

Enhancing Trust and Transparency in AI-Generated Visual Content

Problem Statement: Bridging the Trust Gap in AI-Generated Visual Content: Towards Transparent and Authentic Media Production

Over the past few years, generative artificial intelligence has fundamentally changed how visual content is created. Tools like DALL-E, Midjourney, and Sora can now produce images and videos that are often indistinguishable from real photographs or professionally shot footage (ERTAN & SAHIN CEKEN, 2023). While this technological leap is exciting, it has also created a serious problem: audiences are losing the ability to tell what is real and what is machine-made.

I found this particularly concerning when reviewing the literature for my previous assignment. What stood out to me is that multiple studies—whether they focused on TikTok engagement, YouTube comments, or Gen Z interviews—all pointed to the same issue: people feel uneasy when they cannot trust what they are seeing (Tran, 2025; Duong & Vo, 2025; Calli & Calli, 2025). This "authenticity crisis" is not just a philosophical concern; it has real consequences for journalism, advertising, education, and even personal relationships.

Several solutions have been proposed to address this problem. Some researchers advocate for content labeling systems that clearly mark AI-generated material (Reddy, 2022). Others suggest watermarking techniques or embedding metadata that can be traced back to the source. Explainable AI (XAI) frameworks have also been developed to make AI decision-making more transparent (Zhang & Lu, 2021). However, from my perspective, these solutions are still fragmented and inconsistent. There is no unified framework that considers both technical transparency and human audience perception simultaneously. Moreover, most existing approaches have been tested in controlled environments rather than real-world media platforms where people actually consume content (Ghassemi et al., 2022).

This gap is what motivates my research. I believe that trust in AI-generated content cannot be achieved through technical fixes alone. We also need to understand how different audiences—especially younger generations who grew up with social media—perceive transparency and authenticity. Therefore, my research aims to address the following problem:

"How can we develop a comprehensive and practically implementable framework that enhances trust and transparency in AI-generated visual content, ensuring audience awareness and authenticity while preserving the creative and practical benefits of generative AI?"

Research Objectives

To address the problem stated above, this research will pursue the following five objectives. I have structured them to move logically from understanding the problem to proposing and validating a practical solution:

1. To comprehensively analyze the existing challenges in trust and transparency of AI-generated visual content

This objective focuses on building a solid foundation by identifying what we already know about trust issues in AI-generated media. Specifically, I aim to:

- Identify and categorize the key challenges audiences face when encountering AI-generated images and videos, based on existing literature.
- Examine how issues of authenticity, bias, and platform-specific transparency mechanisms affect user trust.
- Synthesize findings from different audience groups (e.g., Gen Z, general social media users, professional content creators) to identify common patterns and differences.

Justification: Without a clear understanding of the current challenges, any proposed solution would be guesswork. This objective ensures my research is grounded in real issues rather than assumptions.

2. To evaluate existing solutions, including content labeling, watermarking, and Explainable AI (XAI) frameworks

Once the challenges are mapped, the next step is to look at what has already been tried. This objective involves:

- Critically assessing the strengths and weaknesses of current transparency mechanisms such as visible labels, invisible watermarks, and metadata embedding.
- Evaluating how well Explainable AI (XAI) methods—originally designed for general AI systems—can be adapted to generative visual content.
- Exploring alternative validation approaches proposed by scholars such as Ghassemi et al. (2022), who argue that current XAI methods have significant imperfections.

Justification: This ensures that my proposed framework does not reinvent the wheel but instead builds upon and improves existing approaches.

3. To propose a comprehensive framework for enhancing trust in AI-generated visual content

This is the core contribution of my research. Based on the analysis from Objectives 1 and 2, I will develop a framework that:

- Integrates technical transparency mechanisms (labels, watermarks, metadata) with human-centered design principles that consider how audiences actually perceive and respond to transparency cues.
- Addresses the specific needs of different platforms (social media, advertising, cinematic production) and audience groups.
- Aligns with ethical considerations, including informed consent, user autonomy, and fairness.

Justification: This objective transforms the research from a purely analytical exercise into a practical contribution that can be used by content creators, platform designers, and policymakers.

4. To validate the proposed framework in diverse visual content domains

A framework is only useful if it actually works in the real world. Therefore, this objective involves:

- Testing the framework across different types of visual content platforms, including social media (TikTok, YouTube, Instagram), digital advertising, and cinematic production.
- Gathering feedback from practitioners, content creators, and media professionals to refine the framework based on practical experience.
- Assessing the generalizability of the framework and identifying any domain-specific adjustments that may be needed.

Justification: This ensures that my proposed framework is not just theoretical but has practical value and can be adapted to different contexts.

5. To provide practical recommendations for the implementation of the proposed framework

The final objective ensures that my research has a real-world impact. I will:

- Offer actionable guidance for organizations, content creators, platform developers, and policymakers on how to integrate the framework into their existing workflows.
- Provide a roadmap for implementation that considers resource constraints, technical requirements, and potential challenges.
- Contribute to the ongoing academic and industry discourse on responsible and trustworthy AI development and deployment.

Justification: Without practical recommendations, even the best framework may remain unused.

This objective bridges the gap between academic research and real-world practice.

Literature Review

1. The Importance of Trust and Transparency in AI-Generated Visual Content

The rapid advancement of generative AI has transformed visual media production, but it has also created a significant authenticity crisis (ERTAN & SAHIN CEKEN, 2023). Audiences now struggle to distinguish between real and AI-generated images and videos, which directly undermines trust in media.

Key findings from recent audience studies highlight several patterns:

- AI-generated content is perceived as efficient but lacks emotional depth compared to human-made content (Tran, 2025).
- Generation Z users value emotional authenticity and narrative coherence, and they express discomfort when authorship is undisclosed (Duong & Vo, 2025).
- Public discourse on platforms like YouTube reveals emotional and cognitive anxiety, including the "uncanny valley" effect and epistemic uncertainty about what is real (Calli & Calli, 2025).

Reddy (2022) and Zhang & Lu (2021) both emphasize that transparency is essential for building trust, whether in clinical AI or media applications. Without transparency, users become skeptical and disengaged.

2. Current Challenges and Limitations

Despite progress in AI capabilities, several persistent challenges hinder trust and transparency. Key issues include:

- Bias and algorithmic accountability, which raise concerns about fairness and discrimination (Goralski & Tan, 2020).
- Opacity, autonomy, and inscrutability, which make it difficult for users to understand or challenge AI decisions (Berente et al., 2021).
- Difficulty in explaining visual AI decisions, as images are processed emotionally and quickly, making transparency even more critical compared to text-based systems (Ghassemi et al., 2022).
- Fragmented transparency mechanisms that vary across different platforms, leading to inconsistent user experiences (Calli & Calli, 2025).

From my perspective, these challenges are especially critical in visual media because people process images quickly and emotionally. When they see a realistic image, they instinctively trust it—which makes fake content particularly dangerous.

3. Existing Solutions: XAI and Trustworthy AI Frameworks

Several solutions have been proposed to address transparency issues, each with distinct strengths and limitations:

Explainable AI (XAI)

- The goal is to make AI decision-making processes understandable to users.
- For example, Reddy (2022) advocates for XAI in clinical settings to uphold informed consent.
- However, Ghassemi et al. (2022) argue that many XAI methods are imperfect and can provide misleading explanations, calling for alternative validation approaches.

Trustworthy AI (TAI)

- Thiebes, Lins, and Sunyaev (2021) propose five foundational principles: beneficence, non-maleficence, autonomy, justice, and explicability.
- While useful, this framework is general and does not specifically address the unique challenges of visual content where emotional and perceptual factors dominate.

Content Labeling and Watermarking

- These techniques can clearly signal AI-generated material.
- However, adoption is inconsistent across platforms, and users may ignore or distrust labels if not implemented thoughtfully.

4. Gaps in the Literature

Based on my review, I identified three main gaps that this research aims to address:

1. Limited focus on visual content: Most trust and transparency research addresses general AI systems, not generative visual media specifically.
2. Lack of a unified framework: Existing solutions are fragmented and platform-specific. There is no comprehensive model integrating technical transparency with human-centered audience perception.
3. Insufficient real-world validation: Most solutions have been tested in controlled environments rather than actual media platforms where users consume content daily.

Stone et al. (2022) emphasize the need for effective governance mechanisms and interdisciplinary approaches, which reinforces my belief that trust in AI-generated content requires more than just technical fixes.

5. Relevance to This Research

The literature reviewed confirms that trust and transparency in AI-generated visual content are urgent and underexplored issues. While XAI and TAI frameworks offer useful starting points, they require adaptation to the specific context of visual media.

This research directly addresses the identified gaps by:

- Focusing specifically on AI-generated visual content.
- Proposing a comprehensive framework that combines technical transparency with audience perception.
- Validating the framework across diverse real-world platforms (social media, advertising, cinematic production).

References

- Berente, Nicholas, Gu, Bin, Recker, Jan, & Santhanam, Radhika. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Calli, Levent, & Calli, Busra Alma. (2025). Recoding Reality: A Case Study of YouTube Reactions to Generative AI Videos. *Systems*, 13(10), 925. <https://www.mdpi.com/2079-8954/13/10/925>
- Duong, Hoai Lan, & Vo, Thi Kim Oanh. (2025). How do young users perceive and respond to AI-generated short-form videos? An exploration of Generation Z's perceptions, emotional responses, and trust in AI-created video on social media platforms. *International Journal of Human-Computer Studies*. <https://www.sciencedirect.com/science/article/abs/pii/S1071581925002174>
- ERTAN, Ipek Fatma, & SAHIN CEKEN, Kubra. (2023). Generative Artificial Intelligence and Visual Communication: A Systematic Review of Transformations in Image Production, Audience Reception, and Meaning-Making in Digital Media. *Review of Communication Research*, 11, 39. <https://doi.org/10.52152/RCR.V11.39>
- Ghassemi, Marzyeh, Oakden-Rayner, Luke, & Beam, Andrew L. (2022). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 4(11), e806–e811. [https://doi.org/10.1016/S2589-7500\(22\)00107-1](https://doi.org/10.1016/S2589-7500(22)00107-1)
- Goralski, Margaret A., & Tan, Tay Keong. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1), 100330. <https://doi.org/10.1016/j.ijme.2019.100330>
- Reddy, Sandeep. (2022). Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4(4), e214–e215. [https://doi.org/10.1016/S2589-7500\(22\)00029-2](https://doi.org/10.1016/S2589-7500(22)00029-2)
- Stone, Peter, Brooks, Rodney, Brynjolfsson, Erik, Calo, Ryan, Etzioni, Oren, Hager, Greg, Hirschberg, Julia, Kalyanakrishnan, Shivaram, Kamar, Ece, Kraus, Sarit, Leyton-Brown, Kevin, Parkes, David, Press, William, Saxenian, AnnaLee, Shah, Julie, Tambe, Milind, & Teller, Astro. (2022). Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence. *arXiv:2211.06318*. <https://doi.org/10.48550/arXiv.2211.06318>
- Thiebes, Scott, Lins, Sebastian, & Sunyaev, Ali. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(3), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- Tran, Van Cong. (2025). *Bridging Efficiency and Emotion: Evaluating User Engagement with AI-Generated and Human-Generated TikTok Videos Through Media Richness and Uses and Gratifications Theories*. University of Economics Ho Chi Minh City. <https://digital.lib.ueh.edu.vn/handle/UEH/75469>
- Zhang, Ruihan, Yu, Borou, Min, Jiajian, Xin, Yetong, Wei, Zheng, Shi, Juncheng Nemo, Huang, Mingzhen, Kong, Xianghao, Liu Xin, Nix, Jiang, Shanshan, Bahuguna, Praagya, Huang, Peiwen, Chen, Na, Liu, Xuanxuan, & Rao, Anyi. (2025). Generative AI for Film Creation: A Survey of Recent Advances. *arXiv*. <https://arxiv.labs.arxiv.org/html/2504.08296>
- Zhang, Chen, & Lu, Yang. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23, 100224. <https://doi.org/10.1016/j.jii.2021.100224>
-