The Unintended Harms of Banning Based on Misinformation

Mahira Ali (3121909) Murtuza Ali (3123812)
Vyshnavi Molakala Narasimhalu (3012247) Zohar Cochavi (3137686)

May 4, 2023

4TU.Cyber Security
TU Delft UNIVERSITY OF TWENTE.

Contents

1 Introduction

Recently, it has been brought to our attention that users have started to post false information regarding election outcomes. This lead to confusion and mistrust among the public. We have been asked if banning these users is the right action. It is essential to note that simply banning these malicious users who spread misinformation may not be the most effective solution, as it may not address the root cause of the problem. The question of whether to start banning users is a loaded one. On the one hand, we have to secure the user experience and the safety of our community. In contrast, on the other, we have to satisfy every user's right to an opinion and express that opinion. The obvious question is whether one perspective weighs heavier, and if so, which one?

As shown by the case of Facebook banning accounts based on sharing violent imagery, integrating a banning policy into the ToS can have unintended consequences and harms to both the community and the company (Reed & Henschke, 2021a). The long-term effects of such a policy are further stressed by some social media companies becoming increasingly risk-averse after deciding to ban the spread and publication of misinformation on vaccines. This eventually lead to them curtailing freedom of expression and censoring legitimate behaviour (Mills & Sivelä, 2021).

To avoid such unintended harms, we will be using the framework proposed by Chua et al. (2019), *Identifying Unintended Harms of Cybersecurity Countermeasures.* While this framework has been developed for assessing the design of cybersecurity measures in organisations, we believe it will also prove effective in determining if policy regarding our domain produces any unintended harms in the short- as well as long-term. With this framework and the use of case studies, we intend to propose a long-term strategy to promote healthy discourse and avoid harm to our platform and the community.

First, we will study the current 'state' of banning, drawing a landscape of current opinions and existing policies in similar companies, and conclude, by looking at the unintended harms, why we should not ban users (Section 2). Then, a counterargument to the previous section will be provided (Section 3). In the succeeding section, we propose two alternatives to banning and analyse their different advantages and disadvantages (Section 4). Finally, we will conclude the report by giving our preferred alternative (Section 5).

2 The Case Against Banning

To confidently argue why we should not ban users on our platform, it is worth taking a look at current legislation and public opinion on the matter of banning users (Section 2.1). Then, we analyze banning within our domain through the lens of the aforementioned framework (Section 2.2).

2.1 The Current State of Banning

Banning users based on spreading misinformation is, of course, not unique to the political domain. The COVID-19 pandemic showed how the spread of misinformation could dissuade persons from following regulations meant to protect the general public (social distancing, wearing face masks, or taking vaccines[1]) (Bridgman et al., 2020). Even though this information can be shown to be false, or based on falsehoods, with the use of scientific research, it might still be undesirable to ban users (Fetzer, 2004). With this, it seems that simply banning users based on the spread of misinformation, regardless of the domain, is not a clear-cut argument.

[1] There are legitimate healthcare-related reasons why someone would refuse to take a vaccine or wear a mask (Lehmann & Lehmann, 2021). Here, we specifically refer to people who to do so based on false information, regardless of the outcome.

One particularly interesting topic is the distinction between that of *misinformation* and *disinformation*. Where the former is the general term for any type of factual error, the latter is distinguished by the intent of the person spreading the misinformation. In short, if a user spreads misinformation with the purpose of misinforming others it is called disinformation (Andersen & Søe, 2020; Fetzer, 2004). Of course, neither are desirable, but the intent is an important distinction. Most would agree that banning a user based on spreading disinformation is less ethically contested than banning on the basis of spreading misinformation. For the remainder of this report, while we will use both terms interchangeably, we will exclusively talk about disinformation.

2.1.1 Handling Extremism

The spread of this kind of misinformation is not only a philosophical danger but also a physical one (Cohen et al., 2013). While we should be careful in labelling this kind of behaviour as extremist, the definitions of *radicalisation* and *black propaganda* do fit the current scenario well (Gaikwad et al., 2021). Furthermore, this topic touches upon the barrier of platform policy and government legislation and the complexities that come with it (Heldt, 2019; Reed & Henschke, 2021b).

Jackson (2019) demonstrates the difficulty of banning users based on spreading *extremist content* by showing the fragility of the definition of extremism. The author does, however, recommend private companies act on those who spread extremism, adding that it should be clear what the goal of the measures should be. An example would be the goal of reducing *extremist violence* as opposed to that of *extremism*, where they argue the former would not be significantly impacted by banning, as compared to the latter. This emphasises the significance of the policy, either banning or otherwise, having a clear goal and assessing if the policy achieves the *intended* goal.

As mentioned before, the case presented by Reed and Henschke (2021b) shows that the start of a 'reasonable level' of control over what is shared on a platform can set a precedent for the suppression of political debate. They proceed to discuss if the ultimate responsibility for these kinds of hate speech resides with the organization hosting the content, or the government. According to the authors, while legislation is often in place, such as the First Amendment in the USA, the interpretation of this legislation is largely up to the company implementing the policy. Some even go as far as to say that "governments are essentially trying to do the work for them" (Balkin, 2018). In conclusion, they advocate for more better collaboration between government and private companies.

While the field does not immediately provide solutions to this issue, it is worth noting that this intersection between legislation and platform monitoring is a grey area. Active research is being conducted on this issue, and governments, as well as companies, should work together to broaden their knowledge in this research.

With this, it seems clear that there is a necessity to implement policy and take responsibility for user-generated content. This begs the question of how such regulations should be enforced and, more importantly, how we should determine that ToS have been breached.

2.1.2 Monitoring Strategies

This brings us to the practical hurdle of monitoring and moderating. Ideally, we would exclusively employ trained moderators to execute this monitoring. But given the huge throughput of new content found on most social media this is not viable (Gillespie, 2020; Schultz, 2019).

One strategy for screening user-generated content is through the use of statistical, machine learning, or AI models (Al-Ghamdi, 2021). Now, the ethics of artificial have long been a point of discussion (Siau & Wang, 2020), but it does provide a solution to the problem of scale (Gillespie, 2020). Gillespie (2020) also mentions that human moderators "do psychologically scarring work, in sometimes intolerable conditions". This, the author argues, is one of the strongest arguments for the use of even *poorly performing* AI in such contexts.

2.2 Unintended Harms of Banning

At first glance, banning users that compromise the integrity of our social media platform by sharing misinformation, or even other behaviours such as propagating racist, homophobic content or other forms of hate speech seems to be a reasonable opinion under the assumption that it would improve the quality of discourse and interaction on the platform. However, using the work of Chua et al. (2019). on unintended harms, we can make a case for how banning these users may not lead to the best outcomes. The paper discusses how certain technologies or procedures when put into place, could end up having adverse unintended consequences such as affecting user behaviour, their inclusion, or the general perception of the platform itself and proposes a framework to highlight these consequences.

The framework describes the unintended consequences of a measure by categorising them into seven broad categories: Displacement, Insecure norms, Additional costs, Misuse, Misclassification, Amplification and Disruption. Based on these consequences, the framework poses a series of questions that one should ask when considering the addition of a new measure, allowing them to make an informed decision regarding adding any new features and their possible side effects.

This is an essential consideration as it is possible that the harms caused by the policy changes might not be apparent to the company but might serve as a catalyst to exacerbate the behaviour we aimed at curbing in some other location. We will go on to list some possible unintended harms that may occur as a result of a decision to ban users who post election misinformation on our social media platform:

Displacement: Removing users from our platform for peddling misinformation on a topic does not eliminate the problem, we are simply temporarily disrupting the problem, or *displacing* the users to another platform. There is a well documented behaviour of users of extremist groups who are banned moving to other social media platforms that are made especially for them (Horta Ribeiro et al., 2021). The authors go on to claim that this leads to the creation of more extreme albeit smaller communities elsewhere. Banning these groups could also lead to the fragmentation of society with each individual interacting only with other users who agree with them, creating a lot of echo chambers for each viewpoint. If we were to consider the concept of the marketplace of ideas[2], this is the worst possible outcome, as the possibility of discourse and understanding is greatly reduced, and the chances for the 'truth' to surface decrease.

Insecure Norms: While our content moderation strategy is very effective, it is not foolproof and can miss certain posts and users. If we were to follow a strict ban policy on posts and users that we deem to be spreading misinformation, any content that we miss might lend a stamp of authenticity to it (Caulfield et al., 2020), seeing as how most other misinformation was banned. This would lull the users into a false sense of security, causing them to believe misinformation that may not have been flagged by the system, leading to weaker *Security Norms* This might mislead users on the platform who would not have had this misconception otherwise.

Additional Costs: Due to the adverse consequences of sharing misinformation, users might have to put in a lot more effort while linking any form of media, such as articles, videos, or sharing their opinions, to ensure that they abide by the company's Terms of service. Having to go through the Terms of Conduct or such document, and having to make a judgement as to whether their posts violate any such conditions is a significant added cost to the user.

[2] The marketplace of ideas theory holds that the truth will emerge from the competition of ideas in a free, transparent public discourse and concludes that ideas and ideologies will be culled according to their superiority or inferiority and widespread acceptance among the population.

Misuse: Using bans as a punishment leads to users inevitably trying to resort to any means possible to try and get their opponents banned, this is visible especially when figures with large followings engage in discourse. Due to the option of getting someone banned, the focus shifts from having good faith debates into simply trying to "de-platform" the opponent to score online points or to gain favour with their own audience. This might lead to the dilution of discourse online with users resorting to cheap tactics, such as mass reporting, or flagging to get users they disagree with banned.

Misclassification: Due to the scale of our platform, manual moderation is near impossible, requiring us to rely on automated moderating systems, this might often lead to *misclassifications*, as mentioned earlier, this may lend credibility to misinformation that was not flagged, but also has a consequence in accidentally banning innocent users, which could lead to a sense of frustration and being wronged especially when they try to appeal the bans(West, 2018), this could also cause users to resort to alternative platforms which might not have the same moral compass as we do.

Amplification: The advent of social media has brought about several interesting social mechanisms, one of them being the 'Streisand Effect', coined after Barbara Streisand, who tried to subdue photos of her house from being posted online. In trying to do so the photo of her house exploded in popularity. The Streisand Effect details how trying to subdue or hide information about something in a public forum such as the internet could lead to the unintended consequence of increasing awareness of that information. In banning users that post misinformation, we might inadvertently spread it to even more users.

Disruption: As seen in recent times, several extremist groups foster the belief that most forms of popular media are under the control of a select few, who misuse their position of power to suppress the voices of people who do not agree with their ideologies (Engesser et al., 2017) (Major, 2012). As most of the content that is flagged for violating our terms of service is usually from some form of extremist groups who are working to push forward their agenda, simply banning them would only help them push their narrative of being silenced by an oppressive authority. This allows them to undermine the efficacy of other measures such as fact checking by claiming that the platform and related fact checking organisations are biased.

3 The Case for Banning

We can see several advocates for banning users that behave in a way that may be classified as harmful. These advocates often mention that posts which glorify violence, enable or promote violence towards a minority, spread terrorist propaganda, etc should be banned. They often justify these forms of suppression stating that these lead to actual harm to other people. This can be seen in the practices of several social media platforms, which ban users for threatening violence against an individual or a group of people, among other reasons. Taking into account events such as the January 6th insurrection at the US Capitol and the Pro-Bolsonaro riots in Brazil, We can see that spreading misinformation can actually lead to physical harm and damage to people and property. Protecting vulnerable groups, minorities and upholding basic rights to safety are valid reasons to implement measures such as banning users who spread misinformation on such topics. Other reasons that support a decision to ban such users are:

- The prevalence of misinformation on our platform makes it much harder for normal users to consume the news. A study shows that users are more susceptible to fake news, due to a "lack of thinking" rather than preexisting bias (Pennycook & Rand, 2019). This would indicate that users would have to put in additional time to ruminate over each article of news they get in order to ensure that they are not mislead. Considering the amount of posts a user interacts with and the usual use case of scrolling through posts on social media, this becomes an unnatural expectation from our users. This can cause the user to become dejected and jaded, leading them to lose interest in the topic or to become extremely skeptical.

- Users that are more impressionable such as young adults, people from less privileged backgrounds or those who might not have a good educational background have been shown to be especially susceptible to medical misinformation (Nan et al., 2022). We can assume that there also exist similar groups or similar demographics that may also be susceptible to political misinformation as seen by election denial by the right wing in the 2020 elections (Pennycook & Rand, 2021). These sources can be predatory and might exploit these groups by feeding them cherry picked facts, We have an obligation to all our users and must take appropriate actions to ensure that vulnerable groups can also safely interact on our platform.

- There are several users that have friends and family added on their accounts, and have posts dating from the time they joined the platform. They may also have important chats and interactions that they would like to preserve. Some users also use their accounts for work, or to promote their businesses in the case of family owned businesses or self employed individuals. This leads to a value being associated with the account as shown by a study asking users to assign a monetary value to their accounts (Corrigan et al., 2018). This would dissuade the user from performing actions that could jeopardise their account.

  If we were to use a method to verify identity using a unique identifier such as a phone number, we could guarantee that users would not be able to make burner accounts, or even if they could, we could limit this to a reasonable number, as opposed to millions of bot created accounts. This factor combined with the perceived value of a users account would greatly reduce the amount of malicious misinformation spread.

- Using methods such as reducing reach of the users posts are useful only for small accounts, users with existing reach and fame, such as media personalities often have a large amount of followers and these methods will not stop this spread.

- Using automated systems to ban misinformation bots or channels is effective as it usually takes time to accrue a following, repeatedly banning offenders prevents them from increasing or even maintaining their reach.

4 Alternatives

This section presents two alternatives for banning users who spread misinformation. First, we will examine the alternative to suppress messages. Secondly, an analysis of showing the contradicting truth alongside the misinformation will be performed. Here, both alternatives distinguish two types, active and passive.

4.1 Suppress Misinformation Spreading Messages

One of the alternatives to prevent the spread of misinformation is to suppress misleading messages. Specifically, we will discuss message removal upon which they are no longer visible. This paper considers two types of suppression active suppression and passive suppression. Active suppression implies that the platform's users flag messages they perceive as misleading. Subsequently, the flagged messages are manually assessed and removed upon the conclusion of misinformation. Furthermore, passive suppression connotes algorithms detecting misleading information by catching specific phrases or words.

4.1.1 Active Suppression

The study by Mena (2019) showed that active suppression reduces sharing of misinformation. Additionally, this study observed the third-person effect, meaning that users strongly believed that other users were more likely to spread misleading news than themselves (Perloff, 2009). Contrarily, Avaaz (2021, 2020a, 2020b) reports considered flagging users who spread incorrect information instead of messages. Then, a check was performed on the future posts of these flagged users before release.

Whether to flag messages or users, users should be educated on spotting and reporting misinformation. This could include tips on how to evaluate the credibility of sources, how to fact-check information, and how to report false or misleading content. Users often fall to the illusory truth effect, in which the human brain interprets repetitious information as valid (Hassan & Barber, 2021). Or they may succumb to confirmation bias, which is the tendency to be less distrustful of facts that already match one's perspective (Nickerson, 1998). Misleading information is often written using extreme sentiments to engage the reader with the content, making the misinformation viral on social media. The users must be able to differentiate facts from fiction. A survey report reveals the effectiveness of training and educating users on spotting misinformation (Breakstone et al., 2021). About 3400 students were able to assess the credibility of information on social media and differentiate facts from misleading content. News Literacy Project[3] is an educational non-profit US project that educates people on what to trust and what not to trust on social media. By investing in media literacy activities, the impact of misinformation can be reduced and promote critical thinking among users (Kajimoto & Fleming, 2019). This can also make users discerning information consumers who are less likely to share or believe false or misleading content and more likely to report it to the platform or fact-checkers. Critical thinking is essential in a crisis or breaking news event when emotions may run high, and people may be more prone to believing or sharing unverified information (Leigh, 2015a).

4.1.2 Passive Suppression

Passive suppression can be applied instead of manual inspection to cope with the high message flow. This was demonstrated by a study by Leigh (2015b) in which they identified users who are more likely to share fake news based on such users' language upon which linguistic profiles were created. The two main observations were the expression of negative emotions and existential-based needs. Misinformation-detecting algorithms that incorporate these observations caused a remarkable improvement in the accuracy of classifying messages as misinformation.

In addition to the writing style of the message, the remaining aspects considered when assessing misinformation are propagation patterns and the knowledge the message carries (Zhou & Zafarani, 2021). Propagation patterns indicate false information since malicious users usually participate in the circulation of misinformation to amplify their impact. Detection of propagation patterns is more robust against word manipulations. However, it is not possible to detect misinformation before dissemination. Consequently, malicious users will always accomplish their goal of spreading misinformation to some extent. Hence, early detection of misinformation is crucial to reduce the number of people influenced and minimise the damage.
Besides that, algorithms can compare the message's knowledge with existing knowledge. For this purpose, social media giants like Facebook, Google, and YouTube invested in collaborating with the non-profit fact-checking organisation - International Fact-Checking Network (IFCN) (Centre, 2023; Navlakha, 2022). This is because they realised the potential consequences of disseminating election misinformation. Thus they want to identify and label misinformation and impose punishments on repeat offenders. Two reasons for joining forces with fact-checking organisations are as follows. Firstly, collaborating with these organisations can assist social media companies in identifying false or misleading information more quickly and accurately. Fact-checking organisations often have teams of

[3] https://newslit.org/

researchers and journalists trained to investigate and verify the accuracy of claims and information. Secondly, they can aid in providing users with accurate and reliable information. When false or misleading information is identified and flagged, the social media company can provide users with links to check facts or other reliable sources to help users understand the truth about the issue. This can effectively reduce the visibility and impact of misinformation on the platform since it can help prevent it from being shared or endorsed by other users and avert users from being misled.

4.1.3 Unintentional Harms of Message Suppression

Hence, suppressing messages obstruct the achievement of spreading misinformation. Nonetheless, let us now examine the unintentional harms the suppression of messages brings along, according to the framework by Chua et al. (2019).

Displacement: There are two types of displacement. One is the circumvention of the detection. Malicious users can adjust their language usage or create pseudo accounts to bypass the detection of misinformation. As a consequence, they nonetheless accomplish spread of false information. Next, suppressing messages on one platform can cause a shift to other mediums that neglect or are less strict on spreading misinformation. Consequently, the platform loses users while spreading misinformation is still achieved.

Insecure Norms: When a platform ensures its users that false messages are blocked or removed, users are more likely to assume that the remaining accessible information is correct. As a consequence, the alertness of users to detect misinformation drops. Moreover, allying with fact-checking organisations can create a perception of bias on the part of the social media company. Users may view their efforts as biased or partisan if, in their perspective, the company is too closely aligned with certain fact-checking organisations. This could lead to a loss of trust in the platform and reduce its credibility (Rich et al., 2020).

Additional costs: The development of an accurate but also fast algorithm that detects misinformation costs time and money. This is because platforms must expand their team to create such an algorithm. For instance, experiments showed that the initial algorithm designed by Pham et al. (2020) required too much time to cope with the high stream of incoming messages. Hence, extra time was needed to create a faster algorithm.

Misuse: By educating users on detecting misinformation, malicious users learn which indications to circumvent. Or users may try to use the report function to silence or harass others. Besides user misuse, platform misuse is equally essential to forestall. Suppressing messages could lead to censorship or oppress certain viewpoints or perspectives. In consequence, users cannot post messages that do not align with the views of these organisations. Furthermore, this can cause displacement such that users shift to other accommodating platforms where they might create reputational damage to the platform.

Misclassification: A message can be falsely classified to spread false information. This causes invalid blocking or removal upon which users cannot share their feelings or opinion. Consequently, users cannot participate in discussions and current events crucial for personal and global wellness (Gigone & Hastie, 1993; Oeldorf-Hirsch & Sundar, 2015).

Amplification: Removing messages may result in the Streisand effect, where taking down information can draw more attention. As a result, the spread of misinformation is amplified, which is the primary goal of malicious users. Therefore, in preference to removing misleading content, one can consider reducing its visibility on your platform.

Disruption: Educating users on catching misinformation can interrupt passive suppression mechanisms. Including false information can intentionally cause misclassification or biases in algorithms. Additionally, there is a risk that providing too many resources or too much information on misinformation can overwhelm or confuse users. Therefore, finding a balance and focussing on key messages and strategies users can easily understand and apply is essential.

4.2 Expose Facts alongside Misinformation

Another alternative to prevent the spread of misinformation is providing factual information alongside the content with misinformation. This approach allows users to compare the misinformation with accurate information, which can help them to evaluate the credibility of the information they are viewing.

4.2.1 Active and Passive Trigger for Exposure Mechanism

Firstly, we can actively trigger the exposure mechanism by providing users with a feature that allows them to report content they believe contains misinformation. This feature permits users to flag content they believe is false or misleading and provides a space for them to give reasons for their report. Once a piece of content is reported, it could be reviewed by the social media company's team of fact-checkers or AI-based models. If deemed to contain misinformation, alternative information is provided alongside it.

Secondly, the exposure mechanism can be passively triggered by algorithms that automatically identify and flag content containing misinformation. Such an algorithm can use machine learning techniques to analyse the posted text, images, and videos upon which it looks for indicative patterns of misinformation. For instance, the algorithm can recognise specific keywords or phrases commonly used in misinformation or detect images manipulated or doctored to support false claims.

Regardless of active or passive triggering, once the content has been flagged as potentially containing misinformation, the platform can provide alternative information alongside it. For this purpose, we can include a fact-checking label or warning and a link to a credible source providing accurate information on the topic. Alternatively, the platform could also use push notifications, emails, or other means to alert users to the flagged content and encourage them to view the alternative information.

4.2.2 Real-World Examples

Here, we will look at some real-world examples to understand the benefits of providing users with factual information alongside the content of misinformation.

Firstly, we consider Facebook, which developed a feature that automatically fact-checks stories uploaded on the network during the 2020 US Presidential elections. This feature identifies potentially inaccurate or misleading stories and refers them to third-party fact-checkers using a combination of human fact-checkers and machine learning algorithms. Upon discovering an item containing misinformation, it was labelled with a warning label, and users were supplied with alternative information beside it. According to the US 2020 Elections Report of Facebook, the technique helped limit the dissemination of content judged misleading by fact-checkers by an average of 80 per cent globally (FaceBook, 2020). This indicates that the function successfully controlled the spread of disinformation on the platform. Furthermore, Facebook stated that the feature reduced the number of people who would have seen misleading information in their news feeds by 70 per cent on average.

Next, suppose Twitter. In 2019, Twitter announced a new feature called "Birdwatch," which allows users to report tweets they believe spread misinformation (Jha, 2022). Then, such tweets are reviewed by a community of fact-checkers. These fact-checkers add notes to tweets that provide users with alternative information, context, or clarification. Herewith, Twitter aims that this community-driven approach helps combat misinformation on the platform. Several researchers and media outlets praised Twitter's efforts for being innovative and proactive in tackling misinformation. Moreover, the company stated that it saw a reduction in users sharing misinformation after implementing the feature.

Lastly, we consider Google, which operates YouTube. Google also has a fact-checking feature that uses information from third-party fact-checking organisations to provide alternative information to users when they search for specific terms. If a fact-checking organisation has rated a claim as false, the fact-checker's summary will appear next to the search results. Additionally, flagged content is ranked lower in search results to ensure the range of the misinformation is limited.

4.2.3 Unintentional Harms of Fact Exposure

Thus, exposing facts alongside misinformation reduces the spread and restricts the reach of misinformation. Nevertheless, we will now examine the unintentional harms that fact exposure carries again according to the framework by Chua et al. (2019).

Displacement: Malicious users can bypass passive detection by adjusting their language. Consequently, no flag of misinformation is raised, upon which no actions are undertaken against the respective message. Hence, attackers succeed in spreading misinformation. Alternatively, users transfer to platforms with less strict mechanisms against misinformation. Either way, malicious users accomplish spreading misinformation.

Insecure Norms: The users' alertness decreases as they are more likely to assume that the information is valid during the lack of contradictory facts. Consequently, users might develop an incorrect view of certain topics, and their ability to flag misinformation could reduce.

Additional Costs: The platform needs to invest in fact-holding databases to enable correct information alongside misinformation. Additionally, the development of misinformation-detecting algorithms requires the development team to expand.

Misuse: The platform must be cautious not to become biased and partisan. This can happen when they selectively fact-check certain types of content or target specific groups of users, such as politicians. Consequently, marginalised communities or those with certain ideologies become more likely to be targeted by misinformation campaigns.

Misclassification: There is the risk of false positives. These are instances where a user is flagged as malicious, even though they are not spreading misinformation. However, compared to suppressing messages, fact exposure has less severe consequences regarding misclassifications. This is because, here, the quest for contradictory information eventually fails. As a result, the platform discards the flag indicating misinformation, and they undertake no actions against the message.

Amplification: Providing alternative information alongside misinformation could inadvertently amplify the reach of the misinformation. When users see the alternative information, they might also click through to view the original misinformation, which could increase its visibility and engagement. This could be especially true if the alternative information is not presented persuasively or compellingly.

Disruption: Providing users with a feature to flag misinformation requires educating them on how to acquire misinformation. However, attackers can use the educational information to dodge the trigger of misinformation. Consequently, the platform cannot undertake immediate actions whereby the attacker succeeds in spreading misinformation. By the time flags are raised, the attacker has accomplished a certain amount of damage.

5 Conclusion

Let us now return to whether banning users is the optimal response to limit the spread of misinformation. Banning users has the severe consequence that users can no longer express their opinion. The concept is heavily contested but does show some agreement in extreme cases. The core question is how to balance the right to express an opinion with the safety of our platform and its users.

We should consider extrinsic factors such as bad actors who may question our motives and paint us in a bad light, evolving ban evasion measures and the presence of other social media platforms, which may not moderate as strictly as others. With this context, banning users can carry unintended harms which we analysed using the framework proposed by Chua et al. (2019). These unintended harms include but are not limited to: users migrating to other platforms, leading to a false sense of security, and polarization in society. We also considered some arguments made in favour of banning users that spread misinformation. While seemly valid, these arguments often fall apart in considering them in a broader context and the dynamic landscape of social media. Considering these factors among several others, committing to as draconian a measure as banning users seems to be counterproductive. Instead of banning users, we have come up with certain strategies to counter the spread of misinformation.

First, we examined the alternative to suppress messages. Here, messages which spread misinformation are removed or blocked. Suppression can either happen actively, where users flag misinformation, or passively, where algorithms detect misinformation. These algorithms mainly focus on language usage, propagation patterns and type of information. Suppressing messages allows users to share their opinions and thoughts while blocking any misinformation. In contrast to banning, users who spread misinformation remain active on the platform but are restricted in what they can post. In this way, displacement is less likely to occur.

Secondly, we have analysed the possibility of revealing facts alongside the misinformation. Herewith, the user's ability to compare the misinformation with valid information is eased. The exposure mechanism can either be actively or passively triggered. Active triggering occurs upon flag raising by users. Whereas passive triggering happens with algorithms and fact-checking organisations. Unlike banning users and suppressing messages, misuse and misclassification have less severe consequences. This is because fact exposure requires a quest for contradictory information upon identifying misinformation. The misinformation flag is withdrawn when the exploration fails. Hence, no actions are taken against the post in case of misuse or misclassification.

In short, compared to banning users and suppressing messages, fact exposure is the most optimal action to reduce the spread of misinformation. A hybrid approach of active and passive triggering must be implemented since solely actively triggering is infeasible. However, further research is required to develop algorithms that cope with the high flow of messages and deter malicious workarounds.

References

Al-Ghamdi, L. M. (2021). Towards adopting AI techniques for monitoring social media activities [Num Pages: 15-22 Place: Sarajevo, Bosnia And Herzegovina Publisher: Research and Development Academy Section: Articles]. *Sustainable Engineering and Innovation*, *3*(1), 15–22. https://doi.org/10.37868/sei.v3i1.121

Andersen, J., & Søe, S. O. (2020). Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook [Publisher: SAGE Publications Ltd]. *European Journal of Communication*, *35*(2), 126–139. https://doi.org/10.1177/0267323119894489

Avaaz. (2021). Meta-denial: How facebook fails to keep up with the evolving tactics of today's climate misinformers. https://secure.Avaaz.org/campaign/en/climate_briefing_report/

Avaaz. (2020a). White paper: Correcting the record. https://secure.avaaz.org/campaign/en/correct_the_record_study/

Avaaz. (2020b). Facebook's algorithm: A major threat to public health. https://secure.Avaaz.org/campaign/en/facebook_threat_health/

Balkin, J. M. (2018). Free Speech is a Triangle. Retrieved January 10, 2023, from https://papers.ssrn.com/abstract=3186205

Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., & Saavedra, A. (2021). Students' civic online reasoning: A national portrait. *Educational Researcher*, *50*(8), 505–515. https://doi.org/10.3102/0013189X211017495

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-028

Caulfield, T., Spring, J. M., & Sasse, M. A. (2020). Why jenny can't figure out which of these messages is a covert information operation. *Proceedings of the New Security Paradigms Workshop*, 118–128. https://doi.org/10.1145/3368860.3368870

Centre, M. B. H. (2023). About fact-checking on facebook. https://en-gb.facebook.com/business/help/2593586717571940

Chua, Y. T., Parkin, S., Edwards, M., Oliveira, D., Schiffner, S., Tyson, G., & Hutchings, A. (2019). Identifying Unintended Harms of Cybersecurity Countermeasures. *2019 APWG Symposium on Electronic Crime Research (eCrime)*, 1–15. https://doi.org/10.1109/eCrime47957.2019.9037589

Cohen, K., Johansson, F., Kaati, L., & Clausen Mork, J. (2013). Detecting Linguistic Markers for Radical Violence in Social Media. *Terrorism and Political Violence*, *26*, 2014. https://doi.org/10.1080/09546553.2014.849948

Corrigan, J., Alhabash, S., Rousu, M., & Cash, S. (2018). How much is social media worth? estimating the value of facebook by paying users to stop using it. *PLOS ONE*, *13*, e0207101. https://doi.org/10.1371/journal.pone.0207101

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society*, *20*(8), 1109–1126. https://doi.org/10.1080/1369118X.2016.1207697

FaceBook. (2020). A look at facebook and us 2020 elections. https://about.fb.com/wp-content/uploads/2020/12/US-2020-Elections-Report.pdf

Fetzer, J. H. (2004). Disinformation: The Use of False Information. *Minds and Machines*, *14*(2), 231–240. https://doi.org/10.1023/B:MIND.0000021683.28604.5b

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021). Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools [Conference Name: IEEE Access]. *IEEE Access*, *9*, 48364–48404. https://doi.org/10.1109/ACCESS.2021.3068313

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, *65*, 959–974. https://doi.org/10.1037/0022-3514.65.5.959

Gillespie, T. (2020). Content moderation, AI, and the question of scale [Publisher: SAGE Publications Ltd]. *Big Data & Society*, *7*(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research*, *6*(1), 1–12. https://doi.org/10.1186/s41235-021-00301-5

Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy*, *9*, 336–369. https://doi.org/10.5325/jinfopoli.9.2019.0336

Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the$_{d}onaldandr/incels$. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). https://doi.org/10.1145/3476057

Jackson, S. (2019). The Double-Edged Sword of Banning Extremists from Social Media. https://doi.org/10.31235/osf.io/2g7yd

Jha, H. (2022). Twitter community notes feature rolling out to all users globally: How to join. https://www.gadgets360.com/apps/news/twitter-community-notes-feature-global-rollout-fact-check-feature-birdwatch-3599604

Kajimoto, M., & Fleming, J. (2019). News literacy. *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.848

Lehmann, E. Y., & Lehmann, L. S. (2021). Responding to Patients Who Refuse to Wear Masks During the Covid-19 Pandemic. *Journal of General Internal Medicine*, *36*(9), 2814–2815. https://doi.org/10.1007/s11606-020-06323-x

Leigh, M. (2015a). Critical thinking in crisis management. *Emergency Planning College Occasional Paper*, *15*.

Leigh, M. (2015b). Critical thinking in crisis management. *Emergency Planning College Occasional Paper*, *15*.

Major, M. (2012). Objective but not impartial: Human events, barry goldwater, and the development of the “liberal media” in the conservative counter-sphere. *New Political Science*, *34*(4), 455–468. https://doi.org/10.1080/07393148.2012.729737

Mena, P. (2019). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & Internet*, *12*(2), 165–183. https://doi.org/10.1002/poi3.214

Mills, M. C., & Sivelä, J. (2021). Should spreading anti-vaccine misinformation be criminalised? *BMJ*, *372*, n272. https://doi.org/10.1136/bmj.n272

Nan, X., Wang, Y., & Thier, K. (2022). Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, *314*, 115398. https://doi.org/https://doi.org/10.1016/j.socscimed.2022.115398

Navlakha, M. (2022). Google and youtube are investing to fight misinformation. https://mashable.com/article/google-youtube-fact-checking-misinformation

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Oeldorf-Hirsch, A., & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on facebook. *Computers in Human Behavior*, *44*, 240–249. https://doi.org/10.1016/j.chb.2014.11.024

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning [The Cognitive Science of Political Thought]. *Cognition*, *188*, 39–50. https://doi.org/https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). Examining false beliefs about voter fraud in the wake of the 2020 presidential election. https://doi.org/10.31234/osf.io/szdgb

Perloff, R. M. (2009). The third person effect: A critical review and synthesis. *Media Psychology*, *1*(4), 353–378. https://doi.org/10.1207/s1532785xmep0104_4

Pham, D. V., Nguyen, G. L., Nguyen, T. N., Pham, C. V., & Nguyen, A. V. (2020). Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access*, *8*, 78879–78889. https://doi.org/10.1109/ACCESS.2020.2989140

Reed, A., & Henschke, A. (2021a). Who Should Regulate Extremist Content Online? In A. Henschke, A. Reed, S. Robbins, & S. Miller (Eds.), *Counter-Terrorism, Ethics and Technology: Emerging Challenges at the Frontiers of Counter-Terrorism* (pp. 175–198). Springer International Publishing. https://doi.org/10.1007/978-3-030-90221-6_11

Reed, A., & Henschke, A. (2021b). Who Should Regulate Extremist Content Online? In A. Henschke, A. Reed, S. Robbins, & S. Miller (Eds.), *Counter-Terrorism, Ethics and Technology: Emerging Challenges at the Frontiers of Counter-Terrorism* (pp. 175–198). Springer International Publishing. https://doi.org/10.1007/978-3-030-90221-6_11

Rich, T. S., Milden, I., & Wagner, M. T. (2020). Research note: Does the public support fact-checking social media? it depends whom and how you ask. https://misinforeview.hks.harvard.edu/

article/research-note-does-the-public-support-fact-checking-social-media-it-depends-who-and-how-you-ask/

Schultz, J. (2019). How Much Data is Created on the Internet Each Day? Retrieved January 10, 2023, from https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, *31*, 74–87. https://doi.org/10.4018/JDM.2020040105

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 1–40. https://doi.org/10.1145/3395046