Nova Science Press

Journal of Psychology & Education

Vol. 1, No. 5 (2026)

Authenticity Crisis, Personality Replication Risks, and Labeling Governance in Generative-AI Deep Synthesis: Toward a Traceable and Rights-Sensitive Framework

Zhiyong Song[1]*

(1School of Public Health, Ningxia Medical University，Ningxia)

*Corresponding author:Zhiyong Song; email:17585847963@163.com

Journal of Psychology & Education • Vol. 1, No. 5 (2026)

DOI: https://doi.org/10.66581/tbyn7997

Received 6 June 2026 • Accepted: 16 June 2026 • Published 30 June 2026

CITATION

Song, Z. (2026). Authenticity Crisis, Personality Replication Risks, and Labeling Governance in Generative-AI Deep Synthesis: Toward a Traceable and Rights-Sensitive Framework. Journal of Psychology & Education, 1(5), 32.

Abstract

Generative artificial intelligence has transformed deep synthesis from a specialized technique of media manipulation into an ordinary infrastructure for producing text, images, audio, video, virtual scenes, digital humans, and increasingly personalized AI agents. The ethical problem is therefore no longer limited to whether a specific image or video is fake. A deeper challenge concerns the technological blurring of three boundaries: the boundary between authentic and synthetic content, the boundary between a real person and a digital replica, and the boundary between human agency and automated proxy action. This article argues that the governance of AI-generated synthetic content should not be reduced to content labeling alone. Labeling is necessary because it helps users identify the artificial origin of media, but it is insufficient when synthetic content imitates a particular person's voice, likeness, expressive style, behavioral patterns, or decision-making tendencies. In such cases, the ethical issue is not merely deception but the unauthorized occupation of a person's communicative position. The article develops a three-level analysis of ethical risk: epistemic risks to public truth and social trust, personality-rights risks arising from digital replication, and responsibility risks generated by persona-based AI agents. Drawing on current governance instruments in China, the European Union, the United States, and international AI ethics frameworks, the article proposes a five-dimensional governance model: identifiability, provenance traceability, identity authorization, revocability, and accountability. The central claim is that deep synthesis governance must move from a narrow model of "AI-content disclosure" toward a broader

framework of "synthetic personhood governance," in which content origin, identity authorization, agentic behavior, and responsibility allocation are treated as interconnected problems.

Keywords: generative AI; deep synthesis; deepfakes; digital replicas; personality rights; AI agents; content labeling; authenticity; accountability; AI governance

1. Introduction

The rapid development of generative artificial intelligence has made synthetic content increasingly realistic, inexpensive, and accessible. Text can be generated in a human-like style; faces can be synthesized without a real subject; voices can be cloned from short recordings; videos can be manipulated to represent events that never occurred; and conversational agents can be personalized to imitate the linguistic habits, preferences, or decision-making patterns of specific individuals. The ethical concern raised by these technologies is not simply that some people may use artificial intelligence to lie. Deception, forgery, and impersonation existed long before generative AI. What is new is the scale, fidelity, automation, and social embeddedness of synthetic production. Deep synthesis technologies shift the problem from occasional falsification to a persistent uncertainty about whether a given piece of content, identity performance, or communicative act originates from a real human being.

This article focuses on three related problems: the authenticity crisis of AI-generated synthetic content, the personality replication risks of digital replicas, and the limits of labeling-based governance. The argument is that the most serious ethical challenge of deep synthesis is not exhausted by the question "Is this content AI-generated?" A more fundamental question is: "Whose presence, intention, and responsibility does this synthetic content claim to represent?" A synthetic video of an unknown person, a parody image of a fictional figure, a realistic clone of a public official's voice, and a personalized agent trained on a real person's private messages do not raise the same ethical issues. They differ in relation to identity, consent, public harm, and responsibility.

Recent regulatory developments show that governments have begun to recognize the need for specific rules on synthetic content. China's 2025 Measures for the Labeling of AI-Generated Synthetic Content require both explicit and implicit labeling of AI-generated synthetic content under defined conditions, while earlier Chinese rules on deep synthesis and generative AI services already imposed obligations concerning provider responsibility, user management, and content safety. The European Union's Artificial Intelligence Act introduces transparency obligations for certain AI systems, including requirements related to marking AI-generated content and disclosing deepfakes, with Article 50 transparency obligations applying from August 2026. The U.S. Copyright Office has also addressed the legal and policy problems of "digital replicas," emphasizing that AI-generated imitations of a person's voice or likeness create risks not fully captured by traditional copyright doctrine.

These developments confirm the relevance of labeling and disclosure. Yet labeling alone cannot resolve the ethical complexity of deep synthesis. A watermark or metadata label may show that a video, voice, or text was generated by AI, but it does not necessarily tell users whether the depicted person consented, whether the training data were lawfully obtained, whether the content is satirical or deceptive, whether a platform verified the identity claim, or who is responsible if an AI agent acts on behalf of a human. The governance of deep synthesis must therefore combine content transparency with identity authorization, provenance tracking, auditability, and responsibility allocation.

The article proceeds as follows. Section 2 clarifies the concepts of generative AI, deep synthesis, digital replicas, and persona-based agents. Section 3 analyzes the authenticity crisis as an epistemic and social-trust problem. Section 4 examines personality replication as a rights and dignity problem. Section 5 discusses AI agents and responsibility attribution. Section 6 reviews existing governance approaches, including Chinese regulation, the EU AI Act, U.S. digital-replica policy, NIST risk management, and international AI ethics frameworks. Section 7 argues that labeling governance is necessary but insufficient. Section 8 proposes a traceable and rights-sensitive governance framework. The conclusion argues that the future of deep synthesis governance lies in moving from content labeling to synthetic-personhood governance.

2. Conceptual Clarification: From Synthetic Media to Synthetic Personhood

A serious analysis of deep synthesis must distinguish several concepts that are often conflated. Generative AI refers to systems capable of producing new content in response to prompts, examples, goals, or environmental inputs. This includes language models, image-generation systems, audio synthesis models, video-generation tools, and multimodal systems. Deep synthesis refers more specifically to the use of AI to generate, modify, or combine text, image, audio, video, virtual scenes, or other digital content in a way that may simulate reality or human expression. Deepfakes are a narrower subset of synthetic media, usually involving realistic AI-generated or AI-manipulated image, audio, or video content that resembles real people, events, or environments and may mislead viewers.

Digital replicas represent a further development. A digital replica is not merely synthetic content; it is synthetic content that imitates an identifiable person. It may replicate a person's face, voice, gestures, style of speech, artistic manner, emotional tone, or other recognizable attributes. The U.S. Copyright Office's 2024 report on digital replicas uses the concept to address AI-generated representations that realistically portray a person's voice or appearance, especially when such portrayals raise concerns about unauthorized imitation and exploitation. In ethical terms, the decisive feature of a digital replica is not technical similarity alone but recognizability and representational substitution. The replica is ethically significant because it can lead audiences to believe that a real person has spoken, acted, endorsed, consented, or appeared.

Persona-based AI agents extend the problem from representation to action. An ordinary generative model produces outputs. An agentic system can pursue goals, call tools, retrieve information, interact with external systems, and perform multi-step tasks. Recent work on agentic AI emphasizes that such systems are no longer merely reactive text generators but can execute workflows and produce consequences in social, economic, and organizational contexts (Kasirzadeh & Gabriel, 2025; Mukherjee & Chang, 2025). When such systems are personalized with a user's data, memory, voice, writing style, or preference profile, they may operate as digital proxies. The ethical question then shifts from "What content was generated?" to "Who is being represented by this agent, and who is responsible for its acts?"

This conceptual shift is crucial. A deepfake video may deceive viewers about an event. A digital replica may misappropriate a person's identity. A persona-based agent may occupy a person's communicative role. These are related but not identical harms. The first concerns truth, the second concerns personality, and the third concerns agency and responsibility.

The term "synthetic personhood" is used here not to imply that AI systems possess moral personhood, consciousness, or legal personality. Rather, it refers to the social appearance of personhood created when synthetic systems simulate the presence, voice, style, intention, or agency of identifiable persons. The danger is not that a machine becomes a person in the metaphysical sense. The danger is that a

machine may be treated by others as if it were a person, or as if it legitimately represented a person, without adequate authorization, disclosure, or accountability.

3. The Authenticity Crisis: Epistemic Risk and Social Trust

The first ethical risk of generative-AI deep synthesis is an authenticity crisis. This does not mean that every AI-generated output is false or harmful. Synthetic media can support education, accessibility, entertainment, translation, cultural preservation, and creative production. The problem is that generative AI weakens many of the ordinary cues through which people evaluate the origin and reliability of information. In the past, a photograph, audio recording, or video had a special evidentiary status in everyday reasoning. It did not guarantee truth, but it often functioned as an epistemic anchor. Deep synthesis reduces the reliability of that anchor.

Rini (2020) describes this as a threat to the "epistemic backstop" provided by recordings. Social testimony is partly regulated by the possibility that an event could be recorded and checked. If recordings themselves become easy to fabricate, then the trustworthiness of testimony may also be weakened. The problem is not only that fake recordings may be believed. It is also that real recordings may be dismissed as fake. This second effect is sometimes described as the "liar's dividend": the availability of deepfake technology gives wrongdoers a plausible way to deny genuine evidence by claiming that it was generated or manipulated (Chesney & Citron, 2019).

The authenticity crisis therefore has two directions. In one direction, synthetic content may be falsely accepted as real. In the other direction, real content may be falsely rejected as synthetic. Both undermine public trust. Political communication, journalism, judicial evidence, academic integrity, and interpersonal relationships all depend on some shared methods for distinguishing fact from fabrication. When deep synthesis becomes ordinary, the burden of verification shifts from expert institutions to ordinary users, many of whom lack the technical capacity or time to evaluate provenance.

Empirical research supports this concern, while also warning against exaggerated panic. Vaccari and Chadwick (2020) found that synthetic political video can affect deception, uncertainty, and trust in news, but the impact is not simply that all viewers are fooled. The more complex effect is uncertainty. Users may become less confident in their ability to determine what is real. Similarly, research on AI disclosure has emphasized that AI-generated media can become nearly indistinguishable from human-created content, making disclosure design a central human-computer interaction issue (El Ali et al., 2024). The ethical problem is therefore not merely individual gullibility. It is the transformation of the information environment.

This distinction matters because a strong article should avoid technological alarmism. It is imprecise to claim that generative AI simply "reduces the public's ability to distinguish truth from falsehood." A more defensible claim is that deep synthesis increases the cognitive and institutional burden of authenticity judgment. It

raises the cost of verification. It creates incentives for malicious actors to exploit uncertainty. It also pressures institutions to develop provenance systems, labeling norms, and trusted verification channels.

A further risk is that labeling may itself reshape trust in unintended ways. If users are trained to treat labeled content as artificial and unlabeled content as authentic, then the absence of a label may generate false confidence. Yet labels can be removed, metadata can be stripped, screenshots can detach content from its original provenance, and cross-platform circulation can weaken disclosure. Research on synthetic-content warning labels suggests that labels can influence beliefs about whether content is AI-generated, but their effects vary by design and do not necessarily change engagement behavior (Gamage et al., 2025). This means that labeling should be treated as one part of a broader trust infrastructure, not as a complete solution.

4. Personality Replication: Identity, Dignity, and the Unauthorized Occupation of Communicative Position

The second major ethical risk concerns personality replication. Deep synthesis becomes more ethically serious when it imitates an identifiable person. A synthetic image of a non-existent face raises issues of disclosure and possible deception. A synthetic video of a real person saying something they never said raises additional questions of consent, dignity, reputation, and autonomy. A cloned voice used to issue

instructions, endorse a product, solicit money, or express intimate feelings can create harms that are not reducible to misinformation.

Personality replication involves more than the misuse of private data. It concerns the relation between a person and their socially recognizable identity. Human beings appear in the world through bodies, voices, names, gestures, styles of speech, habits of expression, and patterns of interaction. Deep synthesis can detach these attributes from the person and recombine them into synthetic performances. The harm lies not only in economic exploitation but also in the loss of control over how one is represented to others.

This is why traditional intellectual-property concepts are insufficient. Copyright protects works, not persons as such. Privacy law protects certain personal information, but a voice, face, or style may be publicly observable while still being deeply connected to identity. Defamation law addresses false statements that damage reputation, but not every unauthorized digital replica is defamatory. Right-of-publicity law protects commercial value in a person's likeness, but it often developed around celebrities and commercial endorsement. Digital replicas challenge all these categories because they can harm dignity, autonomy, reputation, trust, and emotional security even when no copyrighted work is copied and no explicit false factual claim is made.

The U.S. Copyright Office's digital-replica report is important because it recognizes that existing legal regimes may not adequately address unauthorized

AI-generated voice and likeness imitation. The proposed NO FAKES Act in the United States similarly reflects a policy movement toward giving individuals clearer rights over unauthorized digital replicas of their voice and visual likeness, though the details of such legislation must be balanced against free expression, satire, news reporting, and artistic use.

From an ethical perspective, the key concept is identity self-determination. Individuals have a legitimate interest in controlling whether and how their recognizable personal attributes are used to create synthetic representations. This interest is not absolute. Public interest, parody, historical representation, artistic expression, and political commentary may justify some uses. But the burden should increase when synthetic content is realistic, personalized, commercially exploited, intimate, defamatory, fraudulent, or likely to mislead audiences into believing that the person authorized or participated in the communication.

A crucial point is that an AI system trained on someone's data does not literally reproduce that person's mind. It is misleading to say that a persona agent has "the same thoughts" as the real person. Even if the agent imitates a person's writing style or decision tendencies, it remains a statistical and computational system. The ethical harm does not depend on proving psychological identity between the model and the human. The harm arises when others reasonably interpret the synthetic output as representing the person's intention, endorsement, emotion, or presence. The wrong is the unauthorized occupation of a communicative position.

This distinction helps avoid metaphysical confusion. The issue is not whether the digital replica has consciousness. The issue is whether the replica appropriates a human identity in social interaction. If a synthetic voice of a professor gives academic advice to students, if a replica of a deceased relative sends emotional messages to family members, or if a digital version of a public figure endorses a policy, the ethical question is whether the relevant audiences understand the artificial nature of the interaction and whether the represented person or legitimate rights-holder authorized that use.

Personality replication also raises posthumous and relational questions. A dead person cannot consent in the ordinary sense, but their digital resurrection may affect surviving relatives, cultural memory, and public dignity. A child's digital replica raises additional concerns because children cannot meaningfully consent to long-term uses of their likeness or behavioral data. A public figure's replica may involve stronger free-speech interests but also greater risk of public deception. A private individual's replica may involve stronger privacy and dignity interests. Governance must therefore be context-sensitive rather than based on a simple rule that all synthetic imitation is either prohibited or allowed.

5. Persona-Based AI Agents and Responsibility Attribution

The third risk emerges when generative AI moves from producing synthetic content to performing proxy action. An AI agent may answer emails, negotiate appointments, recommend purchases, manage social media, interact with clients,

provide customer support, conduct educational tutoring, or make preliminary decisions. When personalized, it may do these things in the name, style, or apparent identity of a specific human being.

This development creates a responsibility problem. If a person creates an AI agent based on their own data and authorizes it to communicate with others, who is responsible for what it says or does? If a third party creates an AI agent based on another person's data without authorization, what rights are violated? If an organization deploys a persona-like agent to represent a professional, brand, or public figure, how should responsibility be allocated among the developer, deployer, platform, user, and recipient?

A useful starting point is to reject the idea that responsibility can be assigned to the AI system itself. Current AI agents are not moral agents in the full sense. They do not possess autonomous moral understanding, legal accountability, or the capacity to bear blame and punishment. The relevant problem is not machine responsibility but distributed human and organizational responsibility. Matthias (2004) described the "responsibility gap" created by learning machines whose behavior cannot be fully predicted or controlled by designers. With persona-based agents, the gap becomes more socially visible because the agent may appear to speak or act as a person.

However, the responsibility gap should not become an excuse for responsibility evasion. A governance framework should allocate responsibility according to control, benefit, foreseeability, and role. The user who knowingly deploys an agent to speak

on their behalf should bear responsibility for authorized uses within the declared scope. The platform should bear responsibility for system design, safeguards, identity verification, abuse prevention, logging, and complaint handling. The developer should bear responsibility for foreseeable technical defects, inadequate safety design, or misleading product claims. A malicious third party who creates an unauthorized replica should bear direct responsibility for impersonation, fraud, or personality-rights violations. Organizations that deploy agents in professional or commercial settings should not be allowed to shift responsibility to the system.

The responsibility question becomes more difficult when agent behavior exceeds the user's expectation. Suppose a user authorizes an agent to reply to routine messages, but the agent makes a harmful promise, discloses sensitive information, or expresses a defamatory statement. In such cases, the governance system must examine the authorization boundary, the design of user controls, the clarity of warnings, the availability of human review, the foreseeability of the harm, and the adequacy of logging. Responsibility should be graded rather than all-or-nothing.

The development of AI agents also strengthens the case for interaction disclosure. A person interacting with an agent should generally know that they are interacting with AI, especially when the agent represents a real person, professional role, or institution. The EU AI Act's transparency obligations include requirements that people be informed when interacting with AI systems, unless this is obvious from the circumstances. This principle becomes especially important for persona-based agents

because the harm is not only informational deception but relational misrepresentation. People have a legitimate interest in knowing whether they are speaking with a human being, a generic chatbot, or a personalized proxy.

6. Existing Governance Approaches: Achievements and Limits

Current governance approaches can be grouped into five categories: ethical principles, risk management frameworks, content-labeling rules, digital-replica protections, and agentic accountability mechanisms.

International AI ethics principles provide broad normative foundations. The OECD AI Principles emphasize human-centered values, fairness, transparency, robustness, security, and accountability. UNESCO's Recommendation on the Ethics of Artificial Intelligence similarly stresses human rights, dignity, transparency, fairness, privacy, and human oversight. These frameworks are useful because they articulate general values, but they are too abstract to resolve concrete questions about digital replicas and persona-based agents. They tell us that AI should be transparent and accountable, but they do not specify how a platform should verify consent for voice cloning, how long agent logs should be stored, or when a synthetic representation of a real person should be prohibited.

Risk management frameworks offer more operational guidance. NIST's AI Risk Management Framework and the 2024 Generative AI Profile encourage organizations to govern, map, measure, and manage AI risks across the system lifecycle. The

Generative AI Profile specifically helps organizations identify risks distinctive to generative AI and align mitigation actions with organizational goals. This lifecycle approach is valuable because deep synthesis risks arise not only at the output stage but also during data collection, model training, deployment, user interaction, content distribution, and post-harm remediation.

China's approach is notable for its direct regulation of deep synthesis and labeling. The Provisions on the Administration of Deep Synthesis Internet Information Services, the Interim Measures for the Management of Generative Artificial Intelligence Services, and the 2025 labeling measures establish a regulatory path that emphasizes provider obligations, service management, safety, and content identification. The 2025 labeling rules are especially relevant because they distinguish explicit labels visible to users from implicit labels embedded in metadata or technical formats. This dual model recognizes that human-readable disclosure and machine-readable traceability serve different functions.

The European Union's AI Act adopts a broader risk-based regulatory structure. Its transparency rules require certain AI systems to inform users when they are interacting with AI and require synthetic content to be marked in machine-readable form under specified conditions. The European Commission's Code of Practice on Transparency of AI-Generated Content is intended to support compliance with AI Act obligations concerning marking and labeling of AI-generated content, including deepfakes and certain AI-generated publications. The EU model is significant because

it connects labeling to information-ecosystem integrity, deception prevention, and risk governance.

The United States currently has a more fragmented approach. Instead of a single comprehensive AI law, governance is distributed across copyright policy, right-of-publicity law, privacy law, consumer protection, state legislation, and proposed federal bills. The U.S. Copyright Office's digital-replica report and the NO FAKES Act proposal show growing recognition that voice and likeness replication require specific legal protection. The advantage of this approach is sensitivity to expression and constitutional concerns. Its disadvantage is uneven protection and uncertain enforcement.

Despite these developments, three governance gaps remain. First, content-labeling rules often focus on whether media are AI-generated, not whether a real person's identity was authorized. Second, digital-replica rules often focus on voice and likeness, but may not fully address behavioral style, writing style, personal memory, and agentic imitation. Third, AI-agent governance remains underdeveloped, especially for systems that act across platforms, access personal data, or represent specific persons in social interaction.

7. Why Labeling Is Necessary but Insufficient

Labeling is ethically necessary because users have a legitimate interest in knowing whether content was generated or substantially modified by AI. Without

disclosure, audiences may misinterpret synthetic media as direct evidence of real events or real human expression. Labeling also supports accountability by creating technical and institutional mechanisms for provenance tracking.

However, labeling has at least six limitations.

First, labels are technically fragile. Visible watermarks can be cropped, blurred, overwritten, or removed. Metadata can be stripped when files are uploaded, compressed, screenshotted, or shared across platforms. Machine-readable labels require interoperability standards and platform cooperation. If different systems use incompatible marking methods, the social value of labeling declines.

Second, labels are context-dependent. A small label may be sufficient for low-risk entertainment content, but inadequate for political communication, financial advice, medical information, legal evidence, or synthetic content depicting a real person. The more serious the potential harm, the more robust the disclosure should be.

Third, labels do not prove authorization. A video labeled “AI-generated” may still be an unauthorized imitation of a real person. Conversely, authorized synthetic content may still require disclosure because users deserve to know that they are interacting with a synthetic representation rather than a live person. Therefore, the label “AI-generated” answers only one question: the mode of production. It does not answer the questions of consent, identity, responsibility, or truth.

Fourth, labels may generate false assumptions. Users may infer that unlabeled content is authentic, even though labels may be absent because of noncompliance, technical loss, foreign origin, malicious removal, or platform failure. In this sense, labeling systems can produce a dangerous binary: labeled equals artificial, unlabeled equals real. Governance must therefore combine labeling with media literacy and trusted provenance channels.

Fifth, labels do not resolve responsibility. If an AI agent makes a harmful statement, a label saying "AI-generated" does not explain whether the user authorized the statement, whether the platform failed to implement safeguards, whether the developer misrepresented system capabilities, or whether a third party abused the system.

Sixth, labels may not protect personality rights. A person harmed by an unauthorized replica may not be satisfied by disclosure alone. A synthetic pornographic image, fraudulent voice clone, fake endorsement, or emotional manipulation using a deceased relative's likeness can remain harmful even if labeled. Some uses require prohibition, removal, compensation, or other remedies, not merely disclosure.

For these reasons, labeling should be understood as a threshold governance mechanism. It creates minimum transparency, but it cannot carry the full ethical burden of deep synthesis governance.

8. Toward a Traceable and Rights-Sensitive Governance Framework

This article proposes a five-dimensional framework for governing generative-AI deep synthesis: identifiability, provenance traceability, identity authorization, revocability, and accountability.

8.1 Identifiability

Identifiability means that users should be able to recognize when they are viewing, hearing, reading, or interacting with AI-generated or AI-mediated content. This includes visible labels, audio disclosures, interface notices, and contextual warnings. The level of disclosure should be proportionate to risk. Low-risk creative content may need ordinary labeling. High-risk content involving political speech, public safety, professional advice, or identifiable persons may require prominent and persistent disclosure.

Identifiability also includes interaction disclosure. When a user communicates with an AI system, especially one representing a person or organization, the system should disclose its artificial nature. This is not merely a consumer-protection measure. It protects relational autonomy. People make different judgments when they believe they are interacting with a human being rather than a machine.

8.2 Provenance Traceability

Traceability means that synthetic content should be linked, where feasible, to reliable provenance information: generation time, generating system, account or organization responsible for generation, modification history, and distribution path. This does not require public disclosure of all technical details, which may raise privacy and security concerns. It does require auditable records accessible to appropriate institutions under lawful procedures.

Traceability is especially important because synthetic media often circulate beyond the platform on which they were created. A visible label may disappear; a provenance record may remain available if supported by interoperable standards. Governance should therefore encourage machine-readable provenance infrastructure, platform-to-platform metadata preservation, and audit mechanisms for high-risk synthetic content.

8.3 Identity Authorization

Identity authorization is the missing element in many labeling regimes. When synthetic content imitates an identifiable person, governance should require evidence that the relevant identity attributes were lawfully used. This includes consent for voice cloning, likeness generation, digital human creation, and persona-based agent deployment.

Authorization should be specific, informed, revocable, and purpose-limited. A person who consents to voice cloning for an audiobook has not necessarily consented

to political endorsement, commercial advertising, intimate conversation, or posthumous use. A person who permits a platform to process their data has not necessarily permitted a third party to create a personalized agent that imitates them. Consent should not be hidden in broad terms of service.

Identity authorization should also distinguish between private persons, public figures, professionals, children, and deceased persons. Public-interest exceptions may exist, but the burden of justification should increase when the synthetic representation is realistic, misleading, commercially exploitative, intimate, or likely to affect reputation and trust.

8.4 Revocability and Redress

Revocability means that individuals should have practical mechanisms to withdraw authorization, request removal, challenge unauthorized replicas, and seek correction or compensation. Without revocability, consent becomes a one-way transfer of identity control. This is especially dangerous because synthetic replicas can persist, be copied, modified, and redistributed.

Redress mechanisms should be fast and accessible. Victims of unauthorized replicas often face immediate reputational, emotional, financial, or relational harm. A slow litigation process is insufficient. Platforms should provide complaint channels, identity verification procedures, emergency removal for high-risk harms, appeal

mechanisms to protect legitimate expression, and cooperation with regulators or courts when necessary.

8.5 Accountability

Accountability requires clear responsibility allocation. The governance framework should reject both technological determinism and user-only responsibility. Developers, deployers, platforms, professional users, and end users each have duties corresponding to their control and benefit.

Developers should design systems with safeguards against impersonation, non-consensual sexual content, fraud, and high-risk misuse. Platforms should verify identity authorization for high-risk replication tools, preserve logs, respond to complaints, and prevent repeat abuse. Deploying organizations should conduct risk assessments before using agents in education, health, finance, employment, law, or public administration. Users should be responsible for intentional misuse and for authorized agent actions within reasonably foreseeable scope.

For persona-based agents, accountability requires action logs. If an agent sends a message, makes a recommendation, approves a transaction, or interacts with third parties, the system should record the human instruction, autonomy level, tools used, data accessed, output generated, and review status. Without logs, responsibility cannot be reconstructed.

9. Practical Governance Design: Classification, Standards, and Institutional Coordination

A workable governance system should classify deep synthesis applications by risk rather than treating all AI-generated content alike. Four levels are useful.

Low-risk synthetic content includes clearly fictional, artistic, educational, or entertainment uses that do not imitate identifiable persons or affect significant interests. These uses should generally require light labeling but not heavy licensing.

Medium-risk synthetic content includes realistic media that may be mistaken for real events but does not imitate a specific person in a harmful context. Such content requires visible labeling and provenance preservation.

High-risk synthetic content includes digital replicas of identifiable persons, synthetic political communication, professional advice, commercial endorsement, educational assessment, medical or psychological guidance, financial communication, and AI agents acting on behalf of users. These uses require stronger disclosure, identity authorization, logging, and auditability.

Prohibited or presumptively unlawful uses include non-consensual sexual deepfakes, fraud, extortion, impersonation for financial or legal gain, synthetic evidence fabrication, unauthorized replicas designed to deceive, and agentic systems that make significant decisions without appropriate human oversight in high-stakes settings.

Technical standards are also necessary. Labeling should include both human-readable and machine-readable forms. Provenance systems should be interoperable. Detection tools should be used cautiously because deepfake detection is an adversarial field; detection models can be bypassed, and false positives can harm legitimate speakers (Mirsky & Lee, 2021; Neekhara et al., 2020). The governance aim should not be perfect detection but layered trust: disclosure, provenance, authentication, audit, and institutional verification.

Institutional coordination is equally important. Governments can set baseline rules. Platforms can implement labeling and takedown systems. AI developers can build safeguards and provenance mechanisms. Civil society can monitor abuse. Academic researchers can evaluate label effectiveness and social impact. Courts and regulators can clarify responsibility. No single actor can govern deep synthesis alone.

Education should not be ignored. Users need AI literacy, but literacy cannot replace institutional responsibility. It is unfair to place the entire burden of verification on ordinary citizens. A mature governance model should combine public education with enforceable duties for those who design, deploy, profit from, and distribute deep synthesis systems.

10. Conclusion

Generative-AI deep synthesis has transformed the ethics of digital media. The central issue is no longer merely whether a photo, video, voice, or text is fake. The

deeper issue is whether synthetic systems are beginning to simulate, appropriate, and act through the recognizable identities of human beings. This creates three connected crises: an authenticity crisis in which the boundary between real and synthetic content becomes unstable; a personality replication risk in which a person's voice, likeness, style, or behavioral patterns can be used without authorization; and a responsibility crisis in which persona-based agents act in ways that blur the line between human intention and automated execution.

Labeling is an essential response, but it is not sufficient. It can identify AI-generated content, but it cannot by itself establish consent, prevent identity misuse, preserve social trust, or allocate responsibility. The governance of deep synthesis must therefore move beyond simple disclosure toward a framework that integrates identifiability, provenance traceability, identity authorization, revocability, and accountability.

The normative goal is not to suppress generative AI. Synthetic media and AI agents can produce genuine social value. The goal is to prevent technological simulation from eroding the conditions of truthful communication, personal dignity, and responsible agency. In the age of generative AI, society must not only ask whether content is artificial. It must also ask whose identity is being represented, whether that representation is authorized, what social relationship it creates, and who must answer for its consequences.

References

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

China Cyberspace Administration, Ministry of Industry and Information Technology, Ministry of Public Security, & National Radio and Television Administration. (2022). *Provisions on the Administration of Deep Synthesis Internet Information Services*.

Cyberspace Administration of China. (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services*.

Cyberspace Administration of China. (2025). *Measures for the Labeling of AI-Generated Synthetic Content*.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

El Ali, A., Venkatraj, K. P., Morosoli, S., Naudts, L., Helberger, N., & Cesar, P. (2024). Transparent AI disclosure obligations: Who, what, when, where, why, how. *arXiv*.

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence*.

European Commission. (2026). *Code of Practice on Transparency of AI-Generated Content*.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*, 689–707.

Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. (2025). Labeling synthetic content: User perceptions of warning label designs for AI-generated content on social media. *arXiv*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*, 2672–2680.

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123.

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 149–152.

Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI agents for alignment and governance. *arXiv*.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), Article 7.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Mukherjee, A., & Chang, H. H. (2025). Agentic AI: Autonomy, accountability, and the algorithmic society. *arXiv*.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework 1.0*.

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*.

Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2020). Adversarial threats to deepfake detection: A practical perspective. *arXiv*.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding, 223*, 103525.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.

Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes: The manipulation of audio and visual evidence*. Data & Society.

Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint, 20*(24), 1–16.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, Article 15.

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*.

U.S. Copyright Office. (2024). *Copyright and Artificial Intelligence, Part 1: Digital Replicas*.

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), 1–13.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52.