February 15, 2021

OHCHR Report on Disinformation

Response to the Public Consultation

Authors: Jacob Mchangama, Raghav Mendiratta and Natalie Alkiviadou

Justitia

Justitia is Denmark's first judicial think-tank. Justitia aims to promote the rule of law, human rights and fundamental freedoms both within Denmark and abroad by educating and influencing policy experts, decision-makers, and the public. In so doing, Justitia offers legal insight and analysis on a range of contemporary issues.

Future of Free Speech Project

The Future of Free Speech is a collaboration between Justitia, Columbia University's Global Freedom of Expression and Aarhus University's Department of Political Science. We believe that a robust and resilient culture of free speech must be the foundation for the future of any free, democratic society. Even as rapid technological change brings new challenges and threats, free speech must continue to serve as an essential ideal and a fundamental right for all people, regardless of race, ethnicity, religion, nationality, sexual orientation, gender or social standing.

INTRODUCTION

The line between healthy public debate and conscious manipulation is often blurry.[1] Striking this balance and effectively combatting disinformation appears to be an endless game of whack-a-mole for digital technology companies, who are constantly trying to catch up with technologies and tactics developed by miscreants to amplify disinformation. More than two years after the adoption of the EU Code on Practice of Disinformation in 2018, the measures adopted by platforms have been criticised for being insufficient and unsuitable for countering disinformation and for failing to remove disinformation, especially around the COVID-19 pandemic.[2]

After the onset of the COVID-19 pandemic, platforms were expected to be the mode for effectively communicating reliable health information from local governments and national

[1] https://www.reuters.com/article/cyber-disinformation-facebook-twitter-idINKBN26T2XF

[2] https://www.euractiv.com/section/digital/news/eu-code-of-practice-on-disinformation-insufficient-and-unsuitable-member-states-say/

and international health authorities in real-time to billions of people. At the same time, they had to ensure that their platforms did not become vehicles for amplification of harmful disinformation. In attempting to strike this balance, the fault lines of platforms' content moderation policies came under a greater spotlight.

This submission to the UN OHCHR is an attempt to map the policies, procedures and measures adopted by digital technology companies to counter disinformation and to analyse the extent of their compliance with human rights standards on freedom of opinion and expression.

Issue I: What policies, procedures or other measure have digital tech companies introduced to address the problem of disinformation?

- **Political advertising and political disinformation**

In response to allegations of foreign inference and dissemination of disinformation in the 2016 US Presidential Election Campaign and the 2016 Brexit Vote, major platforms such as Facebook, Instagram and Google have pushed towards increasing transparency for political advertising so users are can understand *why* they are seeing certain ads, *who* is paying for the ads they are seeing, *who* the ads are reaching, the *option* of tweaking the ads they see, and other relevant details.[3]

For the recent 2020 US Presidential Elections, Facebook introduced an Ad Library that showed relevant details of the ads that users were seeing including who paid for those ads.[4] Before the elections in November between March and September 2020 alone, Facebook removed 30 networks engaged in coordinated inauthentic behaviour, displayed warnings on more than 150 million pieces of content, and rejected ad submissions that could have potentially run about 2.2 million times for failing to fulfil Facebook's authorization process.[5] In a similar effort to increase transparency, Google releases an annual Bad Ads report that reports the ads taken down both, at the stage of publication, and also those which were removed after publication.[6] Twitter, on the other hand, took a different approach in 2019 when it decided to ban political advertising altogether.[7] This included appeals for votes, solicitations of financial support for

[3] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54455
[4] https://www.facebook.com/ads/library
[5] https://about.fb.com/news/2020/10/preparing-for-election-day/
[6] https://www.blog.google/products/ads/stopping-bad-ads-to-protect-users/
[7] https://twitter.com/jack/status/1189634371407380480

candidates, advocacy for or against any political content.[8] Twitter also placed a ban on microtargeting whereby advertisers could target specific audiences.[9]

Twitter also has a specific policy on unsponsored misinformation and disinformation affecting civic integrity.[10] This policy is aimed at content that suppresses participation or misleads people about when, where, or how to participate in a civic process. As per Twitter, examples of civic processes include political elections, censuses, major referenda and ballot initiatives. [11] As part of this initiative, Twitter takes a variety of actions based on the severity of the action and frequency of the violation. It labels and reduces the visibility of Tweets containing false or misleading information about civic processes to provide additional context. For repeat offenders, Twitter uses a strike system to determine if further enforcement actions should be applied. For high-severity violations of this policy, including misleading information about how to participate, suppression and intimidation, Twitter requires users to remove the content and temporarily locks them out of their accounts before they can Tweet again.[12]

- **Downranking and removing disinformation**

Platforms such as Facebook and Instagram generally downrank false information so that users did not interact with them as much.[13] However, after the onset of COVID-19, there was a trend in platforms including Facebook, Instagram and YouTube to move towards increased removals. For example, in March 2020, Facebook announced that it would not only downrank but remove COVID-19 related false information if it could lead to *imminent* physical harm.[14] Facebook did not officially elaborate or explain the threshold for *imminent* physical harm.[15] However, in an interview with Jacob Mchangama, Monica Bickert said that imminent physical harm that would warrant removal would be for statements such as "the working class is immune from COVID-19".[16] Similarly, YouTube also announced that it would remove "medically unsubstantiated" information about COVID-19 but did not define what it categorises as medically unsubstantiated.[17]

[8] https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html
[9] https://www.cnbc.com/2019/10/30/twitter-bans-political-ads-after-facebook-refused-to-do-so.html
[10] https://help.twitter.com/en/rules-and-policies/election-integrity-policy
[11] https://help.twitter.com/en/rules-and-policies/election-integrity-policy
[12] https://help.twitter.com/en/rules-and-policies/election-integrity-policy
[13] https://www.facebook.com/communitystandards/false_news/
[14] https://about.fb.com/news/2020/03/combating-covid-19-misinformation/
[15] https://www.facebook.com/communitystandards/
[16] http://justitia-int.org/en/clear-and-present-danger-special-edition-monika-bickert/
[17] https://www.bbc.com/news/technology-52388586

- **Targeting spam, coordinated harm and manipulated media**

Platforms have been tackling disinformation by developing policies that limit the abuse of the platform by inauthentic users who use the service to artificially amplify or suppress information. All leading platforms such as Facebook, Google and Twitter have developed transparency rules around automated applications and activities. For example, Twitter defines Spam as bulk or aggressive activity that attempts to manipulate or disrupt Twitter or the experience of users on Twitter to drive traffic or attention to unrelated accounts, products, services, or initiatives. This includes not just automated Tweeting but also posting manipulated or synthetic media such as deep fakes. Based on the degree of alteration/manipulation and the likelihood of causing harm, Twitter decides between labelling the content and removing it.[18] Similarly, YouTube also follows a similar three-strikes system that results in an account being terminated if it has found to be constantly violating YouTube's spam policies.[19] However, despite these policies, platforms have been criticised for not appropriately checking the dissemination of deep fakes on their platforms that have implications for how people interact with information online and offline.[20]

Issues II and III: To what extent do you find these measures to be fair, transparent and effective in protecting human rights, particularly freedom of opinion and expression? What procedures exist to address grievances and provide remedies for users, monitor the action of the companies, and how effective are they?

- **Political advertising and political disinformation**

Measures adopted by platforms such as increased fact-checking and displaying warning signs on false or manipulated ads help educate viewers and combat disinformation. However, the algorithms that are entrusted in identifying and flagging content are often prone to committing mistakes in their identification.[21] And these algorithms are increasingly becoming responsible for most of the content and accounts flagged by these platforms. For example, of the 6.5 billion

[18] https://help.twitter.com/en/rules-and-policies/manipulated-media
[19] https://support.google.com/youtube/answer/2801973?hl=en
[20]https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/
[21] https://www.osce.org/files/f/documents/6/8/302796.pdf

fake accounts that Facebook removed in 2019, 99.7% were identified by algorithms before anyone flagged them to Facebook.[22]

Further, Twitter's complete ban on political advertising poses some key questions for the future of free speech. For example, it is hard to define what constitutes political advertising.[23] Although Twitter has tried to narrow the scope by restricting ads to banners that discuss elections, candidates, parties and other overtly political content, the definition of what constitutes overtly political may differ from person to person.[24] Over time, this may lead to a scope creep in the issues on whom advertising may be disallowed, including abortion, climate change, increasing funding for medical research, etc.[25] Further, although the aim behind regulating political disinformation amplified by advertising is legitimate, banning ads altogether may not be the least restrictive measure to stop political disinformation. In this regard, it might be better for the future of free speech if a more nuanced approach is adopted, such as promoting fact-checking of ads, displaying warnings on political disinformation, increasing transparency to always display the funder behind an ad etc.

- **Downranking and removing disinformation**

While platforms may need to downrank and remove disinformation, vagueness and opacity in deciding how this is done is problematic for free speech. If private actors moderate content in the absence of clear and publicly-available guidelines, users and organisations have no means to know the threshold of harmful speech that warrants downranking or removal.

Conclusion

Despite the expanding policies on disinformation and the proactive enforcement of these policies, platforms have been criticized for failing to satisfactorily combat disinformation, especially in the context of the ongoing COVID-19 pandemic.[26] For example, Plandemic, a documentary-style conspiracy theory video went viral in May 2020. It asserted, amongst other things, that COVID-19 was manufactured in a lab, that masks and gloves made people sicker and that beaches should remain open because seawater had "healing microbes". It gathered

[22] https://about.fb.com/news/2020/10/preparing-for-election-day/
[23] https://www.bbc.com/news/entertainment-arts-50479568
[24] https://www.nytimes.com/2019/11/15/technology/twitter-political-ad-policy.html
[25] https://www.nytimes.com/2019/11/15/technology/twitter-political-ad-policy.html
[26] https://www.bbc.co.uk/news/technology-52903680

over 7 million views across various platforms despite repeated efforts by platforms to pull it down.[27]

The viral dissemination of the Plandemic video shows that broadened community standards and increased removals do not necessarily guarantee an effective check on disinformation. Given the real harms that could be caused by disinformation during a global pandemic or during an election, the urge to encourage platforms to clamp down on content is strong.[28] However, it would be better for the future of online discourse if platforms explored solutions in media literacy, fact-checking, and moved towards consistent, transparent and publicly accessible guidelines on regulating disinformation instead of simply removing more content through automated algorithms. However, as so often when democracies respond to threats and emergencies, there is a real risk of overreach, threatening basic freedoms, not least freedom of expression.[29]

[27] https://www.technologyreview.com/2020/05/07/1001469/facebook-youtube-plandemic-covid-misinformation/

[28] https://www.sciencedirect.com/science/article/pii/S2590061720300569

[29] https://www.lawfareblog.com/rushing-judgment-examining-government-mandated-content-moderation