GOOD SYSTEMS
A UT Grand Challenge

Moderating Social Media Discourse for a Healthy Democracy

Dr. Josephine Lukito, Kathryn Kazanas, Bin Chen, Dr. Dhiraj Murthy, Akaash Kolluri & Pranav Venkatesh

Policy Brief, Published October 26, 2022
Designing Responsible AI Technologies to Curb Disinformation Project

Executive Summary

The prevalence of hate speech and misinformation on the internet, heightened by the COVID-19 pandemic, directly harms minority groups that are the target of vitriol, as well as our society at large (Müller & Scwarz, 2020). In addition, the intersection between the two only exacerbates their harmful effects leading to an increase in intolerance and polarization (Kim & Kesari 2021). Current platform moderation techniques, as well as Section 230 under the Communications Decency Act, have been insufficient in addressing this problem, resulting in a lack of transparency from internet service providers, clear boundaries on user-platforms relations, and sufficient tools to handle a rapidly expanding internet.

To address this problem space, we advocate for the following solutions:

1. ***Algorithmic governance & transparency***: Internet Service Providers should be more transparent with users about content moderation policies and algorithms, and clarify users' basic rights on the platform.
2. ***Flagging recommendations***: We advocate a more effective, efficient and comprehensive flagging system through a combined strategy of content- and user-based approaches.
3. ***Multiplatform collaboration***: Fighting harmful online content requires a collaborative effort among policy makers, civil society groups, researchers, and different platforms.
4. ***Long-term considerations***: Building a regular and prolonged tracking system is essential to make anti-misinformation efforts more efficient and effective, especially in complex scenarios.

Problem Space

The Problem with Hate Speech & Misinformation

Hate speech incites prejudice against a specific group in society. Often, hate speech is uttered to undermine a minority group's social position by suppressing their voice and isolating them from society. Hate speech encourages discrimination and intolerance, and sometimes will even lead to violence, causing physical, psychological, financial and societal damage to minority groups (Müller & Scwarz, 2020). As hate speech hinders a minority group's participation in society it also hinders their participation in the democratic process.

Misinformation heightens the effects of hate speech (Cinelli et al., 2021). Often, misinformation is used to justify hate speech and discrimination, and widespread misinformation also leads to increased polarization and intolerance. The effects of misinformation are also persistent, as misinformation is resistant to corrections, including the backfire effect, when a person's belief in misinformation can actually strengthen because of corrections (Persily 2020). Additionally, misinformation spreads more virally than fact. All of this allows misinformation to build up, creating echo chambers where people receive only information (even if it's false) that reinforces their beliefs.

This combination of hate speech and misinformation leads to harmful, long-lasting effects that continue to influence a society even after it has been corrected.

Examples of Hate Speech and Misinformation

Some research suggests that hate speech and misinformation interact, or occur together, in many ways (Müller & Schwarz, 2021). Sometimes, stereotypes about a racial minority may feed into more concrete (or specific) forms of misinformation. When hate speech and misinformation occur together during a salient media moment, particularly if a minority group is the perceived enemy, such moments may be prone to violence (Müller & Scwarz, 2020). Finally, hate speech and misinformation often occur in tandem when discussing a political group, event, or individual. In all these instances, hate speech and misinformation can also occur alongside other unwanted or harmful online discourse, such as outlandish conspiracy theories.

Below, we illustrate some examples, drawing from data collected from Parler, a social media platform that is popular among far-right extremists.

Warning: These posts contain racist remarks.

Example 1: Anti-Black

"So Donald Trump who has no proof, data, or facts of being a racist means you're a racist if you vote for him. And Biden who publicly used the term "n[****]" referring to blacks and made the worst crime bill which allowed for mass incarceration for black people in 1994 isn't racist if you vote for him? WHAT THE FUCK IS THAT!!!!!#parler #parlerusa #parlerksa #twexit"

This post demonstrates how the manipulation of facts can also constitute a form of misinformation. Above, there are two claims about Biden which may be factual at face-value, but are deliberately misleading by leaving out context or important details. The former, claiming that Biden used the n-word, was rated false by Politifact. The latter claim about Biden's crime bill was rated as half-true by Politifact, but the above example makes it clear that there is an intent to mislead when combined with the first false claim.

Example 2: Anti-Asian

"Chinese communists have many people working at pfizer to produce the bill gates vaccine. yeah, this vaccine will kill you and your kids. media censoring vaccine information. i don't trust them."

This post demonstrates how a strong, false claim ("this vaccine will kill you and your kids") can be combined with anti-China, anti-elite, anti-media, and anti-vax sentiments. Racist misinformation often emerges during a salient news topic such as the COVID-19 pandemic.

Example 3: Anti-Jewish

"Tainted Swabs. #IToldYa Never give them your #DNA. #BigPharmaGrabblers NEED it to make the next fogging blast more effective. WakeUp! They DO have ethnic specific biowarfare. They use this #Scamdemic to kill off the seniors in the nursing homes, collect on the in$urance, then sit back and laugh at the goyim wearing masks to make us all #NPC."

This example highlights how historical hate speech claims can be applied to spread modern-day misinformation. Many of the conspiratorial claims here can be traced back to centuries of antisemitism, evoking historic stereotypes about genetics and financial activity.

Example 4: Anti-Muslim

"Genocidal Democrat Nazi, George Soros, is SUPPORTING the genocidal Islamist Iranian mullahs in their quest to annihilate America."

"Laura, De Blasio does not "hate" Jews, HE WANTS JEWS ANNIHILATED ! He is a TYPICAL Democrat in bed with genocidal Islamist terrorists. Their ultimate goal is to wipe out Christians and turn America into a socialist state with genocidal Islamists as our overlords."

These two examples highlight how misinformation and conspiracies can be used to attack democratic politicians; in this case, by implying secret partnerships between Democrat politicians/supporters and Islamic organizations. Such conspiratorial claims often use expressions of fear.

To combat the kinds of posts that we see above, social media platforms often engage in platform moderating tactics that are as narrowly focused as deleting one post or as broad as suspending a user's account. Below, we explore these tactics in greater detail.

Existing Platform Moderation Tactics

Presently, social media platforms have a range of strategies they can employ to curb misinformation, hate speech, and other unwanted information on their platform. Some of these methods focus on singular messages, while others are more severe, publishing the user for repeat offenses.

We conceptualize these varied options into four categories.

Moderator	Type	Description
User	**Message Filtering**	Allowing an individual to control the type of content that is on their feed by allowing or *disallowing* certain content. • This can be done by flagging keywords or phrases to avoid and is a more preventative measure.
Platform	**Shadowbanning**	Cutting the user off from certain parts of the site. • This can be used to de-escalate tension.
Platform	**Message Removal**	Completely removing a post from the site. • If a post contains extremely harmful or obscene material (such as direct threats of violence) it should be removed entirely.
Platform	**User Suspension**	Removing a user from posting on a site • If there is a pattern of harmful behavior from a certain user, the social media site can consider suspending or banning the user entirely from the site.

Despite these range of options, the decision to remove a piece of content or to suspend a user is often piecemeal, with relatively little consistency or transparency regarding how platforms will moderate their content. This may stem, in part, from the terms that platforms use to identify harmful content (like coordinated inauthentic behavior).

Section 230

One relevant policy to this discussion is Section 230. Section 230, under the Communications Decency Act, protects websites, platforms and internet service providers from charges related to third-party content. They are given immunity to any charges brought about by content from another provider, unless the website specifically encourages illegal content.

Section 230 was intended to encourage the development and innovation of small to midsize internet service providers (ISPs). However, as it is now, Section 230 is unequipped to handle large conglomerates like Facebook, Google or Twitter nor the central role that the internet plays in everyday life. ISPs have consistently received broad protection under Section 230, but as the line between real life and the internet becomes blurred, it is worth clarifying the exact relationship between user and ISP.

Recommendations

Below, we describe four potential solutions that, combined, would help social media platforms address the problem space. Some of these solutions are inspired by the recently proposed AI Bill of Rights; however, our suggestions for social media companies go beyond automated systems of media platforms.

Algorithmic Governance & Transparency

One necessity that is agreed upon is **the need for more transparency from corporations on issues like moderation and algorithms**, whether through self-reporting or a public audit.

A second recommendation is to **clarify the basic rights of users on the internet.** While the basic rights have thus far focused on privacy issues, policies around discrimination, calls to violence, intimidation, and defamation are also important.

Finally**, there needs to be a consensus on whether a company's algorithm, and the resulting products (recommendation feeds, etc.) are the intellectual property of an ISP**. If an ISP claims a unique algorithm as company property, then it stands to reason that the resulting damage caused by defamation, misinformation, and hate speech is the partial responsibility of the company even if the original content is produced by a third party. The internet is rapidly evolving, so mapping out clear expectations is imperative to achieve an equal and dependable relationship between companies and their users.

Flagging Recommendations

Flagging is a mechanism to alert users that the content they come across is objectionable or violates terms of service. Although flagging is commonly employed by social media platforms as a step preceding message, the effectiveness of the current flagging is limited (e.g., Chopdza & Yan, 2022). **Thus, we advocate a more effective, efficient, and comprehensive flagging system that** is attentive to misinformation and hate speech in non-English.

Many social media platforms use algorithmic and human content detection methods to identify objectional content (misinformation, hate speech, and violence). However these detection methods focus almost exclusively on English-language content, despite the prevalence of non-English misinformation. Additionally, we advocate for user-level flagging systems, as a substantial amount of misinformation spread can stem from only a few users (the Center for Countering Digital Hate, for example, has identified the "Disinformation Dozen," who were said to spread 65% of COVID-19 anti-vaccination misinformation)

Multi-platform collaboration

Mis/disinformation often flows from one platform to another through the sharing or URLs, screenshots, or content shared by the same user on different platforms. **We therefore advocate that social media platforms collaborate with one another to share information about harmful online content.**

There are two approaches to trace and control the spread of misinformation: user-based, and content-based. User-based tracing refers to identifying misinformation spreaders across

platforms using identifiable meta-data, such as screennames. Content-based approaches could begin by building a transparent, shared list of keywords or semantically similar phrases that are indicative of misinformation and can be queried across platforms.

Long-Term Considerations

Were these measures to be enacted, it is important to remember that misinformation detection is an ongoing challenge, and not one that simply emerges around an election. Therefore, social media **platforms should regularize these measures and establish a prolonged tracking system to identify misinformation in complex scenarios.** Combating misinformation is a marathon and requires constant efforts and consistent collaboration between different social media platforms. The suggestions given above including improving algorithmic transparency, strengthening flagging system, and enhancing cross-platform collaboration are mostly temporary measures.