



THE ETHICAL RESPONSIBILITIES OF SOCIAL MEDIA PLATFORMS IN PROMOTING UNVERIFIED INFORMATION:

A QUALITATIVE DOCUMENT ANALYSIS OF PLATFORM GOVERNANCE, TRANSPARENCY, AND FREEDOM OF EXPRESSION

1st Author Name: Li Yufang, 2nd Author: Dr. Mohd Haizra Hashim

1: Undergraduate Student, 2: Lecturer

' Faculty of Business and Communication, INTI International University,

'INTI International University, Nilai, Negeri Sembilan, Malaysia

Abstract

These are important spaces for social media that people are exposed to news, political assertions and public discussion. Still, the accuracy or invalidity of the information may spread if it passes through their recommendation system and moderation policies can impact the extent to which information is shared. This paper takes a look at the moral obligations of social media platforms on recommending or boosting unverified information. It adopts a qualitative document-analysis perspective to make sense of some of the academic literature and public guidance on information disorder, platform responsibility, content moderation, and digital-platform governance. The concept of cooperative responsibility is applied in the analysis so that the responsibility lies neither whole nor exclusively with the platform, the consumer or the public institution. Three ethical dilemmas are then covered: the balance between engaging distribution and accurate information; the need for transparency and accountability in the aspects of moderation; the possibility of misinformation regulation infringing on legitimate expression. The paper states that platforms need to take responsibility in "proportion" to the influence that they have in visibility, recommendation and enforcement, but should not be expected to ensure truth of every post made by its users. These responsibilities involve clear protocols, friction to incentivize caution against risky claims, easy ways to correct claims, informative appeal processes and independent analyses. It has provided a framework for considering how to govern platforms responsibly, although it is made clear the conceptual analysis does not provide a measure of the effectiveness of any single platform intervention.

Index Terms - social media, misinformation, platform responsibility, media ethics, content moderation, freedom of expression

1. Introduction

Social media is no longer just a place to communicate with others. They are also locations where users both find and comment on public affairs, as well as political topics, and provide information to a wide audience. Their design determines what people see, the speed of content distribution, and if disputed statements are still visible. These functions raise ethical issues when information is false or incomplete or hasn't been verified.

Distinguish between the terms used to describe problematic information. Wardle and Derakhshan [1] distinguish between misinformation, false information that is not necessarily meant to cause harm, and disinformation, false information that is deliberately created or spread to deceive. Malinformation – Information that is truthful but can be misapplied in a way that inflicts harm. The differences are significant because an ethical decision does not simply require an assessment of the veracity of a statement, but also an evaluation of intent, context, reach and potential consequences.

An important note is that unverified does not necessarily mean false. Claims can be unverified due to lack of reliable evidence, due to being new or due to the evidence being disputed. To deem every unverified claim "misinformation" would be to conflate uncertainty with falsehoods, and may lead to over-moderation. The moral issue is not, therefore, should platforms be allowed to suppress all unverifiable content but, how should platforms act when there is a possibility of harm in the context of unclear evidence to distribution.

The problem can not be solved by just asking users to be more careful. Recommender systems are created by platforms, the rules governing the distribution of content are set by platforms, how rules are enforced is determined by the platforms and they possess information regarding how content is distributed. Helberger et al. [2] state that public values in platform-

based activities must be shared responsibility between platforms, users and public institutions. UNESCO, too, sees digital-platform governance as a multi-stakeholder issue with the need to safeguard freedom of expression and access to information [3].

The three questions for research in this paper are: (1) What are the ethical issues associated with recommendation or dissemination of unverifiable information? (2) What role can cooperative responsibility play in shaping the responsibilities of platforms, users and public institutions? (3) What are the practical steps that platforms can take, without compromising freedom of expression? The paper adds a targeted ethics lens instead of an additional empirical measurement of platform behavior. It is applicable in the Media Ethics theme of Public Responsibility in Digital Communication.

2. Literature Review

2.1 Information disorder on social media

Rather than a single category of “fake news,” information disorder is a complex challenge that can be explained by the difference in relation to truth, intent, and harm between misinformation, disinformation, and malinformation. The distinction prevents a simplistic one in which each disputed statement is assumed to be deliberate deception. One user might have a legitimate claim and another user might be coordinating the claim to deceive an audience. Policies that do not take this difference into account, may react in a non-uniform way.

Aïmeur et al. [4] discuss the propagation and identification of fake news, misinformation, and disinformation on online social networks. They show that it is technically and socially difficult to spot fake content. The content may be unclear, evidence may evolve, and automated detection can lead to errors. This means that requiring a platform to be responsible for identifying every individual false claim in real-time is unrealistic and would be impossible for a platform. Rather, responsibility should include reasonable processes to address serious harbingers of risk, correct mistakes and explain enforcement decisions.

There are also platforms that serve as governance over the visibility.

2.2 Platforms as governance of visibility.

In addition to the content removal, there's ranking, recommending labelling, and restricting disseminating of posts. As Gillespie [5] describes, content moderation involves a series of decisions with results that affect what is “allowed” to be said and displayed online.

Moderation is thus not simply a technical process. The determination of what is harmful, misleading, and acceptable is value-laden and entails trade-offs.

In this study, Lee and Wei [6] discuss the ethical and responsibility issues in the design of social media based on a literature review and expert review. They recognize conflicts between values, like information sharing and safety; privacy and abuse prevention. Such tensions are pertinent to misinformation because any effort to limit the harmful circulation of information could have an impact on the circulation of legitimate information. It is thus important to think about the harm that is being done by not commenting on content, as well as the harm done by unfairly restricting content with platform governance.

2.3 Cooperative responsibility and platform governance

Helberger et al. [2] suggest cooperative responsibility to tackle the challenge of responsibility allocation in complex platform environment. Their view is based on the fact that information environment is not under the control of a single actor. Platforms determine the conditions of distribution, users produce and disseminate content, governments set the laws interpreting it and civil society, researchers and journalists participate in the public spotlight. Responsibility should be assigned based on the power, knowledge and ability of each actor to prevent harm.

There are also guidelines by UNESCO [3] for the Governance of Digital Platforms which are in line with a multi-stakeholder approach. Ensuring freedom of expression and access to information is at the heart of the platform governance guidelines despite these being important matters to tackle around disinformation and other online harms. This is extremely important to avoid pretextual application of misinformation policies as a result of over- or under-inclusive bans, application by non-facial standards or because of political pressure on platforms to suppress criticism.

Thus, the following principles are emphasised in literature for the present analysis: responsibility: this needs to be proportionate to the effect of the platform; transparency of governance and reviewability; interventions: proportionate to the risk and with respect to ensuring lawful expression. This paper's contribution is to combine these principles as a practical ethics for recommending and amplifying unverified information.

3. Theoretical Framework

This paper mashes the theory of social responsibility with the concept of cooperative responsibility. Social responsibility theory suggests that freedom of the media should go hand in hand with social responsibility obligations to serve the public good, to provide accurate information, and to be accountable for the impact of media practices (McQuail, 2010). But, on the digital space, this does not require that platforms provide the guarantee that all the information created by their users is true. Instead, public values like information quality, pluralism, accountability, and freedom of expression should be kept in mind when judging the design and governance of platforms that actively rank, recommend, and/or monetise content.

The second lens is the cooperative responsibility. Helberger et al. [2] claim that in a platform-based environment no actor is in control of the entire information environment, and therefore responsibility for public values in such environments must be shared between platforms, users, and public institutions. This lens works to prevent two extremes: viewing

platforms as neutral channels and having no responsibility for what they publish or design; and viewing platforms as sole arbiters of truth who can have unlimited discretion in removing content. Instead, duties should be commensurate with each actor's influence, knowledge and ability to prevent harm.

These lenses help to guide analysis through three normative tests: (a) responsibility—does the platform take reasonable precautions for risks posed or exacerbated by the platform? ; (b) procedural accountability—are its rules, decisions, corrections, and appeals sufficiently transparent and reviewable? ; and (c) proportionality—is the intervention proportionate to the seriousness and likelihood of harm, and to the preservation of legitimate expression? These are analytical criteria, not measurements of platform performance.

4. Methodology

Qualitative Document Analysis is utilized as a method in this paper. It does not gather new user data, undertake interviews or measure the performance of any particular platform. Rather, it reads some key, intentionally chosen academic and institutional texts that specifically target information disorder, platform governance, content moderation, and social-media ethics.

The key materials are: Wardle and Derakhshan's [1] work on information disorder; Helberger et al.'s [2] account of cooperative responsibility; Gillespie's [5] account of content moderation; Lee and Wei's [6] review of social-media ethics and responsibility, and Aïmeur et al.'s (2023) review of misinformation research and the UNESCO [3] digital platform governance guidelines. These sources were chosen because they are published with verifiable publication details and for their complementary conceptual, review and policy perspectives.

Thematic reading of documents occurred. The analysis was carried out under three headings: (a) distribution power and reliability of information; (b) transparency, accountability and procedural fairness; and (c) freedom of expression and proportionality. The claims made in the discussion are restricted to the conceptual argument and guidance in these materials. The method is not a systematic review, and does not purport to be comprehensive in coverage of publications, nor does it have a formal database screening protocol to determine frequencies of findings. This is noteworthy in that the findings of this paper are not necessarily evidence that a specific intervention is effective, but are analytical recommendations.

5. Analysis and Discussion

5.1 Distribution power creates responsibilities

The role of a platform is greater the more it contributes to the visibility of information. The choice of publishing the post is the user's choice; the choice of recommending the post to many users is the platform's choice. This separation does not imply that all posts on the platforms are by platform authors. It implies that the reach of information gains is in the context of recommendation systems.

The moral stakes rise even higher when content has a high likelihood of causing harm, and is being disseminated even when there are credible grounds that it is inaccurate. A platform could not necessarily immediately spot truth, especially if it involves a dynamically evolving event. However, uncertainty does not mean that the risk shouldn't be done. A proportionate response can take the form of a delay before a recommendation for mass dissemination of a claim, an addition of context, linking to reliable information, and/or requesting further review of a claim.

However, as Aïmeur et al. [4] have shown the detection of misinformation is still challenging. It means that platforms need to be careful about claiming automated classifications as definitive. While automated systems can be used to pick up patterns and prioritise for review, high impact decisions should be made via human processes, and there should be a path for correction. Instead of looking at individual posts after they turn into controversial, platforms should also assess the risks of their distribution design.

5.2 Transparency and accountability are moral obligations

If rules are not easily accessible, or if enforcement decisions are not explained, then it is impossible for users to effectively interpret platform governance. The information should include what kinds of content can be labelled or downranked, removed, or not included in recommendations, what evidence is taken into account, and what users can do if they disagree with a decision. That doesn't mean that platforms need to disclose all of the security-important information of their systems. It does need clear public rules and helpful explanations of decisions which impact users.

Gillespie [5] notes the social impact of moderation decisions and Lee and Wei [6] demonstrate that moderation of social media is a process of balancing competing values. These views collectively contribute to creating procedural accountability (1) users deserve clear explanations of why important enforcement measures are taken, and (2) independent researchers need safe, privacy-preserving means to assess the impact of platform policies.

Transparency also means owning up to errors. A platform should correct and explain the correction, where practical, in a timely manner if legitimate content is identified as false. A change that is not apparent to the audience who heard it originally may not be able to undo the damage done by the change itself. Similarly, if a platform is able to boost an incorrect assertion, its response should include whether the correction reaches users who viewed the original.

5.3 Freedom of expression and the risk of over-removal

A misinformation policy can inflict moral damage. Political assertions can be challenged, evidence can be incomplete, minority and oppositional voices, challenging dominant accounts can be brought up. Platforms have the power to squelch fair criticism, and restrict access to information when they remove content because it's controversial. UNESCO [3] asserts

that freedom of expression and access to information should be considered as a component of the governance issue, rather than ancillary concerns.

There will be a proportionate response encompassing deletion, withdrawal of recommendation, addition of context and issuance of a corrective notice. Removal should be done when there is obvious child misuse of material which violates the rules, but not as a standard practice for all “questionable” statements. With some exceptions, the following may be less restricted: Labels, context panels, friction before re-sharing and links to credible sources.

The appropriate measure will depend on the nature and extent of the harm and on the likelihood of the harm occurring, the quality of the evidence and the countervailing risks of abridging lawful expression.

It is also important that the principle of proportionality be respected. The identical cases should be judged in the same way, irrespective of the speaker's party affiliation and the popularity of the speaker. The enforcement information that is available to the public can be used to assess whether the rules are consistently enforced. Appeals and independent review are essential as the initial decision made by a platform can be inaccurate.

5.4 Cooperative responsibility: platforms, users and public institutions

The cooperative-responsibility approach is an alternative to two unhelpful extremes. One is the notion of platforms as being neutral pipes, with no responsibility for decisions that affect distribution. The second is to leave platforms as the gatekeepers and allow them unlimited authority to determine what can be disseminated. Helberger et al. [2] offer a more balanced perspective: responsibilities should also be shared, but not necessarily equal.

Platforms should be held accountable for their systems, policies, and enforcement of policies. Users are expected to be responsible in deciding to share information, particularly if it may cause harm to others or may be used to shape public decisions. Public institutions must create clear and rights-respecting legal protections, provide and promote media and digital literacy, and refrain from applying misinformation regulations to legitimate criticism. Independent assessment and public scrutiny is possible on the part of researchers and civil society.

There must not be any shared responsibility which obviates the liability of platforms. Platforms, as they have more technical knowledge and control over the distribution than regular users do, have a special duty to provide explanation and evaluation of their platforms. The users, however, should not be expected to verify all claims with the tools and expertise of professionals. It is necessary to have a fair framework that corresponds to the capacity and power.

5.5 Concrete steps to taking a responsible platform action

As a result of the analysis of the situation, five practical measures are suggested.

First, platforms will make it clear where pushing falsehoods ends and contestation, satire and opinion begin, as well as where things are still being developed. The classification of wetlands can reduce the inconsistencies of enforcement and ease the understanding of the expectations.

Secondly, control of distribution should be risk based in platforms. If there is a potential risk or if there is doubt, the platform may: Reduce the amount of mass sharing (after evaluation); Reduce the automatic recommendation (after evaluation); or Insert context. Such measures should be proportionate – not invisible and all the time penalties.

Thirdly, platforms should have available the possibility for correction and appeal. Mechanism(s) should be put in place to tag potentially harmful misinformation, allow provision of evidence to challenge labels and removals. Corrections should be made immediately, and should be visible to the concerned parties.

Fourthly, there need to be platform publishers of valuable transparency reports. Such reports should include details about the type of actions taken and reasons for significant changes to policy as well as detailing any appeals and their outcomes and limits in automated enforcement. Reports should not include figures on their own.

Fifthly, there is a need to encourage independent research in privacy protected environments. Research and information on effectiveness of recommendation or moderation policies should be made available to researchers, because these policies are supposedly effective. Access should be protected, but so should be personal data and legitimate security interests which should not be used as a general reason for withholding access.

5.6 Technical interventions and algorithmic governance

A more technical discussion of platform responsibility needs to look at how the recommender system works, rather than just the content-moderation determination that is made for a specific post. A typical recommendation system will rank items based on several signals, such as predicted user interest and engagement. Popularity and network signals are given a weight and if they're pretty significant, misleading content that gets a lot of interactions quickly may be given more attention. Some claim that the impact of role of algorithms is going to decide about the propagation of misinformation. This is the opinion of Pathak et al. [8] who state that not all recommender systems can be regarded as the same threat. The moral is that platforms must consider the consequences of their choices when it comes to ranking and rewarding/rewarding content, and can no longer afford to assume that they need only post about a few dubious claims in order for this to happen. In practice, governance should be effective at multiple stages of the recommendation process. First, although the detection systems might be able to determine signs that a claim might be problematic, like coordinated sharing patterns, or claims that have already been fact-checked and are generally considered to be false being shared again and again, or rapid amplification in claims during a developing event, these signs should be used as an indicator for review or a temporary precaution, not definitive proof that a claim is false. Second, a platform can use graduated interventions: context, remind people to read before they reshare, temporarily limit recommendation to new people, or pass on high-impact cases for

human review. Third, the system should allow for reassessment if new evidence arises and the restoration of distribution if the previous restriction is no longer warranted. The field-experimental study conducted by Matias [9] shows that fact checking can impact the subsequent algorithmic ranking, in a news-aggregation environment. The discovery is valuable because it shows that there is a possible interaction between human behaviour and the algorithmic distribution, but it shouldn't be automatically extended to all social-media platforms.

There is a need for governance process that could be audited in these controls. Platforms should record the risk signal used, the intervention selected, duration, who is responsible for the decision to use the intervention, and the criteria for overturning the decision. They should consider factors that may impact results over a longer period of time, such as if a measure prevents people from seeing more exposure to clearly unreliable information, if accurate corrections are able to reach the people who saw the initial information, and if there is a disproportionate downranking of legitimate reporting or minority language discussion. A relevant regulatory example is the European Union's Digital Services Act [10] that mandates platforms with recommender systems to explain the main parameters of the systems in clear language and describe options that the users have to modify or influence them. UNESCO guidance [3] additionally promotes significant transparency, independent consideration, user control, and regular evaluation of automated curation's risks. These are only governance benchmarks and not necessarily indicative of the existence of the same legal requirements in all jurisdictions. Finally, algorithmic intervention should be conducted on a proportionality ladder rather than based on a single automatic algorithm. The recommendation to remove content should be based on low confidence cases on which it is appropriate to provide contextual information or a sharing prompt; cases where credible evidence of significant harm is present, on which it is appropriate to temporarily reduce the recommendation while a review is undertaken; and on a rights-respecting policy, on which it is appropriate to remove the content, if it breaches clearly stated rules or presents a sufficiently serious risk. Every intervention should have a specified review date, an easy appeal system and measures to protect against political or commercial pressure. This does allow for the possibility that recommendation is a style of governance itself, but without the assertion that an algorithm can know truth for itself, without human judgement, without context, and without evidence.

These measures do not imply that any one measure will eradicate misinformation. These are suggestions based on the ethical principles and concerns of governance in the selected sources. The implementation of these should be checked for potential side effects such as for different languages, political groups and communities.

6. Conclusion

In the present paper, the ethics that the social media platforms have when they suggest or promote unverified information were discussed. It shows that not only should platforms be taking down glaringly obvious Deep Fakes-, but platforms have a role to play in terms of the recommendations they provide, as well as the moderation and enforcement of their policies to impact their visibility and ability to access information that people can see. Platforms at the same time may not only be regarded as truth tells.

Cooperative responsibility is a good indicator of where to place functions concerning power/capacity division. Measures of accountability should be put in place for platforms, at the point of their design of and responsible governance of distribution systems, for users, in the form of reasonable constraints when sharing potentially harmful claims, and for public institutions, in the form of providing rights-respecting conditions for accountability and media literacy. Clarity of rules, graduated responses, mechanisms of corrective measures and appeals, and different types of publicity and external opacity checks.

This paper is subject to certain limitations. It is not a systematic review or an empirical study to compare the policies of platforms; it is a qualitative conceptual analysis of selected sources. It isn't a measure of the effects of specific steps taken to combat misinformation. Future research could make use of the framework for a particular platform, or studies could be conducted on the use of label and appeal for specific enforcement actions that have been recorded elsewhere, or on the user experience with label and appeal. This research could provide an opportunity to conduct a test of two major principles- are they applied in a similar fashion and are they in the interests of protecting the reliability of information and freedom of expression?

References

- [1] C. Wardle and H. Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. Strasbourg: Council of Europe, 2017. [Online]. Available: <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>
- [2] N. Helberger, J. Pierson, and T. Poell, "Governing online platforms: From contested to cooperative responsibility," *The Information Society*, vol. 34, no. 1, pp. 1–14, 2018, doi: 10.1080/01972243.2017.1391913.
- [3] UNESCO, *Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information Through a Multistakeholder Approach*. Paris: UNESCO, 2023. [Online]. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000387339>
- [4] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Network Analysis and Mining*, vol. 13, Article 30, 2023, doi: 10.1007/s13278-023-01028-5.
- [5] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press, 2018. doi: 10.12987/9780300235029.

- [6] K. S. Lee and H. Wei, "Design factors of ethics and responsibility in social media: A systematic review of literature and expert review of guiding principles," *Journal of Media Ethics*, vol. 37, no. 3, pp. 156–178, 2022, doi: 10.1080/23736992.2022.2107524.
- [7] D. McQuail, *McQuail's Mass Communication Theory*, 6th ed. London: SAGE, 2010.
- [8] R. Pathak, F. Spezzano, and M. S. Pera, "Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks," *ACM Transactions on the Web*, vol. 17, no. 4, Article 35, 2023, doi: 10.1145/3616088.
- [9] J. N. Matias, "Influencing recommendation algorithms to reduce the spread of unreliable news by encouraging humans to fact-check articles, in a field experiment," *Scientific Reports*, vol. 13, Article 11715, 2023, doi: 10.1038/s41598-023-38277-5.
- [10] European Parliament and Council of the European Union, "Regulation (EU) 2022/2065 (Digital Services Act)," 2022. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>