Scientific Research Publishing

International Journal of Communications, Network and System Sciences, 2026, 19(10), 145-153
https://www.scirp.org/journal/ijcns
ISSN Online: 1913-3723
ISSN Print: 1913-3715

The Impact of AI on Social Media: Enhancing Content Moderation and Combating Misinformation

Andrew Mikhail Anthony[1], Hitendra Chaudhary[1], Olatunde Abiona[1,2], Clement Onime[2]

[1]Department of Computer and Mathematical Science, Indiana University Northwest, Gary, IN, USA
[2]Information & Communication Technology Section, International Center for Theoretical Physics (ICTP), Trieste, Italy
Email: andanth@iu.edu, hchaudha@iu.edu, oabiona@iu.edu, onime@ictp.it

How to cite this paper: Anthony, A.M., Chaudhary, H., Abiona, O. and Onime, C. (2026) The Impact of AI on Social Media: Enhancing Content Moderation and Combating Misinformation. *International Journal of Communications, Network and System Sciences*, 1**9**, 145-153.
https://doi.org/10.4236/ijcns.2026.1910010

Received: February 21, 2026
Accepted: October 28, 2026
Published: October 31, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Social media platforms face a persistent challenge: they must identify harmful or misleading material across immense volumes of text, images, and video while protecting legitimate expression. This paper reviews the potential and limitations of artificial intelligence (AI) in content moderation and proposes a framework that combines automated detection, human review, transparent procedures, and ongoing evaluation. Natural language processing and computer vision can help prioritize content for review, but errors, linguistic variation, bias, and rapidly changing misinformation tactics limit the reliability of automated decisions. The proposed framework routes ambiguous and consequential cases to trained moderators, provides users with explanations and appeals, and uses audits and feedback to improve performance. It is a conceptual proposal rather than an empirically validated system; future research should test its accuracy, fairness, and practical effects across languages and platforms.

Keywords

Artificial Intelligence, Social Media, Content Moderation, Misinformation, Human Oversight, Algorithmic Fairness

1. Introduction

Social media has expanded access to information and participation in public discussion. The same reach allows misleading claims and harmful content to spread quickly. Platforms must respond at a scale that human moderators alone cannot easily manage, yet removal decisions can also affect expression, privacy, and trust. Automated moderation therefore presents a governance problem as well as a tech-

nical one [1] [2].

AI can screen large volumes of posts and identify patterns that warrant review. Its outputs depend on training data, platform rules, and context, and a mistaken classification may either leave harmful material online or remove legitimate speech [1] [3] [4]. This paper examines how AI supports moderation, where it falls short, and how human review and accountability can be built into a practical workflow. It synthesizes the cited literature and proposes a design for further evaluation; it does not report a deployed system or experimental results.

2. AI and the Moderation Problem

2.1. Detection and Prioritization

Machine learning systems can classify text, images, and video, identify suspicious patterns of sharing, and rank potentially harmful posts for review [4]-[6]. Language models may help recognize policy violations in written content, while computer vision systems can identify relevant visual features. These tools can reduce the volume of material that moderators must inspect manually. Detection of a questionable claim, however, is different from establishing whether it is false: factual assessment often requires current evidence and an understanding of context.

Platform examples illustrate both the promise and the limits of automation. Large services use automated signals to help manage content, but policy enforcement still requires decisions about context, severity, and appropriate remedies [6] [7]. Rather than assume that every flagged post should be removed, a moderation system can use its confidence and the potential consequences of an error to decide which cases need human attention.

2.2. Errors and Unequal Performance

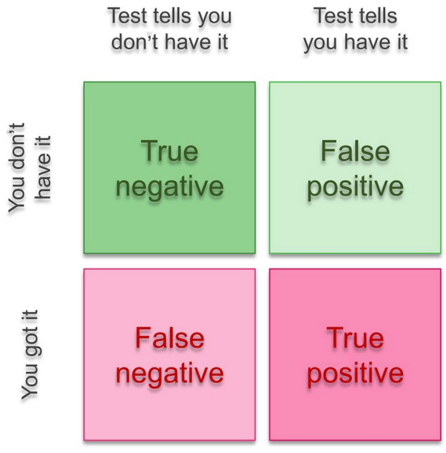


Figure 1. True and false positives and negatives (source: [9]).

A false positive occurs when permissible material is incorrectly flagged; a false negative occurs when harmful material is missed. Satire, quotation, counterspeech, and changing meanings can complicate classification. Performance may also dif-

fer across languages and communities when training data and evaluation sets do not adequately represent them [3] [4] [8]. Reporting only overall accuracy can obscure these differences. Evaluation should include precision, recall, appeal outcomes, and error rates for relevant language and content groups.

Figure 1 maps moderation outcomes by comparing a system's decision with whether content actually violates policy. It highlights false positives and false negatives, the two ways a classifier can make an incorrect decision.

Figure 2 outlines a staged process for addressing algorithmic bias: test and improve data in controlled settings, assess impacts and values, and monitor the system after rollout. Feedback sends failures back for further review.

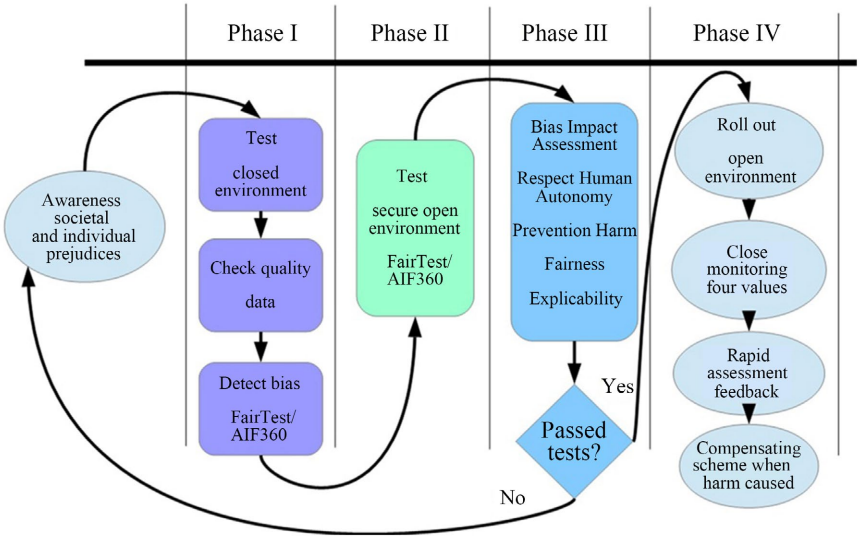


Figure 2. Algorithmic bias (source: [10]).

Adversaries may also alter wording, images, or patterns of distribution to evade detection. Models must be evaluated against emerging tactics without assuming that frequent retraining alone will solve the underlying problems [11]. Research on adversarial attacks in other AI domains points to the broader vulnerability, although results from those domains cannot be transferred directly to social media moderation [12].

2.3. Rights and Accountability

Moderation decisions affect users' ability to participate in public discussions. Platforms should therefore state the rules they enforce, communicate the reasons for significant actions, and offer a meaningful way to challenge mistakes [13] [14]. Data collection for model training and behavioral analysis also raises privacy questions. A responsible system limits access to personal information, documents its purposes, and evaluates whether each data use is necessary. The precise legal obligations depend on jurisdiction and platform context.

Figure 3 places accountability and transparency at the center of linked system modules, including prompts, tools, applications, communications, memory,

knowledge, and data. The diagram shows how oversight can connect these parts of an AI system.

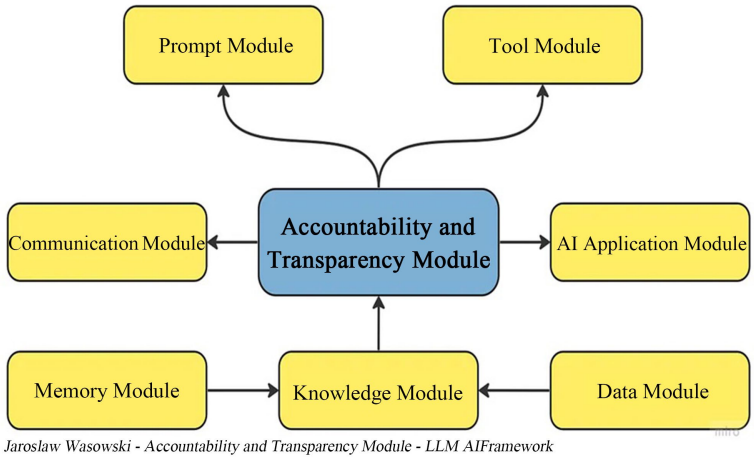


Figure 3. Accountability and transparency (source: [15]).

3. Proposed Human and AI Moderation Framework

The proposed framework has four connected components: automated triage, human review, accountable procedures, and continuous evaluation. It is designed to assign routine screening to AI while reserving contextual and consequential judgments for people [11] [16] [17]. The framework specifies an approach for testing rather than claiming that these components already achieve particular outcomes.

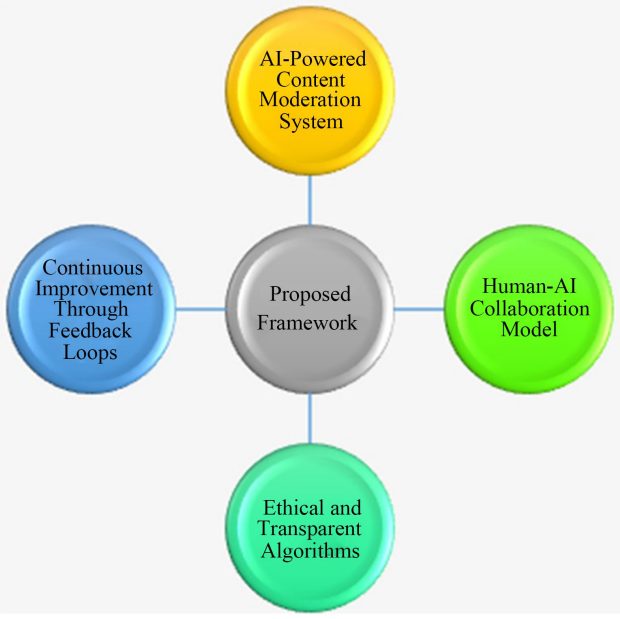


Figure 4. Proposed framework.

Figure 4 summarizes the proposed framework's four pillars: AI-powered moderation, human-AI collaboration, ethical and transparent algorithms, and contin-

uous improvement. Together, they frame automated screening as one part of a broader oversight process.

Figure 5 shows how a prompt and input data guide an LLM response, which an AI moderator checks for accuracy, risks, and policy compliance. Human reviewers add domain, ethics, security, and compliance judgment, with final approval when needed. The diagram depicts review of an LLM response, not the full social media moderation workflow for user posts or misinformation verification.

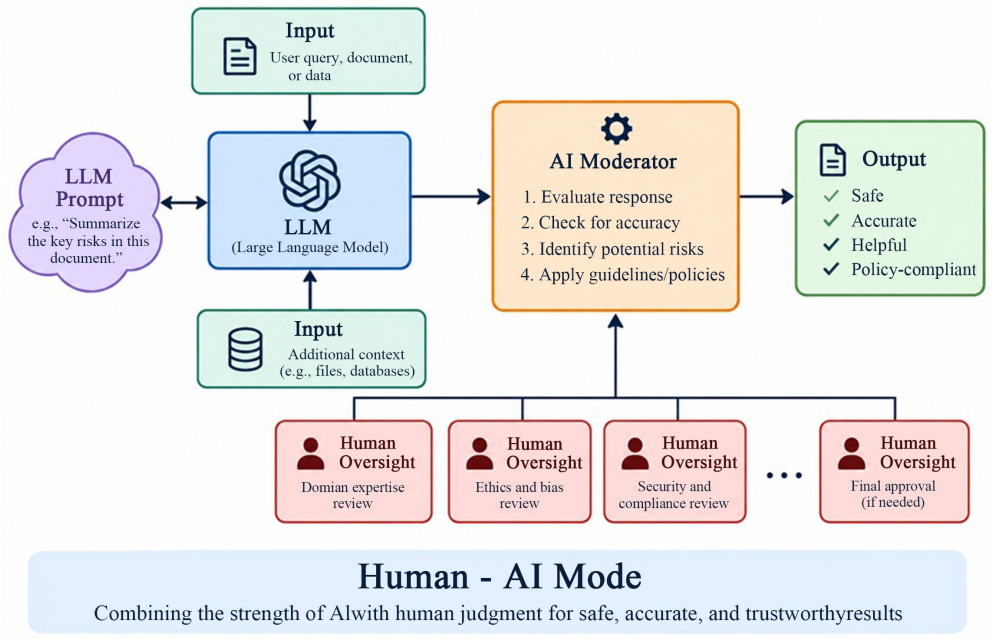


Figure 5. Human and AI oversight model.

3.1. Automated Triage

An AI system first examines content for potential violations or misleading claims and records the relevant policy category and confidence estimate. Text and visual models can contribute to separate signals. A flagged post then moves through a tiered process: low-risk, well-supported cases may receive an appropriate automated action under documented rules; uncertain, high-impact, or disputed cases go to human reviewers. Any use of external evidence for fact checking must identify the source and account for how quickly information may change [5] [18].

3.2. Human Review and Appeals

Moderators examine the original post, surrounding conversation, relevant cultural or linguistic context, and applicable policy. They may confirm, reverse, or modify a proposed action. Training should address platform policy, bias, emerging misinformation tactics, and the limits of AI-generated explanations. Users should receive a clear reason for an action and an accessible appeal process. Appeals provide a safeguard for individuals and a source of information about recurring errors [16] [17].

3.3. Oversight and Transparency

Platforms should document model changes, review criteria, escalation paths, and the people accountable for decisions. Regular audits can compare outcomes across content types and language groups, examine patterns of overturned decisions, and identify whether the system applies rules consistently. Explanations should describe the policy basis for a decision in terms users can understand; technical explanations alone do not guarantee a fair outcome [10] [19] [20]. Independent or cross-functional oversight can review findings and recommend policy changes.

3.4. Feedback and Evaluation

A feedback loop should connect moderator decisions, appeals, audits, and updated evaluation datasets. Before a model is retrained, corrected examples need quality checks so that earlier mistakes do not become new training labels. Measures should include detection performance, time to review, appeal and reversal rates, differences between groups, and user perceptions of the process. Platforms should assess these measures over time and report material changes with appropriate privacy protections [8] [11].

Figure 6 presents the feedback cycle as five steps: gather data, analyze it, generate insights, implement changes, and evaluate results. The evaluation then informs the next round of data gathering.

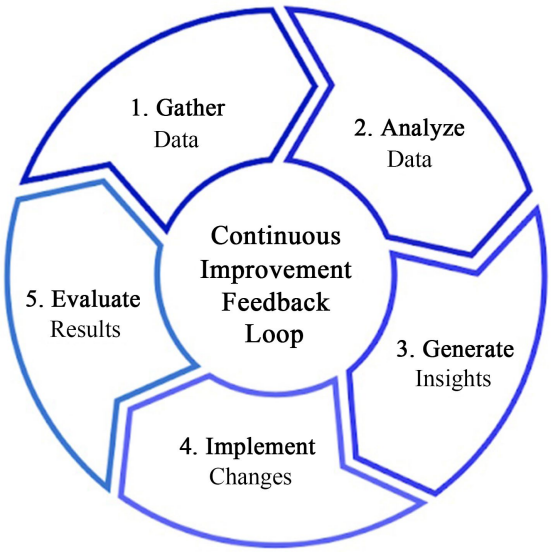


Figure 6. Continuous improvement through feedback loops (source: [21]).

4. Implementation and Research Agenda

Implementation can begin with a limited pilot on a defined policy category and a representative sample of content. Baseline performance should be measured before the AI workflow is introduced. Reviewers can then test routing thresholds, evaluate errors in multiple languages, and examine how often correct decisions appeal. Expansion should depend on evidence that the workflow meets predefined

accuracy, fairness, and response-time targets, as well as applicable privacy and legal requirements [14] [22]-[24].

A simulated public health misinformation scenario could test the framework without presenting hypothetical outcomes as findings. Researchers could assemble time-stamped claims and verified sources, compare automated and human assessments, and measure whether the system adapts when evidence changes. Additional studies should examine moderator workload, the effect of explanations on trust, and the trade-offs between rapid intervention and erroneous removal. Cross-platform and cross-cultural evaluation would help establish where the proposed design is transferable.

5. Conclusions

AI can help social media platforms find and prioritize content that may violate policy or spread misinformation. Its speed does not resolve the difficult questions of truth, context, fairness, or the rights of users. Errors and unequal performance make human review especially important when decisions are uncertain or consequential.

This paper proposes a moderation workflow that combines automated triage with trained reviewers, explanations, appeals, audits, and measured updates. The framework offers a basis for implementation and testing, but its benefits remain to be demonstrated. Platforms that evaluate both technical performance and the effects on users will be better equipped to improve moderation while remaining accountable for their decisions.

Author Contributions

Conceptualization, A.M.A, H.C. and O.A; methodology, A.M.A., H.C., O.A, and C.O; formal analysis, A.M.A., H.C., O.A, and C.O; investigation, A.M.A., H.C. Resources, A.M.A., H.C. Data curation, A.M.A., H.C.; writing—original draft preparation, A.M.A. , H.C. writing—review and editing, O.A., C.O. Visualization, A.M.A., H.C. supervision, O.A., C.O.; project administration, O.A., C.O. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Marsoof, A., Luco, A., Tan, H. and Joty, S. (2022) Content-Filtering AI Systems-Limitations, Challenges and Regulatory Approaches. *Information & Communications Technology Law*, **32**, 64-101. https://doi.org/10.1080/13600834.2022.2078395

[2] Seering, J. (2020) Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, **4**, 1-28. https://doi.org/10.1145/3415178

[3] Nishant, R., Schneckenberg, D. and Ravishankar, M. (2023) The Formal Rationality of Artificial Intelligence-Based Algorithms and the Problem of Bias. *Journal of Infor-*

mation Technology, **39**, 19-40. https://doi.org/10.1177/02683962231176842

[4] Gongane, V.U., Munot, M.V. and Anuse, A.D. (2022) Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions. *Social Network Analysis and Mining*, **12**, Article No. 129. https://doi.org/10.1007/s13278-022-00951-3

[5] Mishra, S., Shukla, P. and Agarwal, R. (2022) Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets. *Wireless Communications and Mobile Computing*, **2022**, Article ID: 1575365. https://doi.org/10.1155/2022/1575365

[6] Bokolo, B.G. and Liu, Q. (2024) Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis. *Electronics*, **13**, Article 1671. https://doi.org/10.3390/electronics13091671

[7] Tong, J. (2022) Artificial Intelligence Is Not a Panacea: Policing Content on Social Media Platforms, Three Dilemmas and Their Ethical and Legal Implications. In: Pihlajarinne, T. and Alén-Savikko, A., Eds., *Artificial Intelligence and the Media*, Edward Elgar Publishing, 123-147. https://doi.org/10.4337/9781839109973.00012

[8] Chanda, S.S. and Banerjee, D.N. (2022) Omission and Commission Errors Underlying AI Failures. *AI & Society*, **39**, 937-960. https://doi.org/10.1007/s00146-022-01585-x

[9] Mozina, M. (2024) Which Is Worse: False-Positive or False-Negative? https://www.linkedin.com/pulse/which-worse-false-positive-false-negative-miha-mozina-phd

[10] Belenguer, L. (2022) AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry. *AI and Ethics*, **2**, 771-787. https://doi.org/10.1007/s43681-022-00138-8

[11] Gorwa, R., Binns, R. and Katzenbach, C. (2020) Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society*, **7**. https://doi.org/10.1177/2053951719897945

[12] Jain, A., Sangeeta, K., Swathi, B., Sadim, M., Dwivedi, S.P. and Albawi, A. (2024) The Impact of Adversarial Attacks on Medical Imaging AI Systems. 2024 *IEEE* 13*th International Conference on Communication Systems and Network Technologies* (*CSNT*), Jabalpur, 6-7 April 2024, 362-367. https://doi.org/10.1109/csnt60213.2024.10545943

[13] Koulu, R. (2020) Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy. *Maastricht Journal of European and Comparative Law*, **27**, 720-735. https://doi.org/10.1177/1023263x20978649

[14] Leerssen, P. (2023) An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation. *Computer Law & Security Review*, **48**, Article ID: 105790. https://doi.org/10.1016/j.clsr.2023.105790

[15] Nayak, S. (2022) Harnessing Explainable AI (XAI) for Transparency in Credit Scoring and Risk Management in FinTech. *International Journal of Applied Engineering & Technology*, **4**, 214-236. https://romanpub.com/resources/Vol%204%20%2C%20No%202%20-%2030%20%20%281%29.pdf?utm_source=chatgpt.com

[16] Lai, V., Carton, S., Bhatnagar, R., Liao, Q.V., Zhang, Y. and Tan, C. (2022) Human-ai Collaboration via Conditional Delegation: A Case Study of Content Moderation. *CHI Conference on Human Factors in Computing Systems*, New Orleans, 29 April 2022-5 May 2022, 1-18. https://doi.org/10.1145/3491102.3501999

[17] Kuo, T., Hernani, A. and Grossklags, J. (2023) The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools. *Proceedings of the ACM on Human-Computer Interaction*, **7**, 1-38. https://doi.org/10.1145/3579530

[18] Kumar, Y., Huang, K., Perez, A., Yang, G., Li, J.J., Morreale, P., *et al.* (2024) Bias and Cyberbullying Detection and Data Generation Using Transformer Artificial Intelligence Models and Top Large Language Models. *Electronics*, **13**, Article 3431. https://doi.org/10.3390/electronics13173431

[19] Naveed, S., Stevens, G. and Robin-Kern, D. (2024) An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI. *Applied Sciences*, **14**, Article ID: 11288. https://doi.org/10.3390/app142311288

[20] Habbal, A., Ali, M.K. and Abuzaraida, M.A. (2024) Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions. *Expert Systems with Applications*, **240**, Article ID: 122442. https://doi.org/10.1016/j.eswa.2023.122442

[21] O3SIGMA (2023) Continuous Improvement Feedback Loop with Smart Manufacturing. https://o3sigma.com/articles/continuous-improvement-feedback-loop-with-smart-manufacturing

[22] Roberts, M.K., Fisher, D.M., Parker, L.E., Darnell, D., Sugarman, J., Carrithers, J., *et al.* (2020) Ethical and Regulatory Concerns in Pragmatic Clinical Trial Monitoring and Oversight. *Ethics & Human Research*, **42**, 29-37. https://doi.org/10.1002/eahr.500066

[23] Kasirzadeh, A. and Gabriel, I. (2023) In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology*, **36**, Article No. 27. https://doi.org/10.1007/s13347-023-00606-x

[24] Xi, Y. (2024) Navigating Technological Shifts: An Examination of User Inertia and Technology Prestige in Large-Language-Model AI Chatbot Transition. *International Journal of Human-Computer Interaction*, **41**, 8167-8183. https://doi.org/10.1080/10447318.2024.2405786