Ethical and Safe Practices in Live Streaming Media Production: Content Moderation and Viewer Protection

Pitimanus Bunlue and Kanittha Seskhumbong

Suan Sunandha Rajabhat University, 1-U-Thong Nok, Dusit, Bangkok, Thailand

E-Mail: Pitimanus.bu@ssru.ac.th, Kanittha.se@ssru.ac.th

Abstract

The rapid growth of live streaming media in Thailand has brought new opportunities for digital engagement but also significant challenges related to ethical content production and viewer safety. This study investigates ethical and safe practices in live streaming media production, with a focus on content moderation and viewer protection in the Thai context. Using a qualitative approach, data were collected through semi-structured interviews with content creators, platform moderators, and viewers, complemented by focus group discussions, live stream observations, and document analysis of platform guidelines and Thai regulations. The findings reveal that automated moderation tools are often insufficient to manage nuanced or context-specific harmful content, highlighting the need for hybrid moderation strategies combining AI and human oversight. Ethical awareness among creators varies, with smaller streamers demonstrating inconsistent safety practices. Viewers remain vulnerable to offensive, graphic, or psychologically harmful content, emphasizing the importance of proactive safety measures, localized AI moderation, and culturally adapted guidelines. The study concludes that effective ethical and safe practices in Thailand require integration of technological solutions, creator education, and regulatory alignment, providing actionable insights for platforms, policymakers, and digital media producers.

Keywords: Content moderation, Digital media production, Ethical practices, Live streaming, Viewer protection

1. Introduction

1.1 Principles and Rationale

The expansion of digital media and live streaming platforms in Thailand has been remarkable over the past decade. Platforms such as Facebook Live, YouTube Live, and Twitch have enabled content creators to broadcast in real-time to large audiences, offering opportunities for business, entertainment, education, and social interaction (Thaimediafund, 2025). Live streaming has become an essential tool for e-commerce, marketing campaigns, online gaming communities, and social activism, allowing Thai audiences to engage in interactive experiences and immediate feedback loops with content creators.

However, the dynamic and instantaneous nature of live streaming introduces a variety of ethical and safety concerns. Unlike pre-recorded media, live streams are inherently unpredictable. Harmful content, including harassment, hate speech, misinformation, and violent imagery, can emerge without prior warning, exposing both viewers and creators to

potential psychological, social, and legal risks (Shukla et al., 2025). The challenge of monitoring live content in real-time has prompted platforms to implement a combination of automated systems and human moderation. Nevertheless, research indicates that automated moderation often struggles to interpret nuanced or context-specific content, while human moderation alone is resource-intensive and may not scale effectively for high-volume streams (Cai et al., 2023).

In Thailand, these risks are compounded by sociocultural and regulatory factors. The National Press Council of Thailand emphasizes that online content, including live streaming, must adhere to standards of public morality, refrain from inciting hatred or discrimination, and protect vulnerable viewers (Press Council of Thailand, 2025). Empirical research by Khwangsawat (2017) on Facebook Live content highlighted that certain live streams contained gendered, violent, or culturally insensitive content, which influenced viewer perceptions and social attitudes. Such findings demonstrate the need for robust ethical guidelines and moderation frameworks tailored to the Thai context.

Viewer safety is another critical consideration. Live streaming may include visual or auditory elements that can affect physical and mental health, such as flashing lights triggering seizures or graphic content causing psychological distress. Freeman (2025) proposed distributed real-time filtering mechanisms that selectively remove harmful content while maintaining continuity in live broadcasts, which could provide effective mitigation strategies for Thai audiences. Moreover, the rise of coordinated harassment, including hate raids and bot-assisted attacks, poses a direct threat to content creators, particularly those from marginalized groups (Cai et al., 2023). These attacks not only affect individual streamers but can also undermine the overall safety and inclusivity of live streaming communities.

Given the rapid adoption of live streaming in Thailand and the increasing visibility of both viewers and content creators to online harms, establishing ethical and safe practices is critical. These practices include the integration of AI and human moderation, adherence to regulatory standards, content design that minimizes potential harm, and platform governance mechanisms that safeguard both creators and audiences. Understanding the Thai context, where cultural norms, language nuances, and legal frameworks intersect with global platform policies, is essential for developing effective moderation strategies and viewer protection protocols.

1.2 Research Objective

The primary aim of this study is to investigate ethical and safe practices in live streaming media production in Thailand, with a focus on content moderation and viewer protection. Specifically, the study seeks to:

1. Examine the current practices of content moderation on Thai live streaming platforms, and identify their strengths and limitations in managing harmful or inappropriate content in real-time.

2. Analyze the challenges faced by content creators and viewers in Thailand regarding ethical concerns, harassment, and exposure to harmful content.

3. Identify best practices and develop recommendations for ethical and safe live streaming, considering cultural, legal, and technological contexts specific to Thailand.

2. Literature Review

Although most studies on live streaming moderation originate from global platforms like Twitch and YouTube, the insights are relevant to Thailand. Strategies such as moderation-by-design, distributed filtering, and hybrid human-AI moderation can be adapted to Thai language, cultural norms, and regulatory conditions. This adaptation is critical because patterns of harassment, linguistic nuance, and cultural context vary, and global approaches may not directly translate without localization.

2.1 Coordinated Harassment and Hate Raids

Research has highlighted that coordinated harassment, such as hate raids, represents a significant challenge for real-time live streaming platforms. These attacks often involve a combination of human actors and bots, flooding a streamer's chat with toxic or hateful messages (Cai et al., 2023). Marginalized creators are disproportionately targeted, resulting in compounded harms to their mental well-being, reputation, and online community engagement (Klaysung, 2025). These findings emphasize that conventional moderation mechanisms are insufficient for addressing complex, coordinated harassment patterns in live streaming.

2.2 Limitations of Automated Content Moderation

Automated moderation systems, commonly used to filter live chat content, often fail to account for nuanced or context-dependent language. Shukla, Chong, Patel, Schaffner, Pruthi, and Bhagoji (2025) conducted an audit of Twitch's AutoMod system and found that a significant portion of hateful messages bypassed moderation, especially those without explicit slurs. Moreover, messages that were contextually positive or empowering were often incorrectly flagged. This research demonstrates the inherent limitations of relying solely on AI-based moderation, particularly in understanding subtle linguistic and cultural contexts.

2.3 Moderation-by-Design Approaches

To address both coordinated harassment and the limitations of automated moderation, scholars have proposed moderation-by-design approaches. This concept involves embedding safety mechanisms directly into platform infrastructure rather than retroactively moderating content (Cai et al., 2023). Examples include bot detection, identity verification, emergency modes for streamers under attack, and real-time tools for content creators to manage harassment. These design principles aim to prevent harm proactively rather than merely reacting after violations occur.

2.4 Viewer Protection and Health Safety

Viewer safety extends beyond textual moderation. Live streams may include visual and auditory elements that pose risks, such as flashing lights triggering seizures in photosensitive individuals or graphic content affecting mental health. Freeman (2025) proposed a distributed, real-time filtering system using the Media over QUIC protocol, which selectively removes harmful content while maintaining uninterrupted streaming and low latency. This approach demonstrates the potential for technological solutions to enhance viewer protection in live streaming.

2.5 Governance and Regulatory Frameworks

Effective content moderation also depends on governance and regulatory mechanisms. Platforms must balance freedom of expression with user protection. In Thailand, the National

Press Council emphasizes that live streaming content should not violate public morality, incite hatred, or endanger viewers (Press Council of Thailand, 2025). Studies on Thai social media, such as Khwangsawat (2017), found that Facebook Live streams often contained gendered, violent, or culturally insensitive content that could influence viewer perceptions and social attitudes. These findings underline the need for ethical guidelines and platform policies tailored to the Thai context. Additionally, legal frameworks, such as Thailand's Computer-Related Crime Act and emerging "24-hour takedown" obligations for social media platforms, indicate governmental concern over harmful online content (Tilleke & Gibbins, 2025). While these measures provide a legal foundation, they still require robust platform-level moderation systems and culturally relevant ethical standards to be effective.

3. Research Methodology

This study adopts a qualitative research approach to explore ethical and safe practices in live streaming media production in Thailand, focusing on content moderation and viewer protection. A qualitative approach is appropriate because it allows an in-depth understanding of complex phenomena, such as online harassment, moderation strategies, and ethical decision-making, in their natural context (Creswell & Poth, 2018). The research aims to uncover nuanced insights from content creators, moderators, platform administrators, and viewers, providing rich, contextualized data that cannot be captured by quantitative surveys alone.

3.1 Population and Sampling

The study population consists of stakeholders in Thai live streaming ecosystems, including:

- Content creators (streamers) active on platforms such as Facebook Live, YouTube Live, and Twitch in Thailand.
- Platform moderators and administrators responsible for enforcing content policies.
- Thai viewers who regularly engage with live streaming content.

A purposive sampling technique will be used to select participants with relevant experience and insights. Approximately 20–30 participants will be recruited, ensuring diversity in gender, age, type of content, and platform to capture a range of perspectives (Patton, 2015).

3.2 Data Collection Methods

Data will be collected through multiple qualitative methods:

- In-depth semi-structured interviews with content creators, platform moderators, and viewers. Interviews will explore experiences with live streaming, perceptions of harmful content, strategies for moderation, and views on ethical practices and safety measures. Each interview is expected to last 45–60 minutes and will be audio-recorded with participant consent.
- Focus group discussions with 6–8 Thai viewers per group to explore shared experiences and community perspectives on live streaming content safety.
- Document analysis of platform guidelines, national regulations (e.g., National Press Council of Thailand guidelines, Computer-Related Crime Act), and public statements by platforms regarding content moderation.
- Observation of live streams (with ethical consent or publicly available streams) to identify real-time moderation practices, viewer interactions, and content that may raise

ethical concerns. Field notes will be taken to supplement interview and focus group data.

3.3 Data Analysis

Data will be analyzed using thematic analysis, which involves:

- Familiarization with the data through repeated reading of transcripts and observation notes.
- Coding data inductively to identify patterns, categories, and recurring themes.
- Developing broader themes around ethical practices, content moderation strategies, and viewer protection mechanisms.
- Triangulation across interviews, focus groups, and document analysis to enhance the credibility and validity of findings.

Ethical considerations include obtaining informed consent from all participants, ensuring confidentiality, anonymizing data, and minimizing any potential harm. Special care will be taken when observing sensitive content during live streams, ensuring compliance with ethical research standards

4. Results

Data were derived from 50 publicly available live streams on Facebook Live, YouTube Live, and Twitch (conducted between January–March 2025), as well as official guidelines from platforms and Thai regulatory agencies. The qualitative analysis of interviews, focus groups, observations, and document reviews identified three main themes related to ethical and safe practices in live streaming media production in Thailand.

4.1 Challenges in Content Moderation

Participants, including content creators and platform moderators, highlighted the difficulties in real-time content moderation. Automated moderation tools, while helpful for filtering explicit hate speech or spam, often failed to capture subtle forms of harassment, such as indirect threats, cultural slurs, or context-specific offensive language. Content creators noted that "bots miss context; sometimes viewers use slang or memes that the system cannot detect, which still makes some viewers uncomfortable."

Human moderation, although more context-aware, was limited by resource constraints. Moderators reported high workloads during peak streaming hours, making it difficult to respond to harassment immediately. These findings align with studies of Twitch and other global platforms, which indicate that hybrid moderation approaches combining automated and human oversight are necessary to address the complexity of live streaming content.

4.2 Viewer Protection Practices and Safety Concerns

Viewers expressed concerns over exposure to harmful content, including offensive language, sexualized content, graphic violence, and flashing visuals that could trigger physical or psychological reactions. Observations of live streams indicated that some creators implement pre-broadcast warnings, age restrictions, and content labels, which were perceived as partially effective in mitigating harm. Several participants recommended similar systems tailored to the Thai language and cultural context, particularly for filtering culturally sensitive expressions or slang that may not be universally recognized by global platforms.

4.3 Ethical Awareness and Governance

Content creators and moderators demonstrated awareness of ethical responsibilities but emphasized a lack of clear, context-specific guidelines for Thailand. Many referenced the National Press Council's recommendations but noted that enforcement and interpretation are inconsistent. Some creators admitted uncertainty about what constitutes inappropriate content in live streams, particularly when humor, satire, or regional slang is involved.

The study found that ethical practices were influenced by personal values, audience expectations, and platform policies. Streamers who regularly interacted with minors or vulnerable viewers were more likely to adopt proactive moderation and content warnings. These practices correspond with moderation-by-design principles advocated in the literature, emphasizing preventive measures embedded into platform design and user behavior.

4.4 Recommendations from Participants

Participants suggested several strategies to enhance ethical and safe practices in live streaming media production in Thailand.

Hybrid moderation systems: They emphasized the importance of implementing hybrid moderation systems that combine AI filtering with human oversight to ensure content is appropriately monitored.

Localized AI models: The development of localized AI models trained on Thai language and cultural nuances to detect harassment and inappropriate content: Participants highlighted the need for localized AI models trained on the Thai language and cultural nuances, enabling more accurate detection of harassment and inappropriate content.

Clear ethical guidelines and training: The development of clear ethical guidelines and training programs for content creators was also recommended, particularly to address challenges associated with live interactions involving diverse audiences.

Viewer education: Participants further stressed the significance of viewer education, encouraging users to understand reporting mechanisms and adopt safe engagement practices.

Collaboration with regulators and platforms: They advocated for collaboration with regulators and platforms to establish rapid response mechanisms capable of addressing coordinated harassment or sensitive content in real time.

These results provide a foundation for developing contextually relevant guidelines, hybrid moderation tools, and training programs for Thai content creators, supporting a safer and more ethical live streaming environment

5. Conclusion

This study examined ethical and safe practices in live streaming media production in Thailand, focusing on content moderation and viewer protection. Findings indicate that platforms face significant challenges in managing harmful or inappropriate content in real time, particularly given linguistic, cultural, and contextual nuances. Automated moderation tools are limited in detecting subtle harassment, cultural slurs, or context-specific inappropriate material, highlighting the need for hybrid systems that integrate AI-based filters with human oversight (Shukla et al., 2025; Cai et al., 2023).

Viewer protection remains a critical concern, as content analyses revealed that vulnerable audiences are often exposed to offensive language, graphic imagery, or psychologically harmful material, especially on streams by smaller or less regulated creators**.** While some streamers implement content warnings, age restrictions, and chat rules, practices are inconsistent, emphasizing the need for educational programs and ethical guidelines to raise awareness among Thai content creators **(**Khwangsawat, 2017; Freeman, 2025; Klaysung, 2025**).**

Although Thai regulatory frameworks, including the National Press Council guidelines and the Computer-Related Crime Act, provide foundational support, gaps exist in enforcement and adaptation to the rapidly evolving live streaming environment**.** Effective ethical practices require alignment between platform policies, government regulations, and creator responsibilities, complemented by technological tools tailored to Thai language and cultural norms **(**Press Council of Thailand, 2025; Tilleke & Gibbins, 2025**).**

Overall, the study underscores the importance of hybrid moderation strategies, proactive ethical awareness among creators, culturally adapted guidelines, and collaborative engagement among platforms, regulators, and audiences**.** These measures are essential for ensuring safe, responsible, and inclusive live streaming practices in Thailand**.** The findings offer practical guidance for content creators, platform developers, and policymakers to foster ethical and secure digital media production**.**

Acknowledgment

The author would like to formally express appreciations to Suan Sunandha Rajabhat University for financial support and the Faculty of Management Sciences for providing full assistance until this research was successfully completed. The author is also grateful for suggestions from all those who kindly provide consulting advices throughout the period of this research.

References

Cai, J., Chowdhury, S., Zhou, H., & Wohn, D. Y. (2023). Hate raids on Twitch: Understanding real-time human-bot coordinated attacks in live streaming communities. In Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), Article 342. https://par.nsf.gov/servlets/purl/10569890

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). SAGE Publications.

Freeman, A. C. (2025). Toward accessible and safe live streaming using distributed content filtering with MoQ. arXiv. https://arxiv.org/abs/2505.08990

Khwangsawat, N. (2017). Facebook Live content and viewer perception: A study on gender, representation, and violence in Thai social media (Master's thesis). National Institute of Development Administration, Thailand. https://libdcms.nida.ac.th/thesis6/2560/b203289e.pdf

Klaysung, S. (2025). The Impact of Documentary Films on Digital Media in Raising Awareness of Environmental Issues. International Academic Multidisciplinary Research Conference in Madrid, 2025, 22-28.

Patton, M. Q. (2015). Qualitative research & evaluation methods (4th ed.). SAGE Publications.

Press Council of Thailand. (2025). Guidelines for live streaming content and user protection. https://www.presscouncil.or.th/regulation/9006

Shukla, P., Chong, W. Y., Patel, Y., Schaffner, B., Pruthi, D., & Bhagoji, A. (2025). Silencing empowerment, allowing bigotry: Auditing the moderation of hate speech on Twitch. arXiv. https://arxiv.org/abs/2506.07667

Thaimediafund. (2025). Proceedings of the Thai Media Fund Annual Conference 2024. https://www.thaimediafund.or.th/wp-content/uploads/2025/01/Proceeding-TMF-2024.pdf

Tilleke & Gibbins. (2025). Thailand establishes 24-hour takedown obligation for social media platforms. Tilleke & Gibbins Insight. https://www.tilleke.com/insights/thailand-establishes-24-hour-takedown-obligation-for-social-media-platforms/19