potential harm** to people's health and wellbeing. (Twitter, 16 Dec 2020)

MISLEADING MISINFORMATION

- Discovering a responsibility to the people using platforms to access reliable information
- ”Soft” moderation
- Positioning platforms as **protectors of public conversation**; central actors in information dissemination, society and even democracy

Over the past several years, **we've seen more and more people coming to YouTube for news and information**. (...) the outbreak of COVID-19 and its spread around the world has reaffirmed how important it is for viewers to get accurate information during fast-moving events. (YouTube, 28 April 2020)

Once a post is rated false [but not posing imminent harm], we reduce its distribution so fewer people see it, and we show strong warning labels and notifications to people who still come across it, try to share it or already have.” (Meta, 25 March 2020)

In serving the public conversation, our goal is to make it easy to find credible information on Twitter and to limit the spread of potentially harmful and misleading content. (Twitter, 11 May 2020)

GOVERNANCE AMIDST AND DUE TO PANDEMIC

- Health emergency as a critical event that necessitates action
- Attributing **failures, errors and glitches** to exceptional circumstances
- Expanding automated systems
- Accepting a **responsible position** – and also positioning others

Ever since COVID-19 was declared a global public health emergency in January, we've been working to connect people to accurate information from health experts and keep harmful misinformation about COVID-19 from spreading on our apps. (Meta, 16 April 2020)

Today, as the unprecedented COVID-19 situation continues, Google outlined how it's reducing the need for people to come into its offices **while ensuring that its products continue to operate for everyone**. (YouTube 16.3.2020).

We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and **this may result in us making mistakes** (…) **We appreciate your patience as we work to get it right** (Twitter 1.4.2020)

We are all in this together, and we will continue to update you on our progress as we strive to **play our part** to protect the public conversation at this critical time. (Twitter, 1 March 2021)

EXPERT AUTHORIZATION

- National and local health agencies, governments and other experts are regularly invoked as…
- **Authorities:** why should misinformation be tackled?
- **Issue-experts:** what is misinformation?
- **Advisers:** how could the systems be improved to better tackle misinformation?

Our **global expert stakeholder consultations have made it clear that**, that in the context of a health emergency, the harm from certain types or health misinformation does lead to imminent physical harm. That is why we remove this content from our platform. (Meta, 1.12.2020)

In order to identify clear bad content, you need a clear set of facts. **For COVID, we rely on expert consensus from health organizations like the CDC and WHO** to track the science as it develops. (YouTube, 25.8.2021)

We're continuing **to work with external experts and governments** to make sure that we are approaching these issues in the right way and making adjustments if necessary. (META, 18.8.2021)

We are regularly **working with and looking to trusted partners, including public health authorities**, organizations, and governments to inform our approach. (Twitter, 1.4.2020)

TAKE AWAYS

Covid-19 emerges as a critical incident for the platforms to re-establish their societal role

- Pandemic as a means of narrating legitimacy in society
- New role of amplifier of good-quality information
- Question remain about misinformation policies' longevety and future uses

REFERENCES

1. Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. Academy of Management Review, 20(3), 571–610. https://doi.org/10.5465/amr.1995.9508080331
2. Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. International Communication Gazette, 80(4), 385–400. https://doi.org/10.1177/1748048518757142
3. Van Leeuwen, T. (2007). Legitimation in discourse and communication. Discourse & Communication, 1(1), 91–112. https://doi.org/10.1177/1750481307071986
4. Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal, 13(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x
5. Moss, G., & Ford, H. (2020). How accountable are digital platforms? In W. Dutton (Ed.), A Research Agenda for Digital Politics (pp. 97–109). Edward Elgar Publishing. https://doi.org/10.4337/9781789903096.00019
6. Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. Internet Policy Review, 9(4).
7. Napoli, P. M., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. First Monday, 22(5). https://doi.org/10.5210/fm.v22i5.7051
8. Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v
9. Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (Illustrated edition). Yale University Press.
10. Tan, A. S. L., Lee, C., & Chae, J. (2015). Exposure to Health (Mis)Information: Lagged Effects on Young Adults' Health Behaviors and Potential Pathways. Journal of Communication, 65(4), 674–698. https://doi.org/10.1111/jcom.12163
11. Bode, L., & Vraga, E. K. (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. Journal of Communication, 65(4), 619–638. https://doi.org/10.1111/jcom.12166
12. Clayton, K., Blair, S., Busam, J. A., …& Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. Political Behavior, 42(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0
13. Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945. https://doi.org/10.1177/2053951719897945
14. Ananny, M. (2018). The Partnership Press: Lessons for Platform-Publisher Collaborations as Facebook and News Outlets Team to Fight Misinformation [A Tow/Knight Report]. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D85B1JG9
15. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
16. Funke, D. (2019, February 28). On Facebook, health misinformation is king. And it's a global problem. Poynter. https://www.poynter.org/fact-checking/2019/on-facebook-health-misinformation-is-king-and-its-a-global-problem/
17. Ananny, M., & Gillespie, T. (2018). Public Platforms: Beyond the Cycle of Shocks and Exceptions. In A. Shaw & D. T. Scott (Eds.), Interventions: Communication research and practice. Peter Lang.
18. Bossetta, M. (2020). Scandalous Design: How Social Media Platforms' Responses to Scandal Impacts Campaigns and Elections. Social Media + Society, 6(2), 2056305120924777. https://doi.org/10.1177/2056305120924777
19. Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of online content. SocArXiv. https://doi.org/10.31235/osf.io/tgnrj
20. Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. Policy & Internet, 13(3), 433–451. https://doi.org/10.1002/poi3.257

THANK YOU

MAPS - https://blogs.helsinki.fi/mapsproject/