# Reflections on AIGC Application and Management in the Platform Governance System

Jiongzhu Liu

School of Culture and Creativity, BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY, Zhuhai, 519000, China

**Abstract.** With the rapid development and widespread application of AI-generated content (AIGC) based on large language models, diffusion models, multimodal generative systems, and so on, the information ecosystem of digital platforms is changing significantly. The current content governance system is based on the traditional model of professional-generated content (PGC), user-generated content (UGC) and algorithmic curation, and is increasingly unable to handle low-cost, large-scale, and difficult-to-attribute synthetic content. Using the combined framework of platform governance theory, the political economy of surveillance capitalism, and new regulatory tools such as the EU AI Act (2024), China's Interim Measures for the Management of Generative AI Services (2023), and the C2PA content provenance standard, this paper studies the governance issues of AIGC. A risk-tiering, multi-stakeholder governance model is proposed to address governance problems of AIGC on digital platforms. Based on the above analysis, the effective governance of AIGC needs to construct a mechanism for verifiable content provenance, distribute accountability throughout the entire AIGC supply chain, and develop an adaptive regulatory system that supports dynamic rule-making and cross-border cooperation.

**Keywords:** AIGC, platform governance, generative AI, content moderation, algorithmic accountability, EU AI Act, disinformation, digital regulation.

## 1. Introduction

### 1.1 Research background

Large-scale generative AI systems can now create text, images, audio and videos that are often indistinguishable from human-made works at near-zero marginal cost. The above functions have provided a new basis for platform management that is suitable for professional content and also covers user-generated content and algorithms. These models generally considered human actors to be the primary creators of content, and the platforms only organised, ranked and moderated the information flow that resulted from this creation. AIGC can be produced in large numbers for specific groups, distributed by recommendation systems, and lacks an identifiable human creator or explicit creative purposes and accountability. At the same time, ethnographic research on the moderation of commercial content has also found that the current moderation system is already under considerable pressure from both the large quantity and the psychological demands of human-generated harmful information [2]. Synthetic content may place extra pressure on the existing system and raise new legal and conceptual issues for the previous framework.

### 1.2 Research Questions And Contributions

The three research questions of this paper are as follows: (RQ1) How do the technical features of AIGC change the problems of platform content governance? (RQ2) What are the patterns and deficiencies in the current AIGC governance models of various platforms and regulatory areas? (RQ3) What core attributes should a governance system for AIGC on platforms have? This paper has the following structure. Section 2 provides the theoretical basis; Section 3 lists typical applications and risk situations of AIGC; Section 4 puts forward a governance system; and Section 5 discusses the implications and deficiencies of this study. In addition to the theoretical analysis of platform governance and surveillance capitalism [1, 3, 4], this paper will also introduce some regulations and technical standards, such as the EU AI Act, China's Interim Measures, the OECD AI principles, the

global partnership on AI, and the C2PA provenance framework [5, 6, 7, 8, 18]. This paper has made the following two contributions: first, it places the governance problems of AIGC in the context of existing platform governance theory; second, it puts forward a risk-tiered, multi-stakeholder framework for future empirical and policy studies.

## 2. Theoretical Framework

### 2.1 Platform Governance As Structural Power

Platform governance refers to the technical system, policies and institutions of a platform that regulate user behaviour and the flow of information.Gillespie thinks that content moderation is a form of normatively loaded governance, and thus it is not a technical problem; rather, the choice platforms make regarding what content to host, promote or remove inherently contains normative values in their design and rules[1, 9]. He has proposed a three-tiered structure for platform governance: the architecture level to embed norms in the design of technology; the rule-and-policy level to set community norms and enforcement measures; and the social-norm level through which expected behavior is gradually internalised by users [1]. The above tiers need to work together well for good governance.

Roberts has extended the above framework to describe the labour infrastructure of moderation and shown that automated systems are built on a large scale of human review work [2]. Klonick believes that the rules of platform content are a form of private constitutionalism for online speech [10]. The rise of AIGC is worsening the above problems; moderation now needs to handle content that is easy to produce and difficult to trace back, and it may exceed the capacity of human-in-the-loop systems.

### 2.2 AIGC and the Platform Economy

Platform capitalism's dominant business model is based on data collection and behaviour analysis [3]. Zuboff refers to this as a system of surveillance capitalism that uses all stages of life as "raw material" to create predictive products through AI, thereby maximising user engagement and data collection [4]. AIGC expands on this by enabling platforms and tool providers to generate content independently of users for the first time. Synthetic content can be directly optimised for engagement and the generation of training data, and thus platforms may be compelled to adopt less transparent optimisation goals and further increase the asymmetry between platform operators and users [3, 4].

### 2.3 Regulatory Frameworks for AIGC and AI

Regulatory responses to AIGC and high-risk AI systems are emerging but scattered. The EU AI Act will use a risk-based approach to classify general-purpose AI models differently and set various obligations such as documentation, copyright information and transparency, and labelling for AI-generated content and deepfakes [5]. China’s Interim Measures for the Management of Generative AI Services specify that service providers shall be responsible for security assessment, labelling of AI-generated content, and complaint and penalty mechanisms [6]. OECD AI principles and the global partnership on AI (GPAI) are non-binding but influential intergovernmental frameworks to promote trustworthy, human-centred AI, and they also explicitly advocate for cross-border cooperation in AI governance [7, 18]. Although the above instruments have jointly built a norm system that promotes openness, accountability and global cooperation, the specific operating mechanism of AIGC in platforms has not yet been determined.

The above theoretical ideas also provide a foundation for the governance model that will be presented later in this paper. Gillespie explains the structural governance power platforms have over visibility and speech; Roberts shows that human moderation infrastructure is already under pressure; Srnicek and Zuboff reveal the commercial incentives that may promote engagement-driven synthetic content, and Klonick demonstrates that platform rules increasingly function as quasi-public governance arrangements. In short, all the above theories indicate that the governance of AIGC

should be all-encompassing and address problems such as technical detection, institutional accountability, platform incentives, and public supervision.

## 3. AIGC Applications And Emerging Governance Challenges

### 3.1 Content Production And Distribution

AIGC is changing how content is produced for social media, news and e-commerce. Generative tools can produce personalised content at a large scale for creators and platforms, thus increasing competition in the creator economy and making the distinction between human and synthetic authorship increasingly blurred [3, 4]. The expansion of AI-assisted content production in journalism has led to new problems regarding the reliability and editorial responsibility of such content, as algorithmically generated or AI-assisted works may be prioritised for popularity over public-interest purposes [15]. Platform recommendation systems will amplify AIGC that can achieve high click-through and watch-time rates, and as a result, optimization goals may take precedence over traditional editorial decisions.

### 3.2 AIGC in Content Moderation

AI systems in the platform have been added to identify hate speech, nudity, spam and other policy violations. Gorwa, Binns and Katzenbach believe that algorithmic content moderation contains political choices in the criteria for classification and deletion, and is therefore subject to technical and normative constraints [14]. Roberts has pointed out that the claims that AI can replace human moderators have repeatedly been overly optimistic, especially for context-dependent judgment of subtle content [2]. When the same generative abilities used to create harmful content are employed for evasion of detection, an adversarial loop emerges in which AIGC is continuously modified to bypass filters, and detection systems lag behind.

### 3.3 Risk Scenarios: Disinformation, Deepfakes, And Copyright

Three risk scenarios show the governance deficiencies of AIGC. First, AI-generated disinformation and impersonation content have been used to disrupt the election. In the 2023 parliamentary elections in Slovakia, AI-generated audio deepfakes of a liberal candidate and a journalist discussing electoral fraud were widely circulated on social media within 48 hours of the election, and at that time, traditional media fact-checking had failed to provide an effective refutation [16, 17]. Therefore, synthetic audio can take advantage of regulatory gaps and the short-term demands for fact-checking in the run-up to elections. Second, deepfake images and videos can seriously harm reputation and undermine the credibility of audiovisual evidence in cases of non-consensual intimate images and political fraud [19, 20]. Third, AIGC poses new copyright risks. Training generative models on copyrighted works and then creating content that is substantially similar to protected works presents a problem for the current intellectual property system, as indicated by the debate over the EU AI Act and other discussions on copyright policy [5, 23].

## 4. Governance Challenges And Management Framework

### 4.1 Detection, Provenance, And Attribution

Good governance of AIGC should have reliable ways to detect synthetic content and trace its origins. Some AI-generated text detectors have shown poor accuracy and have even been taken offline due to high error rates [21]. Provenance of content, for example, in the form of the C2PA technical specification, is also used to add cryptographically verifiable metadata to digital objects that records their creation, modification, and use in artificial intelligence [8]. Leading technology companies have started building C2PA-compatible "Content Credentials", but the rollout is still voluntary. Currently, C2PA-style provenance disclosure has not yet become a widely harmonised legal requirement across

jurisdictions, although transparency provisions in the EU AI Act are broadly in line with such a standard [5, 8]. Diakopoulos's idea of algorithmic accountability proposes adding external auditing and public disclosure of platforms' algorithms to provenance tools to avoid self-reporting [11].

### 4.2 Accountability Gaps And Jurisdictional Fragmentation

Accountability for harmful AIGC remains fragmented. Responsibility can be taken by model developers (for training data and safety guardrails), platform operators (for amplification and recommendation), deploying users (for intentional misuse), and moderation teams (for failure to detect). The current liability system in the EU digital services act and section 230 of the US only covers these types of connections to a limited extent [13, 22]. China's Interim Measures have placed more direct obligations on AIGC service providers but are restricted by cross-border enforcement constraints [6]. The OECD has recommended the following requirements for effectively handling disinformation: openness, responsibility, multi-stakeholder involvement, and policies for all sectors of society [12]. Different regulatory systems around the world have still created enforcement gaps in the global AIGC supply chain. Surveillance capitalism has also failed to hold platforms accountable by giving more importance to the commercial interests of the platforms in increasing user engagement and data collection over regulating content [4].

### 4.3 Toward a Risk-Tiered, Multi-Stakeholder Governance Model

Based on the above, this paper proposes the four components of a governance system for AIGC on platforms. First, mandatory labelling of AIGC should be introduced on major platforms in cooperation with interoperable provenance standards such as C2PA, and enforcement measures and sanctions for non-compliance should be in line with the transparency obligations of the EU AI Act [5, 8]. Second, it will be divided into different risk groups by the system. Synthetic entertainment content, artistic filters, political deepfakes and AI-generated news should be regulated differently according to their potential harm, and stricter standards should be imposed on content that may affect elections, public health and vulnerable groups [5, 6, 7, 12]. Low-risk AIGC, such as entertainment filters and clearly fictional creative works, may only need basic labelling and user education in this model. Medium-risk content, such as AI-generated news summaries, commercial advertisements and synthetic influencer content, should be subject to more stringent disclosure requirements and platform-level monitoring and complaint-handling mechanisms due to the potential harm to the public and choices of consumers. High-risk content, such as political deepfakes, health misinformation, non-consensual intimate imagery and synthetic content targeting vulnerable groups, should be subject to mandatory provenance disclosure, rapid removal procedures, human review and potentially legal sanctions. The aim of the above three-level system is to reduce the likelihood and severity of damage to social order. Third, a multi-stakeholder governance system should be established to build an adaptive rule-making system that includes AI developers, platform operators, regulators, civil society organizations, academic researchers and user representatives, and new risks can be continuously monitored and standards adjusted [1, 7, 18]. Fourth, building on the OECD AI principles and the global partnership on AI that both advocate for international cooperation in the governance of AI, cross-border coordination needs to be strengthened [7, 18].

## 5. Discussion and Conclusion

This paper finds that AIGC exacerbates the existing problems of platform governance and creates new ones. Detection and attribution are still technically difficult; economically, AIGC will continue to be used to optimise engagement, and at the institutional level, regulatory responses are emerging but have not been implemented uniformly. Platform governance theory and the political economy of surveillance capitalism offer theoretical support for the study of the development of these phenomena, but now they need to be extended to address synthetic content that lacks human authorship yet generates behaviour data.

The proposed risk-tiered, multi-stakeholder governance model is not a full-fledged solution, but it has identified the general directions for the development of AIGC governance: verifiable provenance, shared responsibility across the entire supply chain, dynamic institutional learning, and cross-border cooperation. However, the above model is also limited. First of all, provenance standards such as C2PA are voluntary; therefore, they may not prevent malicious removal of metadata or distribution through unregulated channels by bad actors. Second, although mandatory labelling will increase transparency, it is not guaranteed to help users learn about and rely on source-tracing attributes. Third, the type of risk may be different in different political and cultural environments, such as satire, political speech, and public-interest journalism. The above shortcomings show that the governance of AIGC should be adaptable rather than rigid, and technical tools need to be combined with institutional supervision and public education.

In the future, some empirical studies will be carried out to explore how different platforms use AIGC labelling, how users react to provenance signals, and what impact various regulatory systems have had on the deployment and governance of generative AI systems in different linguistic and political environments.

## References


[1] Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

[2] Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

[3] Srnicek, N. (2017). Platform capitalism. Polity Press.

[4] Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

[5] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, OJ L-2024-1689.

[6] Cyberspace Administration of China. (2023). Interim measures for the management of generative artificial intelligence services.

[7] Organisation for Economic Co-operation and Development. (2019, amended 2024). Recommendation of the Council on artificial intelligence (OECD/LEGAL/0449). OECD.

[8] Coalition for Content Provenance and Authenticity. (2024). C2PA technical specification (Version 2.1). https://spec.c2pa.org/specifications/specifications/2.1/specs/_attachments/C2PA_Specification.pdf

[9] Gillespie, T., Boczkowski, P. J., & Foot, K. A. (Eds.). (2014). Media technologies: Essays on communication, materiality, and society. MIT Press.

[10] Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131(6), 1598-1670. https://doi.org/10.2139/ssrn.3135894

[11] Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398-415. https://doi.org/10.1080/21670811.2014.976413

[12] Matasick, C., Alfonsi, C., & Bellantoni, A. (2020). Governance responses to disinformation: How open-government principles can inform policy options (OECD Working Papers on Public Governance, No. 39). OECD Publishing. https://doi.org/10.1787/9789264321132-en

[13] United States Congress. (1996). 47 U.S.C. § 230: Protection for private blocking and screening of offensive material. U.S. Government Publishing Office.

[14] 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/10.1177/2053951719897594

[15] Simon, F. M. (2024). Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena. Tow Center for Digital Journalism, Columbia Journalism School. https://www.cjr.org/tow_center_reports/artificial-intelligence-in-the-news.php

[16] Meaker, M. (2023, October 3). Slovakia's election deepfakes show AI is a danger to democracy. WIRED. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy

[17] United Nations Educational, Scientific and Cultural Organization. (2025). Case study: Can we believe what we hear? Media and information literacy case study on AI-generated audio deepfakes in elections. UNESCO. https://www.unesco.org/mil4teachers/en/toolkit-media/indicator-5/activities/ai-disinformation/case-study-2

[18] Organisation for Economic Co-operation and Development. (2024). Global partnership on artificial intelligence (GPAI). https://www.oecd.org/en/about/programmes/global-partnership-on-artificial-intelligence.html

[19] Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107, 1753-1820. https://doi.org/10.15779/Z38CLR-ART-19-0007

[20] Umbach, R., Henry, N., Beard, G., & Berryessa, C. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-20). https://doi.org/10.1145/3613904.3642382

[21] OpenAI. (2023). New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

[22] European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC. Official Journal of the European Union, OJ L-277, 1-102.

[23] Lucchi, N., & Hunter, S. (2025). Generative AI and copyright: Training, creation, regulation. European Parliament. https://www.europarl.europa.eu/thinktank/en/document/IUST_STU%282025%29774095