Proceedings of ICIHCS 2026 Symposium: Shaping the Future of Media: Audience-Centric Innovation, AI & Digital Transformation
DOI: 10.54254/2753-7064/2026.BA38619

Ethical Controversies in AI-Driven Risk Control and Content Moderation on Social Platforms—A Case Study of Xiaohongshu

Chen Gong

Faculty of Arts and Social Sciences, The University of Sydney, Sydney, Australia
18080459418@163.com

Abstract. The use of artificial intelligence (AI) to identify and manage content risks on social platforms is growing. It raises ethical questions about uncertain classifications, proportionate intervention and opportunities to challenge decisions. This paper examines Xiaohongshu through a qualitative documentary case study comprising a technical publication, a regulatory document and three original news reports. The analysis distinguishes source identification, harmfulness assessment and moderation action, and examines the relationship between internal explanations and user-facing remedies. The selected documents separately describe evidence for reviewers, information for audiences and appeal access, leaving their integration unevaluated. This study argues that technical explanations alone do not establish user-facing procedural protection, understandable reasons and accessible correction are also necessary. Source labels can provide transparency without establishing harm, while stronger restrictions require an additional justification appropriate to their consequences. Differentiated review, bounded disclosure and traceable appeals are proposed as design requirements whose effectiveness remains to be evaluated. The contribution emphasizes linking content-safety objectives with user protections in platform governance.

Keywords: artificial intelligence, risk control, content moderation, platform ethics, Xiaohongshu

1. Introduction

Content moderation turns platform policies into decisions about visibility, participation and permissible expression. Automated tools can support classification, but judgements about context and acceptable conduct remain necessary. Gorwa, Binns and Katzenbach identify technical and political challenges to treating automation as a purely operational solution [1]. Gillespie questions the assumption that increasing scale makes artificial intelligence (AI) an adequate response to moderation problems [2]. Efficiency must therefore be considered alongside the reasons for intervention and the possibility of correcting decisions.

Xiaohongshu provides a relevant setting because lifestyle expression, product recommendations and creator activity intersect with expectations of authenticity. Ju's study of Hong Kong influencers

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.BA38619

documents negotiations over platform rules and content visibility, including uncertainty about the boundary of wealth flaunting [3].

These historical accounts motivate attention to rule interpretation. However, they do not establish how later AI systems operate. Thus, this paper aims to examine how publicly described identification, intervention and appeal arrangements connect technical risk control with procedural protections. The analysis examines what the selected documents establish, what remains unreported and what additional evidence would be needed to evaluate actual outcomes. The study offers platform designers and governance researchers a basis for assessing procedural safeguards and identifying priorities for user protection.

2. Research foundations and method

2.1. Conceptual scope and ethical framework

AI-driven risk control here concerns the identification and management of content-safety risks, payment fraud and other unexamined business functions fall outside its scope. AI-assisted moderation concerns the tools used to assess content, while governance of AI-generated content concerns the material being assessed. These functions can intersect without sharing a single model. Similarly, an informational label, a distribution restriction and an account sanction differ in purpose and consequence.

Procedural justice emphasizes that rules should be intelligible, consequential decisions should be explained and errors should be challengeable. Suzor et al. connect meaningful transparency to reasons that enable accountability [4]. Jhaver's observational Reddit study associates removal explanations with subsequent moderation outcomes, without making causal claims [5]. Are's interview research illustrates difficulties in accessing effective appeals [6]. These studies inform analytical questions rather than supply evidence about Xiaohongshu's users or current performance.

2.2. Documentary selection and analysis

Following documentary analysis, this study selected five records on Xiaohongshu's governance: one technical paper, one regulatory instrument and three original news reports cover March 2025 to April 2026 [7]. Inclusion required an identifiable publisher, retrievable text and relevance to risk identification, labelling or redress. Unattributed claims and accounts without a clear platform identity were excluded. Passages were recorded by date, speaker, governance stage, explanation recipient and stated remedy. Explicit disclosures, unreported information and unsupported outcome judgements were recorded separately. Platform-attributed statements were distinguished from journalists' observations and system authors' reports. Cross-document comparison examined scope and chronology without independently verifying platform claims. The analysis retained qualifying evidence, including appeal access and the distinction between labels and restrictions. Missing disclosures were treated as documentary limitations, not evidence of absent practices. Public commitments received no performance score, unrelated accounts were not combined into individual controversies. This analysis is not an algorithm audit, representative user study or legal-compliance evaluation.

DOI: 10.54254/2753-7064/2026.BA38619

3. Xiaohongshu governance practices and case materials

3.1. AI-assisted review and internal explanations

Li et al. present Hierarchical Guard (Hi-Guard), a multimodal moderation framework incorporating policy definitions and structured risk categories [8]. Their deployment account describes classification outputs and structured evidence intended to support human analysis and re-examination. This provides a documented example of author-reported AI-assisted moderation practice. It also makes the recipient of an explanation analytically important: an output useful to a professional reviewer does not, by itself, establish what an affected user receives.

These findings concern one system and evaluation setting, not independently verified user outcomes. This study does not attribute a disputed post's treatment to Hi-Guard or identify it as the AI-content detector. The records therefore cannot establish a single automated decision pipeline.

3.2. AI content labels and reported appeal arrangements

China's 2025 Measures for Labelling Content Generated or Synthesized by Artificial Intelligence distinguish several identification routes. Article 6 addresses metadata, user declarations and detected indications of generation, including a category of suspected synthetic content [9]. The measures took effect on 1 September 2025. This distinction provides a regulatory context for labelling uncertainty, it does not make every labelled item harmful or require its removal.

A March 2025 report attributes to Xiaohongshu a classifier-based identification process, machine and human involvement, and a feedback mechanism for challenging results [10]. The account predates the measures' effective date. A September report records a platform representative distinguishing labels for suspected AI content from further distribution restrictions where misleading or other problematic effects are identified [11]. It also reports broader user concerns about misclassification and distribution, without independently establishing a particular Xiaohongshu error. In April 2026, Yang reported the platform's governance position and a dedicated channel for appealing AI labels [12]. This should not be described as the first opportunity for any appeal. Table 1 summarises these distinct statements. Together, the materials document identification and redress commitments, but they provide no complete sequence linking an individual original post, disputed decision, appeal and final correction. Across the reports, the observable development is a more specific public description of label-related redress. This, however, does not establish that the dedicated channel replaced general feedback or made correction faster. Descriptive specificity and practical accessibility are separate dimensions of procedural progress. The comparison reveals differences in what is made inspectable: technical documentation explains evidence for reviewers, labelling provisions inform audiences, and appeal announcements address creators. The documents fail to clarify data flow between these elements, positioning the link between detection and redress an object of analysis rather than an assumed property of the system.

Table 1. Selected labelling and appeal records [9-12]

Record	Publicly described arrangement	Evidential boundary
March 2025 measures	Differentiated identification routes	Normative requirements
March 2025 report	Classification and general feedback	Platform-attributed statement

DOI: 10.54254/2753-7064/2026.BA38619

Table 1. (continued)

September 2025 report	Labels and additional restrictions distinguished	No individual outcome verification
April 2026 report	Dedicated label appeal channel	Effectiveness unmeasured

4. Ethical controversies

4.1. Protective efficiency and proportionate intervention

Identifying synthetic material can help audiences understand its origin without establishing whether it contains harmful claims. Treating every source label as censorship would overlook this informational purpose. Conversely, identifying AI involvement does not establish deception, infringement or another substantive violation without further assessment. The platform's reported distinction between labelling and additional restrictions is relevant because it describes separate grounds for these actions [11].

An ethical difficulty arises when an uncertain source judgement produces consequences that users cannot distinguish from a judgement about credibility or misconduct. This is a risk to examine, not an observed effect established by this corpus. The appropriate safeguard depends on the action: a correctable informational label and account exclusion do not warrant identical procedures. Efficiency remains a legitimate consideration, especially where rapid intervention protects other users. Nevertheless, more consequential measures require a clearer connection between evidence, the applicable rule and the action selected. This proportionality argument does not require assuming that the platform currently treats all AI content alike. It specifies what would need justification if source identification were used to support stronger restrictions.

4.2. Meaningful transparency and resistance to evasion

Transparency serves different audiences. Internal explanations help reviewers assess a classification, decision notices help users identify the affected material and understand its treatment, aggregate reporting enables external scrutiny. Hi-Guard's technical account concerns the first relationship [8], whereas the public reports mainly describe governance positions and channels. The corpus does not establish whether users routinely receive explanations sufficient to challenge individual decisions. This question remains open, it does not demonstrate that explanations are absent.

Gorwa et al. describe evasion through reverse-engineering keyword rules [1]. This supports caution about disclosing operational details, but does not justify abandoning reasons. A notice can identify the governing rule, the action and whether a label follows a declaration or an inferred source judgement without exposing the detection model. Where withholding details is necessary, the remaining explanation must still allow a meaningful response. Meaningful transparency is evaluated by what a recipient can understand and challenge, consistent with its accountability function [4]. Publishing rules or technical papers can contribute to that function, but neither alone establishes that affected users can exercise it.

4.3. Responsibility and effective redress

Human involvement alone does not settle responsibility. A reviewer may confirm an automated output without sufficient context, while a model may provide useful evidence without authority to

DOI: 10.54254/2753-7064/2026.BA38619

resolve a contested rule interpretation. Ruckenstein et al. describe the contextual and care-oriented work of human moderators [13]. Their research supports attention to the conditions of review without establishing how Xiaohongshu organises its staffing or labour practices.

The reported appeal arrangements constitute a positive public commitment to redress. Effectiveness requires separate evidence about access, response times, reasons and correction. The selected materials do not supply an outcome dataset capable of establishing these properties. Findings from other platforms cannot fill this gap [6]. Normatively, the platform should retain responsibility for the decisions it implements, even when several technical and human actors contribute. An appeal should examine the disputed basis rather than merely repeat the initial classification. Users should also be able to distinguish reversal of a source label from reversal of a separate content-policy decision, since correcting one does not necessarily resolve the other.

5. Governance recommendations

Review should consider the severity, uncertainty and reversibility of the action. Low-impact source labels can use accessible correction routes, while substantial distribution restrictions or account sanctions require stronger reasons and access to contextual reassessment. Requiring human approval for every item would impose costs without demonstrating better outcomes. A more defensible design prioritises consequential or contested decisions and gives reviewers access to the relevant content and policy basis. Where mixed human and AI production complicates classification, the process should allow contextual evidence without assuming that a creator can disclose an entire workflow. These are proposed safeguards, but their benefits require evaluation using operational records and user research.

Notices should identify the content concerned, the source or rule judgement, the action taken and the available challenge. When a determination concerns suspected generation, uncertainty should remain visible instead of presenting an inference as an established fact. Public explanations need not disclose exploitable features. Further information could be retained for authorised reviewers under appropriate safeguards. Transparency reporting should likewise distinguish label corrections, reversals of restrictions and account restoration rather than merge them into a single appeal figure. Table 2 connects these disclosure needs to the relevant governance decisions, supporting future evaluation without suggesting existing reporting.

Table 2. Proposed safeguards by decision type

Decision	Proposed explanation	Review focus
Source label	Identification basis and uncertainty	Provenance evidence
Content restriction	Applicable rule and affected material	Context and proportionality
Appeal outcome	Decisions changed and remaining restrictions	Correction and implementation

Appeal interfaces should include submission confirmation, status tracking and a reasoned outcome. Evidence requests should be relevant and proportionate, limiting unnecessary personal information. Where an appeal succeeds, the platform should specify which status or restriction has changed and which separate decisions, if any, remain in force. Target response periods should be stated, with priority for severe or time-sensitive restrictions. Any later effectiveness assessment should distinguish submissions, completed reviews and reversals, and explain their denominators. These measures would support accountability without assuming that a high reversal rate necessarily

DOI: 10.54254/2753-7064/2026.BA38619

indicates either successful protection or poor initial moderation. Interpretation requires information about selection, case severity and unresolved appeals.

6. Conclusion

Using public technical, regulatory and journalistic materials, this documentary case study examines ethical controversies surrounding AI-driven risk control and content moderation on Xiaohongshu. The records describe an AI-assisted review system, differentiated grounds for AI content labelling and announced opportunities to challenge labels. They establish public accounts of governance arrangements rather than independently verified moderation outcomes. In particular, the availability of an appeal route cannot guaranty that consumers receive prompt and efficient correction, and the technical publication cannot pinpoint the reason behind a challenged user decision.

The analysis distinguishes judgements about content origin from judgements about harm and the selection of a governance action. Source labelling can advance an audience's understanding without treating the content as unlawful or harmful. Restrictions with greater consequences require an appropriate additional justification. A parallel distinction separates explanations intended for internal reviewers from reasons that users can understand and use in an appeal. These distinctions make specific procedural obligations assessable without demanding disclosure of every detection feature or assuming that automation is inherently illegitimate.

The recommendations connect differentiated review, understandable notices and traceable correction. Their effectiveness remains an empirical question. The study draws on a small purposive corpus, includes platform-attributed statements and does not use internal logs or interviews with affected users. It therefore cannot measure error rates, disparate impacts or appeal success. Future research could examine actual notices and completed appeals, with appropriate consent and privacy protections, to assess whether stated procedures translate into usable remedies. The present contribution is a bounded account of the evidence and ethical reasoning needed to connect content-safety objectives with accountable decisions in a social platform setting.

References

[1] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 1–15. https://doi.org/10.1177/2053951719897945

[2] Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2), 1–5. https://doi.org/10.1177/2053951720943234

[3] Ju, R. (2022). Producing entrepreneurial citizens: Governmentality over and through Hong Kong influencers on Xiaohongshu (Red). *Policy & Internet, 14*(3), 618–632. https://doi.org/10.1002/poi3.324

[4] Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication, 13*, 1526–1543. https://ijoc.org/index.php/ijoc/article/view/9736

[5] Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter?: User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Article 150, 1–27. https://doi.org/10.1145/3359252

[6] Are, C. (2025). 'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation. *Information, Communication & Society, 28*(11), 1997–2014. https://doi.org/10.1080/1369118X.2024.2396621

[7] Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

[8] Li, A., Jin, W., Tong, J., Qin, P., Li, J., & Lu, G. (2026). Towards trustworthy multimodal moderation via policy-aligned reasoning and hierarchical labeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge*

DOI: 10.54254/2753-7064/2026.BA38619

Discovery and Data Mining V.1 (pp. 2278–2289). ACM. https://doi.org/10.1145/3770854.3783934

[9] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, & National Radio and Television Administration. (2025, March 14). Notice on issuing the Measures for Labeling Content Generated or Synthesized by Artificial Intelligence [in Chinese]. https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

[10] Yang, L., & Fan, W. (2025, March 18). AI content labeling comes under stronger regulation: Testing implementation across 23 mainstream platforms [in Chinese]. *Southern Metropolis Daily (Nanfang Plus)*. https://static.nfnews.com/content/202503/18/c11095123.html?enterColumnId=2556

[11] Xiao, X., Zhang, X., & Bai, J. (2025, September 4). First test of AI content "carrying identification": Which of 35 applications slip through the net? [in Chinese]. *21st Century Business Herald*. https://www.21jingji.com/article/20250904/herald/ce1753b0474be884a904a4c5aef8d53e.html

[12] Yang, M. (2026, April 27). Focusing on AI governance and supporting quality creation: Xiaohongshu announces its AI governance position for the first time [in Chinese]. *Economic Information Daily Online*. https://www.jjckb.cn/20260427/a1a0a208bc244043bbfe3b2d2b5f9dec/c.html

[13] Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society, 22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990