CITACONOMIA : Economic and Business Studies

Volume 5 No.2
ISSN: 2963-492x
The article is published with OpenAccess at:
https://ejournal.citakonsultindo.or.id/index.php/CITACONOMIA/indexdex

THE PARADOX OF DIGITAL ACCOUNTABILITY: A CRITICAL ANALYSIS OF META'S ETHICAL OUTSOURCING IN CONTENT MODERATION

Alya Salsabila Zahra[1], Tsabitah Lintang Permata[2], Desy Dwi Nugrahaini[3], Adam Farid Hidayat[4], Pramudya Ibnu Subrata[5]
[1,2,3.4.5]Universitas Slamet Riyadi, Surakarta
Email: alyazahra2207@gmail.com

Abstrak: *This conceptual article examines the ethical crisis in the global social media ecosystem following Meta's 2025 policy shift, which replaced human fact-checkers with AI-driven Llama automation and "Community Notes." Using a qualitative descriptive method and the TARES test framework, the study analyzes the phenomenon of "ethical outsourcing"—the delegation of moral responsibility from professional curation to unverified algorithms and crowdsourced labor. Findings reveal a profound paradox where operational efficiency compromises informational integrity, as AI systems frequently fail to grasp cultural nuances and linguistic complexities, leading to the systematic verification of disinformation. Furthermore, the reliance on community-based moderation introduces risks of collective bias and manipulation by organized interest groups, violating the principles of equity and social responsibility. This research offers a theoretical contribution by redefining digital transparency as "radical honesty" concerning technological limitations. It argues for the restoration of the "Human-in-the-Loop" model, repositioning public relations practitioners as "ethical guardians" who provide the necessary moral oversight to automated systems. Ultimately, the study concludes that restoring public trust requires a structural reintegration of human conscience into the algorithmic decision-making process to ensure that digital public spheres remain accountable and humanity-oriented.*

Keyword: *Algorithmic Accountability, Community Notes, Digital Ethics, Ethical Outsourcing, Meta, Public Relations, TARES Test.*

INTRODUCTION

The current global social media ecosystem is facing a critical digital ethics crisis, characterized by systemic failures in mitigating the massive spread of disinformation. This phenomenon is no longer viewed merely as a technical anomaly but is understood as a fundamental erosion of the social responsibility inherent in communication corporations (Lattimore et al., 2021). Public safety and informational integrity are frequently compromised in favor of high user engagement driven by non-transparent algorithms. Consequently, public trust in major digital platforms, particularly Meta, has reportedly declined significantly due to the system's inability to distinguish between subjective opinions and verifiable objective facts (Ward, 2024). International regulators are increasingly applying pressure, taking decisive action against platforms perceived to allow false narratives to flourish without adequate human oversight.

Corporate ethical responsibility in safeguarding the digital public sphere is

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

increasingly questioned when technological efficiency is prioritized over informational accuracy. Fundamental public relations ethical principles, such as honesty and accountability, are often found neglected in content moderation practices relying on full automation (Parsons, 2023). This situation was exacerbated by Meta's official policy in early 2025, which reduced the involvement of human fact-checkers in favor of AI-driven systems and Community Notes (The Guardian, 2025). This shift is analyzed as an attempt to delegate moral responsibility to users and machines, a concept described as "ethical outsourcing." Experts warn that the sociopolitical impact of this policy poses a tangible threat to global information stability (Seitel, 2024).

Meta's radical transformation in content moderation is marked by the extensive termination of partnerships with expert third-party fact-checkers. The role of human curation, previously a pillar of information validation, has been replaced by Llama-based AI automation and the Community Notes feature (The Guardian, 2025). While management justifies this move as a means to increase response speed for massive data volumes, academics critique it as a transfer of verification accountability from the organization to algorithms and general users (Ward, 2024). The full-scale implementation of AI is feared to overlook contextual nuances and cultural sensitivities that can only be grasped through human intuition and moral judgment (Parsons, 2023). Thus, this strategic shift represents a crucial turning point where public accountability is traded for operational efficiency.

A significant literature gap exists regarding the absence of human conscience in ethical decision-making processes within Meta's environment. While previous studies have extensively evaluated algorithmic effectiveness in detecting harmful content, the intuitive moral dimension is often neglected in academic discourse (Ward, 2024). The role of professional communication practitioners, who should function as the "organizational conscience," appears increasingly marginalized by rigid automated systems (Parsons, 2023). This phenomenon highlights a research gap concerning how human values can be preserved when ethical control mechanisms are surrendered to artificial intelligence. Therefore, exploring the psychological and sociological impacts of the absence of human moral consideration in content moderation is essential.

Furthermore, there is a lack of evaluation regarding social responsibility delegated to the masses through community features. User-based oversight has primarily been studied through the lens of digital participation without considering the risks of collective bias that may damage information integrity (Lattimore et al., 2021). The absence of a centralized and competent moral authority within Meta's public relations structure creates uncertainty regarding global ethical standards (Seitel, 2024). This study aims to fill these gaps by

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

offering a new perspective on the importance of human elements in maintaining objectivity and fairness. Through this conceptual approach, a more comprehensive understanding of the ethical boundaries of technology in the public sphere can be achieved (Lattimore et al., 2021; The Guardian, 2025; Parsons, 2023).

THEORETICAL STUDY AND HYPOTHESIS DEVELOPMENT

A. The Principle of Truthfulness in Digital Communication

The first dimension of this ethical evaluation framework focuses on the principle of truthfulness regarding information disseminated to the public via digital channels. In the context of Meta's policies, truthfulness is measured not merely by the absence of falsehoods but by a sincere intent to provide complete and accurate facts (Baker & Martinson, 2001). However, the practice of spreading information through imperfectly tested algorithms often blurs the boundaries between verifiable facts and subjective user opinions. Consequently, the integrity of corporate communication relies heavily on the extent to which informational truth can be accounted for without intentional distortions made for technical gain (Parsons, 2023).

Furthermore, the ethical weight of truthfulness in public relations demands a commitment to transparency that transcends algorithmic efficiency. When platforms prioritize engagement metrics over factual precision, they risk violating the fundamental trust established with the global audience. As such, the alignment of automated content delivery with the objective reality of information remains a critical benchmark for evaluating the ethical performance of social media giants (Baker & Martinson, 2001; Parsons, 2023).

B. Authenticity and Sincerity in Automated Moderation

The second criterion, authenticity, emphasizes the sincerity of an organization in consistently fulfilling its communicative responsibilities. This principle requires a clear alignment between the values promised by the corporation and the actual actions implemented within its content moderation features (Lattimore et al., 2021). However, the deployment of the Llama AI automation system in early 2025 has reportedly produced artificial responses that fail to reflect human empathy (The Guardian, 2025). As a result, Meta's authenticity as a provider of public dialogue is increasingly questioned, as its ethical decisions are perceived as mechanical rather than grounded in genuine moral conscience.

This perceived lack of sincerity suggests a disconnect between corporate branding and operational reality. Authenticity in the digital age requires that automated systems do not function as a shield for corporate indifference but as an extension of the organization's

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

core ethical identity. Without the presence of human-like discernment and empathy, the transition to fully AI-driven moderation risks being viewed as a performative gesture rather than a substantive commitment to ethical communication (Lattimore et al., 2021; The Guardian, 2025).

C. Respect for Human Dignity and User Rights

The third principle, respect, is directed toward the dignity of users as the primary audience of information. Respect for human rights in the digital sphere is manifested through the protection of user privacy and safeguarding the community from harmful content (Seitel, 2024). Nevertheless, Meta's policy of reducing human moderation staff is viewed as a disregard for the users' right to a safe and dignified communication environment. This shift threatens to damage the relationship of trust that should be built through respectful, two-way communication between the platform and its global community.

Moreover, respect involves acknowledging the vulnerability of users in an algorithmically curated environment. By delegating the duty of care to automated processes, a platform may inadvertently expose its audience to psychological or social harm. Ethical public relations must, therefore, ensure that technological advancements do not compromise the fundamental dignity of the individuals they serve, maintaining a human-centric approach to digital safety (Seitel, 2024).

D. Equity and Algorithmic Justice

The fourth dimension, equity, demands equal access and fair treatment for all segments of society without discrimination. In the social media ecosystem, ethical equity is tested by the system's ability to remain impartial and prevent algorithmic biases from disadvantaging specific groups (Ward, 2024). However, the failure of features like Community Notes to identify coordinated information manipulation campaigns often harms minority groups who lack a significant power base on the platform. Therefore, the balance of informational power must be maintained to prevent the domination of public opinion by those with superior resources.

The pursuit of algorithmic justice requires proactive measures to identify and mitigate systemic biases inherent in automated code. Equity is not merely a passive state but an active obligation to ensure that the "digital commons" remains inclusive. When platforms fail to protect marginalized voices from algorithmic silencing or targeted disinformation, they breach the ethical requirement of fairness that is central to modern public communication (Ward, 2024).

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

E. Social Responsibility and Global Stability

The final criterion is social responsibility, which encompasses the long-term impact of corporate communication on the common good. Large communication firms like Meta hold a moral obligation to contribute positively to social stability and democratic integrity (Parsons, 2023). Yet, the priority given to user engagement growth over the quality of public information is considered a violation of the social contract between the corporation and society. Effective digital public relations ethics can only be achieved if all technological policies are directed toward supporting broad and sustainable public welfare.

This social mandate implies that corporations are stewards of the public discourse, not just neutral service providers. The ethical repercussions of their policy choices resonate far beyond corporate balance sheets, influencing the health of global democracies. Consequently, social responsibility must serve as the overarching framework within which all other ethical dimensions operate, ensuring that innovation remains subservient to the needs of the collective society (Parsons, 2023; Seitel, 2024).

F. Algorithmic Accountability and Moral Responsibility

Algorithmic accountability in the digital public sphere is understood as the moral and legal obligation of technology providers to answer for the social impacts produced by their programming. This responsibility is no longer limited to functional technicalities but includes the ethical consequences that arise when algorithms act as arbiters of truth (Diakopoulos, 2020). In modern public relations, accountability is often distorted when companies hide behind the complexity of technology that is difficult for the general public to comprehend (Parsons, 2023). Thus, any automated system designed to manage public narratives must be integrated with transparency principles to ensure systemic biases are openly identified and corrected.

The implementation of Llama AI by Meta in early 2025 serves as a case where algorithmic accountability is being intensely tested by global digital ethics watchdogs. The system's failure to recognize local cultural contexts has triggered the spread of harmful content that could have been prevented through competent human oversight (The Guardian, 2025). Accountability is often lost when companies claim that moderation decisions are made independently by machines without direct professional intervention, creating a "moral vacuum" where no party is explicitly responsible for systemic disinformation (Ward, 2024). Therefore, strengthening ethical standards in code development is imperative to ensure that technology remains humanity-oriented rather than merely focused on operational efficiency.

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

RESEARCH METODOLOGY

The research design for this study employs a qualitative conceptual approach, focusing on the intersection of digital ethics and algorithmic accountability within the social media ecosystem. This method involves a critical analysis of existing frameworks, such as the TARES test, to evaluate the ethical dimensions of truthfulness, authenticity, and respect in Meta's current communication strategies (Baker & Martinson, 2001). By examining foundational public relations principles alongside contemporary digital ethics, the study identifies systemic failures in automated content moderation that compromise informational integrity (Lattimore et al., 2021; Parsons, 2023). This approach allows for a deep exploration of how corporate "ethical outsourcing" to AI affects the social contract between platforms and the global public.

Data collection is based on a comprehensive literature review of academic journals and contemporary case reports regarding Meta's 2025 policy shifts. The analysis integrates multidisciplinary reviews on algorithmic accountability to understand the technical and moral responsibilities of technology providers (Diakopoulos, 2020). Specific attention is given to the real-world implications of replacing human fact-checkers with Llama-based AI systems and the Community Notes feature, as documented in recent journalistic accounts (The Guardian, 2025). This secondary data provides the necessary evidence to critique the transition from human-centric verification to automated processes that lack cultural and contextual nuance.

The analytical framework utilizes a critical-evaluative lens to assess the erosion of the "organizational conscience" in the digital age. By synthesizing current theories on algorithmic responsibility, the study evaluates the sociopolitical impacts of Meta's strategic shift on global information stability (Ward, 2024). The methodology further explores the gaps in social responsibility delegated to user-based oversight, referencing established public relations practices to highlight the marginalized role of human moral judgment (Seitel, 2024). Ultimately, this methodology seeks to offer a new perspective on maintaining objectivity and fairness in a public sphere increasingly governed by non-transparent algorithms (Parsons, 2023; Ward, 2024).

RESEARCH RESULT

A. The AI Paradox: Operational Efficiency vs. Ethical Integrity

The implementation of Meta's Llama artificial intelligence in early 2025 has highlighted a growing paradox between operational efficiency and ethical integrity. While the technology excels in processing massive data volumes at high speeds, this capability is frequently used to justify the minimization of human curation (The Guardian, 2025).

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

However, findings indicate that the system consistently fails to comprehend complex cultural contexts and linguistic nuances, such as sarcasm and satire. This technical limitation has led to an ironic phenomenon where hoaxes are systematically "verified" by algorithms. Such inability to perform deep moral judgment poses a severe threat to the quality of public information, as technical efficiency ultimately sacrifices the objective truth that serves as the foundation of public relations ethics (Baker & Martinson, 2001).

This ethical dilemma is further complicated by the opaque nature of automated decision-making, which obscures public accountability. Although response times to harmful content may improve, the validity of these automated decisions often falls below the accuracy standards established by human experts (Ward, 2024). The neglect of linguistic complexity and social sensitivity within Meta's algorithms reflects a corporate priority leaning toward operational cost reduction rather than the protection of the information ecosystem (Parsons, 2023). Consequently, this AI paradox demonstrates that efficiency without the oversight of human conscience results in a mechanization of communication that disregards human values and corporate social responsibility (Diakopoulos, 2020).

B. The Dynamics of Digital Democratization and Ethical Outsourcing

1). Challenges of Community-Based Moderation

Meta’s shift toward mass-based moderation through the Community Notes feature is framed as an effort toward digital democratization. Within the framework of dialogic communication, public participation in verifying truth is seen as a progressive step toward reducing the dominance of a single authority over public narratives (Lattimore et al., 2021). However, the effectiveness of this system depends heavily on the integrity and objectivity of the contributing users. The success of this democratization is questionable when the curation of information is surrendered to a crowd that lacks clear journalistic competence or professional ethical standards (Ward, 2024). As a result, the potential for informational distortion remains high, as truth-assessment mechanisms are based on majority consensus rather than objective facts.

2). Risks of Collective Bias and Systematic Deception

The vulnerability of Community Notes to manipulation by "cyber armies" represents a significant threat that can transform a clarification tool into an instrument of disinformation. Findings show that large user collectives can be exploited to suppress specific narratives or spread propaganda through organized systemic voting **(**The Guardian, 2025). This phenomenon is a form of deception where truth is falsified through the

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

impression of artificial public support. Such practices violate the principle of Equity in the TARES Test, as they allow parties with superior technological resources to dominate public opinion (Baker & Martinson, 2001). Relying on community moderation without stringent expert oversight risks exacerbating polarization and damaging the integrity of global public relations ethics (Parsons, 2023).

3). Conceptualizing Ethical Outsourcing

Meta's communication strategy, which appears to empower users through active participation, is critically analyzed as a form of "ethical outsourcing." This narrative of public empowerment is often used as a justification for the organization to divest itself of moral responsibility or liability when widespread information crises occur (The Guardian, 2025). Conceptually, this phenomenon represents a corporate attempt to shift the burden of high-risk ethical decision-making from professional practitioners to algorithms and communities that lack legal accountability (Parsons, 2023). Consequently, corporate public relations integrity is jeopardized by the absence of a clear moral authority to bear the consequences of systemic failures (Ward, 2024). This divestment of responsibility constitutes a breach of the Social Responsibility principle within the TARES framework, which mandates that large communication organizations must remain the vanguard of public truth rather than mere technological facilitators (Baker & Martinson, 2001).

CONCLUSION

In conclusion, the strategic shift in Meta's content moderation policies toward full-scale AI automation and community-led verification marks a significant departure from established ethical communication standards. The "AI Paradox" reveals that while technological efficiency allows for the management of vast data volumes, it fundamentally lacks the human moral judgment necessary to preserve truthfulness and authenticity (Baker & Martinson, 2001; The Guardian, 2025). This study finds that replacing human experts with algorithms like Llama often results in a "moral vacuum," where the nuanced understanding of cultural contexts and social sensitivities is sacrificed for operational speed. Consequently, the integrity of the digital public sphere is compromised, leading to a decline in public trust and a violation of the core principles of honesty and accountability that should define professional public relations (Parsons, 2023; Seitel, 2024).

Furthermore, the delegation of moderation duties to the general public through features like Community Notes, while framed as democratization, introduces significant risks of collective bias and systematic deception. This "ethical outsourcing" allows corporations to divest themselves of direct responsibility, shifting the burden of truth-

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.

seeking onto users who may lack professional ethical competence or be susceptible to manipulation by organized interest groups (Lattimore et al., 2021; Ward, 2024). Such practices directly contradict the principles of equity and social responsibility within the TARES framework, as they empower those with superior technological resources to dominate public discourse. Without centralized and competent moral authority, these systems facilitate the spread of disinformation under the guise of public participation (Baker & Martinson, 2001; Diakopoulos, 2020).

Ultimately, this study emphasizes that algorithmic accountability must be re-integrated with human oversight to safeguard global information stability. Technology should serve as a tool for humanity rather than a replacement for the "organizational conscience" provided by professional communication practitioners (Parsons, 2023). As digital platforms continue to evolve, it is imperative for international regulators and corporate leaders to prioritize ethical integrity over mere technical efficiency. To ensure a safe and dignified digital environment, the preservation of human-centric verification processes remains essential in maintaining the delicate balance between innovation and social responsibility (Seitel, 2024; Ward, 2024).

REFERENCE

Baker, S., & Martinson, D. L. (2001). The TARES test: Five principles for ethical persuasion. *Journal of Mass Media Ethics*, *16*(2-3), 148–175. https://doi.org/10.1207/S15327728JMME1602&3_6

Diakopoulos, N. (2020). Accountability in algorithmic decision making: A multi-disciplinary review. Journal of Business Ethics, 167(2), 231–245. https://doi.org/10.1007/s10551-019-04213-z

Lattimore, D., Baskin, O., Heiman, S. T., & Toth, E. L. (2021). Public relations: The profession and the practice (5th ed.). McGraw Hill Education.

Meta faces backlash over withdrawal of human fact checkers in favor of AI driven Community Notes. (2025, January 15). The Guardian. https://www.theguardian.com/technology/2025/jan/15/meta-ai-moderation-human-fact-checkers-backlash

Parsons, P. J. (2023). Ethics in public relations: A guide to best practice (5th ed.). Kogan Page.

Seitel, J. P. (2024). The practice of public relations (15th ed.). Pearson.

Ward, S. J. A. (2024). Digital ethics and the future of public communication: A critical analysis of algorithmic responsibility. Journal of Media Ethics, 39(1), 12–28. https://doi.org/10.1080/23736992.2023.2285141

Copyright at authors some right reserved this work is licensedUnder a Creative Commons Attribuation-ShareAlike 4.0 International License.