Government regulation can effectively curb social media dangers

August 14 2023

Credit: CC0 Public Domain

Government legislation to flag and moderate dangerous content on social media can be effective in reducing harm, even on fast-paced platforms such as X (formerly Twitter) new research shows.

Social media posts such as those that promote terrorism and hate, dangerous challenges that put teen lives at risk, or those that glamorize suicide, pose a significant threat to society. And this harm spreads exponentially, like an infectious disease.

Dr. Marian-Andrei Rizoiu from the University of Technology Sydney (UTS) Behavioural Data Science Lab and Philipp J. Schneider from École Polytechnique Fédérale de Lausanne harnessed state-of-the-art information spread modeling to analyze the dynamics of content dissemination.

The study, The Effectiveness of Moderating Harmful Online Content, has been published in the journal *PNAS*. It finds that even with a 24-hour turnaround time, government mandated external moderation is likely to be effective in limiting harm.

"Social networks such as Facebook, Instagram and Twitter, now X, don't have much incentive to fight harmful content, as their business model is based on monetizing attention," says Dr. Marian-Andrei Rizoiu from the University of Technology Sydney (UTS) Behavioural Data Science Lab.

"Elon Musk acquired Twitter with the stated goal of preserving free speech for the future. However, alongside free speech, mis- and disinformation spreads and prospers in this unregulated space."

In response, many countries including Australia, are looking to intervene and regulate. The European Council has taken the groundbreaking step of introducing the Digital Services Act and the Digital Markets Act to combat the dissemination of dangerous content.

The EU legislation requires trusted flaggers to identify harmful content, which platforms must then remove within 24 hours. However, critics have suggested the legislation might be ineffective given the speed at

which social media content spreads.

"We've seen examples on Twitter where the sharing of a fake image of an explosion near the Pentagon caused the US share market to dip in a matter of minutes, so there were doubts about whether the new EU regulations would have an impact," says Dr. Rizoiu.

To better understand the relationship between the moderation delay and the likely harm reduction achieved, the researchers examined two key measures: potential harm and content half-life.

Potential harm represents the number of harmful offspring generated by a single post, and content half-life denotes the time required for half of all offspring to be generated.

Prior research has determined the half-life of social media posts on different platforms. X (Twitter) has the fastest half-life at 24 minutes, followed by Facebook at 105 minutes, Instagram 20 hours, LinkedIn 24 hours, and 8.8 days for YouTube.

"A lower half-life means that most harm happens right after the content is posted, and content moderation needs to be performed quickly to be effective. We found the reaction time required for content moderation increases with both the content half-life and potential harm," he said.

The study has implications for policymakers looking to introduce similar legislation in other countries. It can provide a framework and valuable guidance around mechanisms for content moderation by indicating where to focus fact-checking efforts and how quickly to react.

"The key to successful regulation includes appointing trusted flaggers, developing an effective tool for reporting harmful content across platforms, and correctly calculating the necessary moderation reaction

time," Dr. Rizoiu says.

"By understanding the dynamics of content spread, optimizing moderation efforts, and implementing regulations like the EU's Digital Services ACT, we can strive for a healthier and safer digital public square where harmful content is mitigated, and constructive dialog thrives."

More information: Schneider, Philipp J. et al, The effectiveness of moderating harmful online content, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2307360120. doi.org/10.1073/pnas.2307360120

Provided by University of Technology, Sydney