The Authenticity Spectrum Framework: Classifying Deepfake and Generative AI Risks in Synthetic Media

Francis Martinson
Department of Computer Science,
North Dakota State University, USA
ORCID: 0009-0007-2235-2516

ABSTRACT

The rapid advancement of generative artificial intelligence technologies—including large language models, diffusion models, and deepfake systems—has created unprecedented capabilities for synthetic media generation while simultaneously enabling novel vectors for fraud, disinformation, and exploitation. However, existing governance frameworks fail to differentiate between beneficial applications such as AI-generated marketing content, accessibility tools, and creative expression, and harmful uses including identity fraud, non-consensual intimate imagery, and political disinformation. This paper introduces the **Authenticity Spectrum Framework (ASF)**, a novel five-level classification system for AI-generated content based on three dimensions: disclosure transparency, creator intent, and harm potential. Building on prior research examining exploitation architectures in gaming systems [1] and smartphone vulnerabilities [2], the ASF extends dual-use technology analysis to synthetic media governance. Through systematic analysis of current synthetic media platforms including AI avatar generators, video generation models, and voice cloning services, we demonstrate the framework's practical application to real-world governance challenges. The framework provides regulators, platform operators, and AI developers with a standardized taxonomy for risk assessment aligned with emerging requirements under the EU AI Act, NIST AI Risk Management Framework, and FTC guidelines.

General Terms

AI Governance, Security, Ethics, Regulation

Keywords

Deepfakes, Generative AI, Synthetic Media, AI Governance, Large Language Models, Risk Classification, EU AI Act

1. INTRODUCTION

The emergence of generative artificial intelligence has fundamentally transformed digital content creation across all media modalities. Large language models (LLMs) such as GPT-4, Claude, Gemini, and LLaMA generate human-quality text across diverse domains including creative writing, technical documentation, and conversational interaction. Diffusion models including Stable Diffusion, DALL-E 3, Midjourney, and Flux produce photorealistic images from natural language descriptions with increasing fidelity and controllability. Video generation systems—Runway Gen-3, Kling 2.0, Pika, Luma Dream Machine, and OpenAI's Sora—now create synthetic video sequences that challenge human perceptual capabilities and approach cinematic quality. Voice cloning services including ElevenLabs, Resemble AI, and Speechify enable generation of synthetic speech from minimal training samples, often less than sixty seconds of reference audio.

The dual-use nature of these technologies presents a fundamental governance paradox familiar from other technology domains. Prior research has demonstrated how game engine vulnerabilities enable both legitimate modding communities and malicious exploitation through identical technical mechanisms, with the same DLL injection techniques serving creative modification and competitive cheating [1]. Similarly, smartphone camera architectures designed for user convenience simultaneously enable surveillance attacks that capture biometric data without consent through APIs intended for legitimate photography applications [2]. Synthetic media technologies exhibit the same dual-use characteristic with heightened stakes: the diffusion model generating marketing imagery operates on identical principles to systems producing non-consensual intimate images; the voice cloning service enabling audiobook production utilizes the same neural architectures facilitating voice phishing attacks.

Current regulatory frameworks, including the European Union Artificial Intelligence Act and proposed United States legislation, struggle to address this duality. Regulators often default to binary permitted/prohibited classifications that fail to capture the nuanced spectrum of applications and associated risks. This approach creates dual failure modes: overly restrictive regulations that impede beneficial innovation in accessibility, creative expression, and commercial efficiency, alongside insufficient protections against genuinely harmful applications that exploit regulatory gaps.

This paper addresses this governance gap by introducing the Authenticity Spectrum Framework (ASF)—a structured classification system that enables proportionate governance responses calibrated to actual risk levels. The framework contributes to the emerging field of AI governance by providing a common taxonomy for researchers, policymakers, platform operators, and practitioners working on synthetic media regulation.

1.1 Research Objectives

This research pursues four primary objectives. First, to develop a theoretically grounded classification framework for synthetic media that captures the full spectrum of applications from beneficial to harmful. Second, to operationalize the framework through concrete classification criteria applicable to current generative AI platforms and use cases. Third, to demonstrate alignment between the proposed framework and existing regulatory requirements, enabling practical compliance implementation. Fourth, to provide guidance for organizations deploying synthetic media technologies in assessing and managing associated risks.

1.2 Key Definitions

Synthetic Media refers to digital content wholly or partially generated by artificial intelligence systems, including content produced by large language models, diffusion models, GANs,

VAEs, NeRFs, and audio synthesis systems. Deepfake refers specifically to synthetic media depicting identifiable individuals saying or doing things they did not actually do, typically created using deep learning techniques. The Authenticity Spectrum represents the continuum from transparently artificial to maliciously deceptive content.

2. BACKGROUND AND RELATED WORK

2.1 Technical Evolution of Synthetic Media

The technical capabilities enabling current synthetic media emerged through several distinct research trajectories. Generative Adversarial Networks, introduced by Goodfellow et al. in 2014, established the adversarial training paradigm where generator and discriminator networks compete to produce increasingly realistic outputs [3]. This architecture enabled early deepfake systems and remains influential in current face-swapping applications. The transformer architecture, introduced for natural language processing in 2017, revolutionized text generation and subsequently influenced all modalities of synthetic media. Diffusion models, particularly latent diffusion as implemented in Stable Diffusion and DALL-E, achieved breakthrough image generation quality in 2022 [4]. Video diffusion models extended these techniques to temporal sequences with rapid capability improvements throughout 2023-2025.

2.2 Detection-Focused Research

Early academic response to synthetic media focused primarily on detection. Rossler et al. established the FaceForensics++ benchmark for deepfake detection in 2019, providing standardized datasets and evaluation protocols [5]. However, detection approaches face fundamental limitations that constrain their governance utility. Detection operates in a perpetual arms race with generation capabilities—a dynamic observed across security domains including anti-cheat systems in gaming [1] and malware detection in mobile devices [2]. As detection systems identify specific artifacts, generation systems are optimized to eliminate them. Furthermore, detection provides only binary classification (synthetic/authentic) without capturing the nuanced risk factors that should inform governance responses.

2.3 Governance-Focused Research

Governance-focused research has emerged more recently as the limitations of detection-only approaches became apparent. Chesney and Citron analyzed deepfakes as threats to privacy, democracy, and national security [6]. The Partnership on AI published a Framework for Responsible Practices in Synthetic Media [7]. Vaccari and Chadwick examined deepfakes' impact on political trust and democratic processes [8]. The NIST AI Risk Management Framework provides general AI governance guidance applicable to synthetic media systems [9]. The EU AI Act establishes disclosure requirements for AI-generated content [10]. The FTC has issued guidance on AI-generated content in advertising [11]. However, these frameworks lack granular risk classification enabling proportionate responses.

2.4 Research Gap

Existing frameworks address synthetic media governance through either technical detection (which fails to capture risk factors) or broad regulatory categories (which fail to enable proportionate responses). This paper proposes a structured classification framework specifically designed for synthetic media risk assessment, drawing on lessons from exploitation architecture analysis in related technology domains [1][2].

3. METHODOLOGY

3.1 Framework Development Approach

The Authenticity Spectrum Framework was developed through iterative analysis combining regulatory document review, technical capability assessment, and use case cataloging. Initial dimensions were derived from examination of existing regulatory language in the EU AI Act, FTC guidance, and academic governance literature. These dimensions were refined through systematic analysis of documented synthetic media incidents, including both beneficial applications and harm cases reported in academic literature, news media, and regulatory filings.

3.2 Use Case Analysis

Systematic analysis of current synthetic media platforms informed framework application guidance. Platforms were selected to represent major technology categories: AI avatar generators (Synthesia, HeyGen, D-ID), video generation models (Runway Gen-3, Kling, Pika, Sora), voice cloning services (ElevenLabs, Resemble AI), image generators (Midjourney, DALL-E, Stable Diffusion), and synthetic UGC platforms (MakeUGC, Creatify, Arcads). For each platform, documented use cases were classified according to framework dimensions.

3.3 Validation Approach

Framework validation employed expert review methodology. Draft framework specifications were circulated to practitioners in AI governance, content moderation, and regulatory compliance roles. Feedback was incorporated through two revision cycles, with particular attention to operational feasibility and alignment with existing compliance workflows.

4. THE AUTHENTICITY SPECTRUM FRAMEWORK

The Authenticity Spectrum Framework classifies synthetic content across five risk levels based on three primary dimensions: disclosure transparency, creator intent, and harm potential. This multi-dimensional approach enables more precise risk assessment than single-factor classifications while remaining operationally tractable.

4.1 Framework Dimensions

Dimension 1 (Disclosure Transparency) measures the degree to which synthetic content is identified as AI-generated to its intended audience. This dimension ranges from explicit labeling through visible watermarks, metadata tags, and disclosure statements, through contextual indicators where platform context implies synthetic origin, to active concealment where synthetic origins are deliberately obscured or misrepresented.

Dimension 2 (Creator Intent) examines the purpose behind synthetic content generation as reasonably inferable from context, stated purposes, and deployment patterns. Intent categories span a spectrum: creative expression, accessibility applications, education, entertainment, commercial promotion, satire, persuasion, deception for gain, and deliberate harm or exploitation.

Dimension 3 (Harm Potential) assesses probable negative consequences across three categories: individual harms (identity theft, reputation damage, financial losses), collective harms (erosion of epistemic trust, democratic manipulation, social division), and economic harms (fraud, market manipulation, IP violations).

4.2 Classification Levels

The framework defines five classification levels representing positions along the authenticity spectrum from fully transparent to weaponized deception. Each level implies different governance responses. Table 1 presents the complete classification scheme.

Table 1. ASF Classification Levels

Level	Name	Disclosure	Intent	Examples	Governance
1	Transparent	Explicit labeling	Creative/educational	AI art with attribution	Permissive
2	Contextual	Platform-implied	Entertainment	Film VFX, game NPCs	Standard oversight
3	Ambiguous	Undisclosed	Commercial	Synthetic UGC ads	Active monitoring
4	Deceptive	Actively concealed	Financial gain	Fake reviews, astroturfing	Prohibition
5	Malicious	Weaponized	Harm/exploitation	NCII deepfakes, fraud	Criminal liability

5. RESULTS: PLATFORM ANALYSIS

Systematic analysis of synthetic media platforms reveals distinct risk profiles across technology categories. This section presents detailed evaluation results demonstrating the framework's practical applicability. Table 2 summarizes classification results for representative platforms across five major categories.

Table 2. Platform Analysis Results

Platform Category	Representative Platforms	ASF Level	Primary Risk Factors
AI Avatar Generators	Synthesia, HeyGen, D-ID	Level 2-3	Undisclosed commercial use
Synthetic UGC Platforms	MakeUGC, Creatify, Arcads	Level 3-4	Testimonial deception
Voice Cloning Services	ElevenLabs, Resemble AI	Level 1-5	Vishing, impersonation
Video Generation	Runway, Sora, Kling, Pika	Level 1-4	Event fabrication
Image Generators	Midjourney, DALL-E, SD	Level 1-3	Misattribution

5.1 AI Avatar and Synthetic UGC Analysis

Platforms such as MakeUGC, Creatify, Arcads, Synthesia, and HeyGen represent a paradigm shift in digital marketing and corporate communications. These services enable generation of synthetic user-generated content—AI avatars delivering scripted testimonials, product demonstrations, and promotional messages visually indistinguishable from recordings of human presenters. When combined with video generation models, marketers can produce unlimited advertising variations without human creators.

Analysis reveals these platforms typically occupy Level 3 (Ambiguous) when deployed without explicit disclosure—the content is not labeled as AI-generated, but intent is commercial rather than deceptive. However, risk escalates to Level 4 (Deceptive) when deliberately designed to appear as authentic customer reviews, particularly for products with safety implications, financial services, or healthcare applications. Platform operators face dual-use risk management challenges mirroring observations from game security research [1]. The same avatar technology enabling efficient training video production can generate fraudulent testimonials for investment scams.

5.2 Voice Cloning Analysis

Voice cloning services generate synthetic speech from minimal training data—often less than sixty seconds of reference audio for basic cloning. Analysis reveals the widest classification range (Level 1-5) of any technology category, demonstrating that identical technical capabilities produce vastly different risk profiles depending on deployment context. Legitimate applications including audiobook narration, accessibility features for individuals who have lost their voices, and content localization classify as Level 1-2 with appropriate disclosure. However, identical capabilities enable Level 5 applications including vishing attacks impersonating family members, CEO fraud schemes, and synthetic evidence fabrication. This finding parallels smartphone camera research showing identical APIs serve both photography and surveillance [2].

5.3 Video Generation Analysis

Text-to-video and image-to-video models including Runway Gen-3, Pika, Kling 2.0, and Sora generate increasingly photorealistic video from text prompts or reference images. Analysis reveals classification spanning Level 1-4 depending on application context: creative expression and prototyping (Level 1-2), commercial advertising production (Level 2-3), and fabrication of events or statements that never occurred (Level 4-5). The rapid capability improvements observed throughout 2024-2025 increase the urgency for governance frameworks that can accommodate evolving technical capabilities.

6. REGULATORY ALIGNMENT ANALYSIS

Analysis demonstrates strong alignment between ASF classification levels and existing regulatory requirements, enabling practical compliance implementation. Table 3 maps ASF levels to specific regulatory provisions.

Table 3. Regulatory Framework Alignment

ASF Level	EU AI Act Alignment	NIST AI RMF	FTC Compliance
Level 1-2	Compliant (Art. 52 satisfied)	GOVERN, MAP	Compliant
Level 3	Potential violation	MAP, MEASURE, MANAGE	Review required
Level 4-5	Non-compliant (prohibited)	MANAGE (prohibit)	Section 5 violation

6.1 EU AI Act Alignment

The EU AI Act (Article 52) establishes transparency requirements for AI-generated content, mandating disclosure when content depicts persons appearing to say or do something they have not [10]. ASF Levels 1-2 satisfy these requirements through explicit or contextual disclosure. Level 3 indicates potential non-compliance requiring remediation through enhanced disclosure mechanisms. Levels 4-5 represent clear violations requiring enforcement action.

6.2 NIST AI RMF Integration

The NIST AI Risk Management Framework emphasizes proportionate risk management across four functions: GOVERN, MAP, MEASURE, and MANAGE [9]. The ASF operationalizes this principle for synthetic media: Levels 1-2 require minimal intervention focused on maintaining disclosure mechanisms (GOVERN function); Level 3 warrants active monitoring and risk assessment (MAP, MEASURE functions); Levels 4-5 require prohibition and enforcement (MANAGE function).

6.3 FTC Compliance

FTC guidance addresses deceptive practices in AI-generated advertising under Section 5 of the FTC Act [11]. The ASF provides classification criteria for determining when synthetic UGC crosses the threshold from permissible commercial speech (Level 2-3) to deceptive advertising (Level 4). Synthetic testimonials presenting AI avatars as real customers constitute Level 4 regardless of product quality claims.

7. DISCUSSION

7.1 Framework Contributions

The ASF contributes to AI governance literature by providing a standardized taxonomy for synthetic media risk assessment that balances precision with operational tractability. The three-dimensional classification approach captures risk factors that single-factor frameworks miss. Content with identical technical characteristics can occupy different ASF levels based on disclosure status and deployment context, enabling nuanced governance that recognizes legitimate uses of powerful capabilities while preventing weaponization.

7.2 Limitations and Future Work

Several limitations constrain the current framework. Intent assessment can be difficult to establish definitively, particularly for content encountered outside its original deployment context. The framework requires ongoing maintenance as synthetic media capabilities evolve. Classification boundaries may require refinement as new application patterns emerge. Future work should develop technical indicators correlating with ASF levels, enabling automated pre-screening to support human classification.

7.3 Implementation Guidance

Organizations deploying generative AI should conduct ASF assessments as part of responsible AI governance processes, evaluating both intended applications and foreseeable misuse patterns. Assessment should occur at product design, prior to deployment, and periodically during operation as use patterns emerge. Classification documentation supports regulatory compliance demonstrations and provides audit trails for enforcement inquiries.

8. CONCLUSION

This paper introduced the Authenticity Spectrum Framework (ASF), a five-level classification system for synthetic media based on disclosure transparency, creator intent, and harm potential. Through systematic analysis of current synthetic media platforms and alignment with regulatory requirements, we demonstrated the framework's practical applicability to real-world governance challenges.

Drawing on prior research examining exploitation architectures in gaming [1] and mobile systems [2], the framework addresses a critical gap in AI governance by enabling proportionate responses to generative AI capabilities that neither stifle beneficial innovation nor permit harmful applications to proliferate. The framework's alignment with the EU AI Act, NIST AI RMF, and FTC guidance enables immediate practical application while supporting future regulatory development.

As generative AI continues to advance, governance frameworks must evolve correspondingly. The ASF provides a foundation for this evolution, offering researchers, regulators, platform operators, and practitioners a common language for discussing synthetic media risks and appropriate safeguards. Future research should explore technical implementations supporting ASF classification and integration with content moderation systems.

9. REFERENCES

[1] Martinson, F., & Rangel, D. (2023). A Comprehensive Analysis of Game Hacking through Injectors. International Journal of Computer Applications, 185(33), 56-63.

[2] Abukari, A. M., Amini, M., & Martinson, F. (2023). A Revealed Architecture of Camera-based Attacks for Smartphones. International Journal of Computer Applications, 185(27), 45-49.

[3] Goodfellow, I., et al. (2014). Generative adversarial nets. NeurIPS, 27.

[4] Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR, 10684-10695.

[5] Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. IEEE/CVF ICCV.

[6] Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge. California Law Review, 107, 1753-1820.

[7] Partnership on AI. (2023). Framework for Responsible Practices in Synthetic Media.

[8] Vaccari, C., & Chadwick, A. (2020). Deepfakes and

disinformation. Social Media + Society, 6(1).

[9] NIST. (2023). AI Risk Management Framework. NIST AI 100-1.

[10] European Union. (2024). AI Act. Official Journal of the EU.

[11] FTC. (2024). Guidance on AI in Advertising.

[12] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes. ACM Computing Surveys, 54(1), 1-41.

[13] Kietzmann, J., et al. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135-146.

[14] Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. Cyberpsychology, Behavior, and Social Networking, 24(3).

[15] Westerlund, M. (2019). The emergence of deepfake technology. Technology Innovation Management Review, 9(11).