Article

Not peer-reviewed version

A Framework for Ethical AI-Generated Content Governance

Joseph Michael Odhiambo * and Kennedy Ondimu

Posted Date: 3 September 2025

doi: 10.20944/preprints202509.0271.v1

Keywords: AI-generated content; ethical governance; bias mitigation; content provenance; transparency; ICT policy; AI ethics

 doi:10.20944/preprints202509.0271.v1

Article

A Framework for Ethical AI-Generated Content Governance

Joseph Michael Odhiambo [1,*] and Kennedy Ondimu [2]

[1] Lukenya University

[2] Technical University of Mombasa

* Correspondence: obijis@gmail.com

Abstract

Current applications of generative AI technologies—large language models and diffusion models-based picture generators—have been linked with profound ethical and societal issues. Misinformation, intellectual property hijacking, algorithmic bias, and content integrity problems must be well-regulated. Existing frameworks address these issues in disconnected and missing a clear actionable framework. This study conceives a broad-ranging ethical AI-content creation regulation regime, founded on three pillars: top-level ethical guidelines, process working norms derived from guidelines, and technical facilitation mechanisms. The regime emphasizes transparency, accountability, fairness, and human governance with implementation avenues through data regulation, explainability, content provenance, and user interaction. Guided by positive research methodology and synthesis of literature-based, the idealized regime redresses mistakes in existing regulatory and technology practice. Policy recommendations are presented for policymakers, developers, and users as inputs to be used by them for accountable AI output generation, deployment, and use. Incorporating ethical theory and functional ICT mechanisms, the project is evolving on reliable, fair, and human-centered AI systems.

Keywords: AI-generated content; ethical governance; bias mitigation; content provenance; transparency; ICT policy; AI ethics

1. Introduction

The development of highly advanced artificial intelligence (AI) engines that are capable of generating more advanced and lifelike content in an immense variety of modalities – from text, image, and sound to video and even code – is a revolution in digital creation and dissemination (Kaur et al., 2025). These generative AI technologies are laced with promise for invention, industry-wide optimization, and the democratization of creative work (Bansal et al., 2024). From optimizing content creation workflows and tailoring user experiences to facilitating creative expression and scientific inquiry, applications of AI-authored content are multiplying and permeating various aspects of our lives (Morris, 2023).

But this technological advancement is accompanied by an array of ethical challenges that must be afforded thorough consideration and active oversight. The same characteristics that make AI-produced content so powerful – its scalability, speed, and increasing indistinguishability from human-created content – also pose the analogous threats (Sieg, 2023). These risks include a broad spectrum, ranging from the potential spreading of misinformation and disinformation on a scale never seen before to the reinforcement of pre-existing societal biases embedded in training datasets, intellectual property rights violations, erosion of trust in online media, and malicious use in developing harmful or misleading content.

Existing regulatory and legal frameworks, constructed mainly with pre-AI eras in mind, consistently fall short to deal with the new complexities and subtleties introduced by AI-generated

 doi:10.20944/preprints202509.0271.v1

content (Shandilya et al., 2024). Traditional concepts of authorship, originality, and accountability are being challenged, and a rewriting of established conventions and the formulation of new rules of governance must take place (Saunders, 2023). Furthermore, the speeding pace of technological developments assures that any static regulatory solution will have an opportunity to quickly become outdated, thereby giving rise to the need for dynamic and prospective resolutions.

Recognizing this watershed juncture, this paper proposes an end-to-end and multi-layered governance model for ethical AI-generated content. Our approach exceeds a narrowly reactive or a purely technical solution, rather advocating an integrated solution which brings together ethical principles, executable processes, and supporting technical facilities. We believe that effective governance should emanate from collaborative efforts of various stakeholders, such as AI developers, deploying organizations, policy makers, legal specialists, ethicists, and end-users.

This paper begins with the outlining of the key ethical connotations that arise in the instance of AI-generated content. We delve into the transparency issues, examining the call for transparency in the extent to which AI is involved in content creation and the failures of current disclosure regimes. We frame the crucial issue of accountability, expounding on how accountability is to be attributed in the case of AI-generated content that is harmful or violates rights. The pervasive problem of reducing bias is discussed, considering how training data biases could become amplified in the generated output and how to design fair and equitable AI systems. We also discuss the complexities surrounding intellectual property rights for content generated by AI, particularly authorship and ownership. Finally, we identify key public risks of abuse and misinformation, as well as the risk of AI misuse and the imperatives of immediate protection of public trust. On this ethical basis, the framework provides actionable strategies along the lifecycle of AI. Recommendations on data governance are provided, emphasizing the need for good data sourcing, curation, and documentation. We discuss principles for ethical model development, advocating consideration of fairness, explainability, and transparency in generative AI model training and design. The framework further emphasizes the critical function of labeling and content provenance tracking, proposing mechanisms for transparent marking of AI-generated material and monitoring its source. Secondly, we look at the necessity to establish redress and oversight mechanisms like complaint channels on harmful content, appeals against rulings, and accountability.

By a systematic and adaptive structure, the paper attempts to be a guide to establishing the responsible development and deployment of AI-created content (Leong et al., 2024). Our goal is to assist in creating robust ethical safeguards that not only inhibit harm, but also facilitate the wise application of this revolutionary technology to advance society (Leon, 2025; Pöyhönen, 2024). The structure described seeks to be a proactive and dynamic instrument that can adapt to continuing developments in AI and shifting awareness of its ethical footprint. Finally, this paper contends that proactive and systemic governance is necessary to realize the vast potential of AI-created content while protecting intrinsic moral values and promoting a reliable digital environment (Basyoni et al., n.d.).

Objective of the Paper

The overall aim of this paper is to conceptualize an integrated, multi-layered ethical model of regulation of AI-generated content. The model needs to fill in gaps in existing ethical and regulatory frameworks by combining normative rules, operational standards, and technical facilitators in an integrated whole. This paper will specifically:

a) Define and dis-aggregate the ethical concerns posed by generative AI technologies, including misinformation, bias, and ownership disputes;

b) Synthesize peer-reviewed literature, policy reports, and international case studies to determine best practices and areas of governance deficit;

c) Propose a clear and structured framework connecting ethical theory with operational ICT mechanisms for responsible AI content generation

 doi:10.20944/preprints202509.0271.v1

d) Provide actionable recommendations to policymakers, AI developers, and users for the promotion of transparency, accountability, and human agency in AI-generated content.

In closing the ethical possibility gap and technological capability, the framework is designed to facilitate the creation of responsible, equitable, and accountable AI systems.

2. Background

The Artificial Intelligence field has experienced unprecedented growth in generative models' ability in the last few years (He et al., 2025; Kılınç & Keçecioğlu, 2024). Spurred on by advancements in deep architectures, such as transformer architectures, and large datasets, today's AI systems are capable of producing more sophisticated as well as natural-sounding material previously the exclusive domain of human writers (Poddar, 2024). This change is a paradigm shift from the earlier AI applications that previously focused on analysis, classification, and prediction. (Omankwu, 2023)

The journey to high-end AI-driven content creation has been one of evolution and increasing speed. Initial content creation through AI was either statistical or rule-based and used to generate content that was extremely simple and quite discernibly so compared to human content (Zhu et al., 2025). Improvements in the neural network, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), were the core for more complex forms of generation, such as in natural language processing and image generation (Dhruv & Naskar, 2020; Mienye et al., 2024).

The paradigm shift was in the year 2014 when Generative Adversarial Networks (GANs) came into the picture (Porkodi et al., 2023). GANs, in their competitive architecture of a generator and discriminator network, were shown to be capable of producing extremely realistic synthetic data like images, video, and audio (Hughes et al., 2021). It triggered more research and development activity, and generative models became even more capable.

Later years have witnessed the improvement of training large language models (LLMs) with the transformer architecture revolutionizing text generation (Annepaka & Pakray, 2025). GPT-3 and its later releases have exhibited a record high ability to produce coherent, contextually sound, and many times perceptually indistinguishable text across a huge array of styles and topics (Akhtarshenas et al., 2025). Similarly, improvements in diffusion models have witnessed an exponential increase in the quality and life-likeness of images and video produced by AI. (Kishnani, 2025)

The availability of these powerful generative AI technology has seen them increasingly used across sectors. In the creative arts, AI is used for content creation, idea generation, and artistic exploration. In advertising and marketing, AI is used to create personalized content as well as streamline campaign planning. In education and research, AI can generate artificial training data as well as simulate data. Even in science, AI is researched for synthesizing new molecules and materials (Sapkota et al., 2025).

But with this rapid evolution and wider use comes the growing acknowledgment of the fundamental ethical concerns such as AI-generated content. With these technologies more integrated into our Internet life, the potential for abuse and unintended consequence has become more difficult to ignore. Several key events and trends have served to alert us to the need to address these ethical concerns sooner rather than later:

a) The ability of AI to produce extremely realistic synthetic audio and video, popularly referred to as "deepfakes," has generated very serious concern for their deployment in spreading misinformation, influencing the public's opinions, and damaging reputations. Increasing sophistication and accessibility of the technology surrounding deepfakes is a serious digital media trust threat (George & George, 2023; Tiwari, 2025).

b) Generative AI models learning from big data used to train them will tend to replicate existing social biases by gender, race, and other protected characteristics. Such biases are thereby automatically or even intentionally amplified in the output, perpetuating discriminatory stereotypes and inequality (Kirk et al., 2021; Vázquez & Garrido-Merchán, 2024).

 doi:10.20944/preprints202509.0271.v1

c) Using copyrighted materials to train artificial intelligence models and generating outputs that share similarities with existing works are challenging intellectual property right, ownership, and fair use concerns. Lack of clear legal provisions to that end leaves creators and consumers of AI-generated content in limbo (Stransky, 2023).

d) As it is becoming increasingly difficult to differentiate between content created by humans and AI-created content, online information can be eroded in terms of trust (Ou et al., 2024). This "authenticity crisis" influences journalism, science communication, and democracy, and makes it increasingly hard for individuals to differentiate between authentic sources and artificial copies (Migisha & Hagström, 2025).

e) Generative AI can be used for malicious purposes by malicious intent, for example, the creation of spam, phishing, hate speech, and other types of harmful material in quantity (Shibli et al., 2024). The ability to automate and tailor these types of attacks makes them potentially more effective and more difficult to detect (Kurtović et al., 2025).

f) The "black box" of some sophisticated AI systems is hard to pierce, and it is impossible to determine how they produce material and assign blame when they cause harm. The secrecy makes it hard to recognize and confront bias and to hold anyone accountable for AI system impacts (Sayre & Glover, 2024).

These new ethical issues highlight the need for a holistic and anticipatory approach to regulating AI-generated content. Current legal and regulatory schemes, as best they offer some applicable principles, will usually be insufficient to capture the peculiar nature and richness of this technology. A specialized framework taking into account the peculiar ethical issues of AI-produced content is thus essential to ensuring responsible innovation and preventing potential harms. This essay attempts to contribute to this work of critique by mapping a multi-layered structure that can be applied to guide future ethical AI-generated content design, deployment, and utilization.

3. Literature Review

The rapid advancement of generative AI has spurred a growing body of literature that reflects on its technical promise, its social impact, and most significantly, its ethics. The following review scans significant research in the controversy over ethical AI-generated content regulation, citing existing frameworks, acknowledged challenges, and proposed solutions.

Early arguments were more inclined to focus on the philosophical implications of artificial creativity and authorship. Authors like Boden (1990) explored what creativity entails in computational systems, laying the groundwork for valuing what AI can do and what it cannot do with creating original content. As power generated by AI evolved, arguments shifted more towards practical matters, for instance, the social repercussions of media generated by AI.

Finding Key Ethical Issues

Most of the existing literature has been centered on identifying and analyzing the ethical problems of AI-generated content specifically. Floridi & Cowls, (2022)developed a building block for AI ethics, where core principles like transparency, justice, non-maleficence, and responsibility are directly applied to the regulation of the generated content.

Several researchers have raised the issue of bias especially. Crawford & Paglen, (2021) highlighted the inherent biases in large data sets and how such can be exaggerated in AI output leading to biased outcomes. Noble's "Algorithms of Oppression" (2018) also names the social harms that result from biased algorithmic systems as an issue that will directly resonate with AI-created text capable of disseminating adverse stereotypes (Harrison, 2021).

The potential for disinformation and misinformation facilitated through AI-generated content has been at the center of concern. Mustak et al., (2023) have coined "deep fakes" and the risks that

 doi:10.20944/preprints202509.0271.v1

their potential for undermining trust in information and destabilizing democratic institutions present. Vaccari & Chadwick, (2020) looked at the broader environment of online disinformation and how far AI can be seen to contribute to its production and circulation at scale.

Intellectual property rights of AI-generated content have also been the subject of widespread discussion. Dornis, (2021) issued the challenge of acknowledging AI as an inventor, shattering conventional thinking on authorship. Other scholars (e.g., Benzie & Montasari, 2022), however, have focused on the need to use current copyright legislation to address the unique problems of AI-generated works, most particularly with regards to ownership and originality.

Demand for transparency and explainability in AI systems, such as generative systems, has been emphasized by many researchers (e.g., Balasubramaniam et al., 2023). Knowing how AI generates content will aid in the identification of biases and mitigation thereof, rendering the technology accountable, and establishing trust in the technology.

Current Governance Proposals and Frameworks

There have been several attempts to advance general frameworks and guidelines for the ethical development and application of AI, some of which can be extended to generative AI. The European Union AI Act (2023) is a significant regulatory initiative, adopting a risk-based framework for AI regulation, with specific provisions for high-risk AI systems. While not specifically addressing generative AI, its principles of transparency and accountability apply.

Moral standards and industry-led initiatives have also emerged. Organizations like the Partnership on AI have developed best practices and principles for responsible AI. Research organizations and universities have offered ethical frameworks that value human control, equity, and public good.

Certain concrete proposals for the regulation of AI-generated content have also been made. Technical solutions, such as watermarking and provenance tracking (e.g., Rijsbosch et al., 2025), for detecting and tracing the origin of AI-generated content have been suggested by some authors. Others have called for labeling content and disclosure requirements to alert users to the use of the AI in creating the content (e.g., Gamage et al., 2025).

Legal scholars have discussed the possible use of existing legal doctrine and the need for novel legal paradigms to address issues such as liability for faulty information produced by AI (e.ge., Novelli et al., 2024). Algorithmic responsibility and designing redress mechanisms are recurring themes in the literature.

Gaps and Future Directions

Although the literature is growing, there remains room to complete some gaps in terms of ethical AI-generated content regulation. The majority of frameworks in existence are broad and do not provide specific guidance on the particular challenges posed by different forms of generated content (text, image, audio, video). Furthermore, the evolving character of AI technology implies that it requires ongoing change and calibration of regulation approaches.

There is more work to be done in the domain of socio-technical analyses of regulation of AI-generated content in order to learn about how technical devices can be integrated into social norms, law, and users' practices. There must be investigation of implications of educating users and media literacy as a method of managing a world with more and more AI-generated content.

Also to be underscored is the international and inter-jurisdictional nature of AI-generated content regulation. Owing to the internet's borderless character, harmonized ethical standards as well as regulatory approaches in various countries need to be defined.

This paper adds to existing literature by proposing a multi-layered and holistic approach that introduces ethical standards, procedural processes, and technological tools explicitly targeting the pitfalls of AI-generated content regulation. The paper aims to address some of the identified gaps by furnishing actionable strategies for various stakeholders and emphasizing the need for a nimble and team-based approach in this rapidly evolving field.

 doi:10.20944/preprints202509.0271.v1

4. Methodology

Process used in the research is constructive where developing an integrative framework to efficiently govern AI-generated content is being done. The process is structured and includes searching for the underlying problems, reorganizing the current knowledge base, conceptualizing fresh solutions, and combining them into an integrative and functional framework. Process development involves the following primary steps:

4.1. Integrative Literature Review and Ethical Analysis

After the above discussion, an overall review of all academic literature, industry reports, policy briefs, and ethical standards forming the foundation of AI ethics, media ethics, intellectual property law, and social implications of AI-generated content was conducted. The review aimed to describe the key ethical concerns, dominant models of governance, as well as gaps in dominant solutions.

From the learnings derived from the review of literature and current ethics guidelines (e.g., Cerratto Pargman & McGrath, 2021), the major ethical principles that could be considered to apply to content produced by AI were enumerated and assessed. The general principles of transparency, accountability, fairness, non-maleficence, and respect for privacy and autonomy are the primary principles on which the framework being put forward is rooted.

It was conducted in an exhaustive stakeholder analysis to understand all the stakeholders in the field of AI-generated content. The stakeholders were diverse and consisted of AI developers, deploying organizations (e.g., ad agencies, media agencies), end-users, policymakers, legal scholars, ethicists, and civil society organizations. Understanding the roles, responsibilities, and likely impacts on each of these sets of stakeholders was extremely crucial in developing an all-encompassing solution.

4.2. Structuring and Design of the Framework

Multi-Layered Solution is the way to go. Taking note of the multi-layered nature of the ethical issues, a multi-layered framework was designed. The solution transcends monocentric (e.g., technical or legal) and brings together various levels of intervention in a more secure, efficient system of government. The above-mentioned layers are:

a) Ethical Principles: Foundational principles constituting the core of the AI content generation, usage, and implementation.
b) Process Guidelines: best practices for everyday use and stakeholder tactics for all stakeholders across all stages of the AI lifecycle.
c) Technical Mechanisms: technical mechanisms and existing technology that can be utilized to facilitate ethical governance.
d) Iterative Refining: the initial draft framework was iteratively improved through internal debate and critical examination. This included testing for consistency, completeness, and feasibility of the proposed components and compatibility with accepted ethical principles and stakeholder needs.

4.3. Actionable Strategy and Mechanisms Development

Process Guidelines Development: Subsequent to the aforementioned ethical issues and principles, process guidelines were established step by step below for significant phases of the AI life cycle, i.e., data governance, model development, content creation, and deployment. The guidelines are intended to bridge the gap between overall ethical principles and actionable step for stakeholders.

Technical Mechanism Identification and Evaluation: New and current technical mechanisms of concern to ethical governance were compared and contrasted based on their effectiveness and

 doi:10.20944/preprints202509.0271.v1

limitations. These included content label technologies, origin tracking, bias detection and correction, and user remedy and feedback architectures.

Integration with Legal and Regulatory Implications: Although the commonality is ethical governance, legal and regulatory implications present and potential were considered so that the framework offered is compatible with law's large-scale architectures and capable of supplementing legislation law.

4.4. Framework Validation and Future Directions

Conceptual Validation: Conceptual validation of the conceptual framework illustrated was also done by examining its internal consistency, logical consistency, and consistency with current ethical theories and principles. This involved careful thinking through of how different layers and elements of the framework relate to each other. Future Work: Empirical Validation Possibilities

Although the present paper tries to set out theory building of the framework, future studies would be tasked with finding potential areas of empirical testing via expert interviews, pilot trials, and case studies to try out its practice effectiveness as well as implementability.

Flexibility and Future Work: The strategy outlines rapid development in the field of AI technology. To that effect, the strategy is flexible and shall be receptive to modifications as time progresses. The article further outlines key areas for future work, including developing some means to assess ethical AI-generated material and determining global harmonization of governance approaches.

Overall, the strategy employed in this instance is build and iterative with overwhelmingly thorough literature review, ethical analysis, stakeholder consideration, and systematic framework development. The goal is to create a complete, actionable, and adaptive framework that will be ready to contribute to the ethical and responsible management of the accelerated world of AI-content.

5. Proposed Framework

Enactment of the outlined methodology has resulted in the development of a three-layered model of ethical AI content governance. The model is made up of three interdependent layers: Ethical Principles, Process Guidelines, and Technical Mechanisms, all branching into a broad strategy for promoting responsible innovation and avoiding harm.

These three layers are not separate but interconnected and supportive. Ethical principles are the root values which guide the creation of process guidelines. Technical mechanisms provide the tools and techniques that may be employed for the application of such guidelines as well as for enforcing compliance with the origin ethical principles.

Conclusion of Results:

The designed multi-layered framework builds a systematic and adaptive model to the ethical regulation of AI-created content. By addressing ethical concerns at levels of core principles, professional practice, and facilitation technologies, this framework is designed to encourage responsible innovation, mitigate possible harms, and foster greater public trust in AI-generated material. The framework should be a dynamic document that can adapt to change in AI technology as well as the evolution of its ethics. Future work includes ongoing refinement and testing of this framework in application and through consultation with stakeholders.

 doi:10.20944/preprints202509.0271.v1

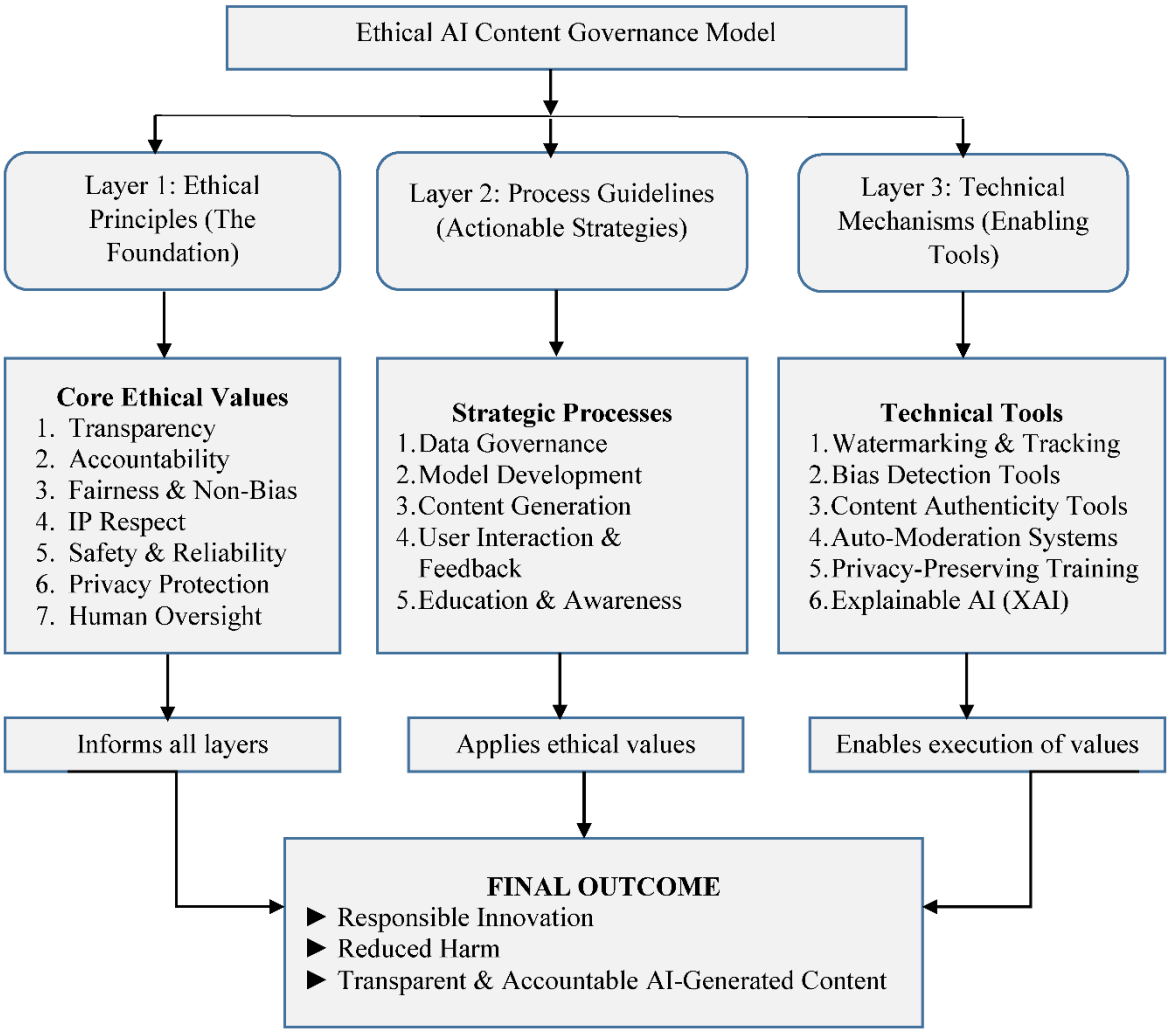

Figure 5.1. Proposed Framework (own designed).

6. Discussion

The multi-level format of this document sets out an integrated solution to the complex ethical landscape of AI-generated content. By combining necessary ethical standards with procedural guidelines and enabling technical provisions, it attempts to move beyond broken measures and build a less fractured environment for this new technology.

6.1. Strengths of the Proposed Framework

One of the advantages of this framework is that it is comprehensive. It recognizes that ethical governance is more than a technical, legal, or social issue, but rather a cross-road of all these. By addressing ethics at different levels, from principles to practical applications, the framework provides a more robust and elastic solution than single-shot suggestions.

The emphasis on stakeholder responsibility is another key advantage. The model explicitly assigns responsibility to various stakeholders in the AI environment, from developers and deployers to end-users and policymakers. Shared responsibility is the only form of regulation because it would be impossible for an individual person to address the multi-dimensional ethical issues resulting from content created by AI.

On top of that, flexibility of the framework is also crucial in the ever-evolving field of AI. Grounding the governance in well-established ethical principles, but providing the flexible process

 doi:10.20944/preprints202509.0271.v1

guidance and focusing on facilitation technologies, enables the framework to evolve when future developments and new ethical challenges occur without necessarily becoming obsolete all at once.

Including technical mechanisms as an applied value. Ethics standards and principles are needed but may rely on technology to function effectively. Listing and including technical methods relevant to the issue like watermarking, bias detection, and content verification, the framework outlines concrete ways of using ethical issues.

6.2. Solving Critical Ethical Problems:

The framework addresses very critical ethical issues mentioned in the literature review:

a) Transparency: Emphasis on open labelling and the disclosure of the AI role will resolve the transparency issue and aims to allow users to make responsible decisions.

b) Accountability: Through establishing responsibility for different stakeholders and suggesting redress systems, the framework allows for the inclusion of accountability in the marketplace for AI-generated content.

c) Bias Mitigation: Data governance and model development principles explicitly take care of the serious problem of bias as an express principle towards pre-emptive measures to ensure fairness and no discrimination.

d) Intellectual Property: In not necessarily providing end-of-line legal solutions, the framework does understand the importance of upholding respect for IP rights and promotes the development and utilization of technical solutions for tracing provenance.

e) Misuse and Disinformation: Safety and trustworthiness emphasis, coupled with content guidance and authenticity checking technology, should minimize the danger of improper use and disinformation propagation.

6.3. Comparison with Current Methods

The system surpasses and enhances existing approaches to AI ethics and regulation. It provides more concrete advice in the form of process guidelines and technical tools compared to principle-only schemes. While acknowledging the place of legal and regulatory codes (such as the EU AI Act), it gives a broader ethical guidance that can be used to inform and direct the creation and application of such legislation. Unlike purely technical solutions, this approach gives prominence to the underlying ethical norms and the fundamental importance of human scrutiny and societal values.

6.4. Limitations and Future Directions

All its goodness apart, even the proposed framework has its critics. Technical realizability of some guidelines and mechanisms will generally be at odds with technical feasibility, scalability, and user acceptance. For instance, the need for well-accepted and robust watermarking techniques and their global adoption presumes large industry players' cooperation and standardization.

Secondly, use and deployment of the ethical considerations could be dependent on context and therefore will elicit ongoing argument and debate among the stakeholders. The model provides a base, but ultimate deployment will rely on end to end interaction and tuning to specific usage situations and cultural norms.

Future research will have to cover a number of themes:

a) Empirical Validation: Demonstrating the effectiveness and impact of the framework in use through case studies and pilot exercises.

b) Metrics Development: Crafting measurable metrics for assessing the ethics of AI content and the efficacy of governance frameworks.

 doi:10.20944/preprints202509.0271.v1

c) International Harmonization: Exploring international collaboration and creation of harmonized ethical standards and governance frameworks to AI content across borders.

d) User Education and Media Literacy: Examining effective mechanisms for enhancing public awareness and media literacy that enable users' critical reception of AI content.

e) Resolution of the "Intent" Problem: Still trying to figure out how to deal with the ethical problem of AI-generated content where intent in its generation or dissemination is ill-intentioned, though the content may be innocent.

f) Reformed "Human-Created": As more and more creative endeavors are based on AI, the traditional human/AI creation dichotomy will be more and more unsettled, demanding new thinking about concepts like authorship and originality.

The proposed multi-layered model makes a central contribution to constructing a more responsible and ethical system of AI-generated content. Through the integration of ethical values, actionable processes, and facilitation technologies, it offers an integrated and responsive way of addressing the intricate challenges and actualizing the extraordinary potential of this transformative technology. Its effective operation, however, relies on ongoing cooperation, adaptability, and constant commitment to ethical values on the part of all parties. As AI advances, so must our frameworks for its ethical regulation, so that innovation benefits the wider interests of society.

7. Conclusion and Recommendations

The framework presented here sets forth a holistic response to the abundant ethical AI content environment. Prioritizing transparency, accountability, fairness, intellectual property respect, safety, privacy, and human agency, the framework is designed to establish a robust ethical foundation. The related process guidelines provide actionable steps for stakeholders along the AI lifecycle, from data management to interacting with users. Finally, the technical solutions presented fulfill the function of providing operational means of implementing these guidelines and the ethical foundation upon which they were based. The general outline allows for recognition that governance needs to be a participatory process and, most importantly, a multi-pronged approach with provisions for treating the ethical, procedural, and technological aspects of AI content.

Based on the model provided and discussion presented hereinabove, the following are suggested recommendations to the various stakeholders:

a) To AI Developers

- Embed Ethical Considerations by Design: Proactively integrate ethical norms and guidelines into the process of AI-based content model and application development.
- Prioritize Explainability and Transparency: Move towards model development and output generation with transparency and utilize transparent labeling approaches to keep the users informed about the use of AI.
- Aggressive Bias Mitigation: Employ robust data governance techniques and bias detection and mitigation technologies across the life cycle of model development.
- Invest in Safety Controls: Develop and deploy AI systems with top priority for safety and reliability without the creation of unsafe or misleading content.
- Put in Place Clear Mechanisms of Accountability: Define clear roles and responsibilities for development, deployment, and impact of AI content and put in place mechanisms of redress and control.
- Assure Constant Oversight and Review: Periodic observation of the operation and ethics of AI content generation systems and adapt governance as needed.

 doi:10.20944/preprints202509.0271.v1

- Formulate Adaptive and Principle-Based Regulations: Formulate the regulatory foundation that is adaptive to the fast-paced changing nature of AI but grounded on underlying ethical principles.
- Inviting Standardization and Interoperability: Encourage the creation and utilization of standardized marking, watermarking, and provenance tracking processes.
- Invite Global Cooperation: Encourage international dialogue and cooperation to develop aligned ethical principles and governance processes for AI-created content.
- Invest in Scholarship and Education: Support scholarship on the ethics of AI and encourage media literacy education and public awareness of AI-generated content.
- Consider Sector-Specific Rules: Develop sector-specific rules and standards for specific industries where AI-created content poses unique ethical challenges (e.g., media, healthcare, finance).

b) For End-Users:

- Advocate Critical Media Literacy: Promote critical literacy to examine the veracity and credibility of online information, e.g., being able to identify possible AI-created content.
- Enforce Disclosure and Labeling: Promote and enable transparent marking and tagging of AI-generated content.
- Provide Feedback and Report Abuse: Utilize discussion forums made available to provide feedback on content created by AI and report probable offending or unethical usage.
- Be Educated: Educate themselves periodically regarding the benefits and drawbacks of AI-generated content and its probable impact on society.

c) For Research Community:

- Create Technical Means of Ethical Regulation: Create efficient watermarking, bias detection, authenticity of content, and explainability AI methods.
- Explore the Socio-Technical Impacts: Study the overall social and cultural impacts of AI-generated material and develop models of comprehension and moderation.
- Enable Interdisciplinary Cooperation: Promote collaboration among AI researchers, ethicists, legal scholar, and social scientists to address ethics.

By embracing these standards and a common spirit of collaboration and vision, we can strive toward a future where responsibly and ethically the value of AI-generated content is reached with fewer harms and a safe digital world for all. Ethical AI governance is an adaptive process requiring constant conversation, evolution, and common commitment to essential human values.

References

Akhtarshenas, A., Dini, A., & Ayoobi, N. (2025). *ChatGPT or A Silent Everywhere Helper: A Survey of Large Language Models* (No. arXiv:2503.17403). arXiv. https://doi.org/10.48550/arXiv.2503.17403

Annepaka, Y., & Pakray, P. (2025). Large language models: A survey of their development, capabilities, and applications. *Knowledge and Information Systems*, *67*(3), 2967–3022. https://doi.org/10.1007/s10115-024-02310-4

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197. https://doi.org/10.1016/j.infsof.2023.107197

Bansal, G., Nawal, A., Chamola, V., & Herencsar, N. (2024). Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, *20*(11), 356:1-356:22. https://doi.org/10.1145/3689641

 doi:10.20944/preprints202509.0271.v1

Basyoni, L., Qayyum, A., Shaban, K., Elmahjub, E., Al-Ali, A., Halabi, O., & Qadir, J. (n.d.). *Generative AI-Driven Metaverse: The Promises and Challenges of AI-Generated Content*. Retrieved July 18, 2025, from https://www.authorea.com/doi/full/10.36227/techrxiv.174062953.37029023?commit=ff09d7ff31d869ea728a3256c6a3cd374afc52d5

Benzie, A., & Montasari, R. (2022). Artificial Intelligence and the Spread of Mis- and Disinformation. In R. Montasari (Ed.), *Artificial Intelligence and National Security* (pp. 1–18). Springer International Publishing. https://doi.org/10.1007/978-3-031-06709-9_1

Cerratto Pargman, T., & McGrath, C. (2021). Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research. *Journal of Learning Analytics*, *8*(2), 123–139.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY*, *36*(4), 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

Dhruv, P., & Naskar, S. (2020). Image Classification Using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A Review. In D. Swain, P. K. Pattnaik, & P. K. Gupta (Eds.), *Machine Learning and Information Processing* (pp. 367–381). Springer. https://doi.org/10.1007/978-981-15-1884-3_34

Dornis, T. W. (2021). *Of 'Authorless Works' and 'Inventions without Inventor' – The Muddy Waters of 'AI Autonomy' in Intellectual Property Doctrine* (SSRN Scholarly Paper No. 3776236). Social Science Research Network. https://doi.org/10.2139/ssrn.3776236

Floridi, L., & Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City* (pp. 535–545). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119815075.ch45

Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. K. (2025). Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–29. https://doi.org/10.1145/3706598.3713171

George, D. A. S., & George, A. S. H. (2023). Deepfakes: The Evolution of Hyper realistic Media Manipulation. *Partners Universal Innovative Research Publication*, *1*(2), Article 2. https://doi.org/10.5281/zenodo.10148558

Harrison, L. M. (2021). Algorithms of Oppression: How Search Engines Reinforce Racism by Safiya Umoja Noble (review). *College Student Affairs Journal*, *39*(1), 103–105.

He, R., Cao, J., & Tan, T. (2025). Generative artificial intelligence: A historical perspective. *National Science Review*, *12*(5), nwaf050. https://doi.org/10.1093/nsr/nwaf050

Hughes, R. T., Zhu, L., & Bednarz, T. (2021). Generative Adversarial Networks–Enabled Human–Artificial Intelligence Collaborative Applications for Creative and Design Industries: A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.604234

Kaur, H., Kishor Kumar Reddy, C., Manoj Kumar Reddy, D., & Hanafiah, M. M. (2025). Single Modality to Multi-modality: The Evolutionary Trajectory of Artificial Intelligence in Integrating Diverse Data Streams for Enhanced Cognitive Capabilities. In A. Singh & K. K. Singh (Eds.), *Multimodal Generative AI* (pp. 297–322). Springer Nature. https://doi.org/10.1007/978-981-96-2355-6_13

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., & Asano, Y. (2021). Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. *Advances in Neural Information Processing Systems*, *34*, 2611–2624. https://proceedings.neurips.cc/paper/2021/hash/1531beb762df4029513ebf9295e0d34f-Abstract.html

Kishnani, D. (2025). *The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/159096

Kılınç, H. K., & Keçecioğlu, Ö. F. (2024). Generative Artificial Intelligence: A Historical and Future Perspective. *Academic Platform Journal of Engineering and Smart Systems*, *12*(2), Article 2. https://doi.org/10.21541/apjess.1398155

Kurtović, H., Šabanović, E., Almisreb, A. A., Saleh, M. A., & Ismail, N. (2025). Exploring the Dark Side: A Systematic Review of Generative AI's Role in Network Attacks and Breaches. In B. Duraković, A. A. Almisreb, & J. Šutković (Eds.), *Recent Trends and Applications of Soft Computing in Engineering (RTASCE)—Sarajevo* (pp. 27–51). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-82881-2_3

Leon, L. D. (2025). Developing a Framework for Implementing AI in Education and Evaluating Its Use. In *Teaching and Learning in the Age of Generative AI*. Routledge.

Leong, W. Y., Leong, Y. Z., & Leong, W. S. (2024). Evolving Ethics: Adapting Principles to AI-Generated Artistic Landscapes. *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, 242–247. https://doi.org/10.1109/ICITRI62858.2024.10698905

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), Article 9. https://doi.org/10.3390/info15090517

Migisha, M.-G., & Hagström, T. (2025). *Seeing is Believing? : TikTok Users Perception, Trust, and Engagement with AI-Generated Visual vs. Human-Generated Visual Content*. https://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-105905

Morris, M. R. (2023). *Scientists' Perspectives on the Potential for Generative AI in their Fields* (No. arXiv:2304.01420). arXiv. https://doi.org/10.48550/arXiv.2304.01420

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, *154*, 113368. https://doi.org/10.1016/j.jbusres.2022.113368

Novelli, C., Floridi, L., Sartor, G., & Teubner, G. (2024). *AI as Legal Persons: Past, Patterns, and Prospects* (SSRN Scholarly Paper No. 5032265). Social Science Research Network. https://doi.org/10.2139/ssrn.5032265

Omankwu, O. C. (2023). Artificial Intelligence a Paradigm Shift in Healthcare Sector Past, Present and Future Prospects. *NIPES - Journal of Science and Technology Research*, *5*(3). https://doi.org/10.5281/zenodo.8348838

Ou, M., Zheng, H., Zeng, Y., & Hansen, P. (2024). Trust it or not: Understanding users' motivations and strategies for assessing the credibility of AI-generated information. *New Media & Society*, 14614448241293154. https://doi.org/10.1177/14614448241293154

Poddar, H. (2024). From neurons to networks: Unravelling the secrets of artificial neural networks and perceptrons. In *Deep Learning in Engineering, Energy and Finance*. CRC Press.

Porkodi, S. P., Sarada, V., Maik, V., & Gurushankar, K. (2023). Generic image application using GANs (Generative Adversarial Networks): A Review. *Evolving Systems*, *14*(5), 903–917. https://doi.org/10.1007/s12530-022-09464-y

Pöyhönen, M. (2024). *Human-AI Integration in Long-Established Organizations*. https://aaltodoc.aalto.fi/handle/123456789/127134

Rijsbosch, B., Dijck, G. van, & Kollnig, K. (2025). *Adoption of Watermarking Measures for AI-Generated Content and Implications under the EU AI Act* (No. arXiv:2503.18156). arXiv. https://doi.org/10.48550/arXiv.2503.18156

Sapkota, R., Raza, S., Shoman, M., Paudel, A., & Karkee, M. (2025). *Multimodal Large Language Models for Image, Text, and Speech Data Augmentation: A Survey* (No. arXiv:2501.18648). arXiv. https://doi.org/10.48550/arXiv.2501.18648

Saunders, D. (2023). *Authorship and Copyright*. Routledge. https://doi.org/10.4324/9781003370437

 doi:10.20944/preprints202509.0271.v1

Sayre, M. A., & Glover, K. (2024). Machines Make Mistakes Too: Planning for AI Liability in Contracting. *Case Western Reserve Journal of Law, Technology and the Internet*, *15*, 357.

Shandilya, S. K., Datta, A., Kartik, Y., & Nagar, A. (2024). Role of Artificial Intelligence and Machine Learning. In S. K. Shandilya, A. Datta, Y. Kartik, & A. Nagar (Eds.), *Digital Resilience: Navigating Disruption and Safeguarding Data Privacy* (pp. 313–399). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53290-0_6

Shibli, A. M., Pritom, M. M. A., & Gupta, M. (2024). AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns. *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–6. https://doi.org/10.1109/ISDFS60797.2024.10527300

Sieg, S. (2023). *AI Art and Its Limitations: Narrow Intelligence and Visual Indulgence*. https://soar.suny.edu/handle/20.500.12648/14056

Stransky, C. (2023). *A legal approach to whether ai generated content should be protected under copyright*. https://hdl.handle.net/10539/38576

The effectiveness of the combined problem-based learning (PBL) and case-based learning (CBL) teaching method in the clinical practical teaching of thyroid disease | BMC Medical Education. (n.d.). Retrieved June 29, 2025, from https://link.springer.com/article/10.1186/s12909-020-02306-y

Tiwari, S. (2025). *The Impact of Deepfake Technology on Cybersecurity: Threats and Mitigation Strategies for Digital Trust* (SSRN Scholarly Paper No. 5259359). Social Science Research Network. https://doi.org/10.2139/ssrn.5259359

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Vázquez, A. F. de C., & Garrido-Merchán, E. C. (2024). *A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence* (No. arXiv:2407.01556). arXiv. https://doi.org/10.48550/arXiv.2407.01556

Zhu, C., Cui, L., Tang, Y., & Wang, J. (2025). *The Evolution and Future Perspectives of Artificial Intelligence Generated Content* (No. arXiv:2412.01948). arXiv. https://doi.org/10.48550/arXiv.2412.01948