Universal Library of Engineering Technology Research Article

ISSN: 3064-996X | Volume 2, Issue 2

Open Access | PP: 21-26

DOI: https://doi.org/10.70315/uloap.ulete.2025.0202004

Universal Library Open Access Publications LLC

AI and Deepface Influencers: The Challenge of Authenticity in the Online Space

Samoilenko Vladyslava

Brand Ambassador, The Noli Shop, New York, United States.

Abstract

The article investigates the phenomenon of AI influencers generated with DeepFake technologies and examines their impact on the authenticity of digital content, audience trust, and the economic models of marketing. It outlines the technological foundations—generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, transformers, text-to-speech systems, and voice conversion—and proposes a classification of AI influencers by modality (visual, audio, textual, and multimodal). The methodological framework rests on a comparative analysis of previous studies in the field, which identifies authenticity challenges in the online environment arising from advances in artificial intelligence and the proliferation of deepfakes. To counter these threats, the study recommends multimodal detectors of synthetic content and legal measures such as mandatory labeling, international certification standards, and legislative adaptation. The conclusion offers guidance on developing real-time detection methods, harmonizing legal regulation, and expanding media-literacy programs in the "post-truth" era. The findings will interest an interdisciplinary community of scholars and practitioners—from digital-ethics theorists and media sociologists to specialists in artificial intelligence and marketing—who seek a deeper understanding of the transformation of identity and trust in the age of software-generated personas. Legislators and platform developers shaping regulatory and technological mechanisms for verifying online content authenticity will likewise find the results valuable.

Keywords: *AI-Influencers, DeepFake, Online Trust, Synthetic Content Detection, GAN, Legal Regulation, Media Literacy.*

INTRODUCTION

The rise of social media and artificial intelligence (AI) has produced a new class of influencers—fully synthetic personas and "clones" of real individuals generated with DeepFake technologies [2]. These AI-influencers already engage millions of users in brand campaigns, virtual shows, and educational initiatives. At the same time, their widespread adoption poses a serious challenge to the authenticity of online content: public distrust is growing, the "plausibility paradox" is intensifying, and the "liar's dividend" emerges, whereby any genuine publication may be perceived as fake [1, 3].

Contemporary scholarship on deepfake influencers and online authenticity can be divided into four thematic clusters. The first concentrates on the marketing and behavioural dimensions of synthetic content. Alsharairi N. A. and Li L. [1] examine techniques for attracting and motivating youth toward healthier lifestyles through digital campaigns; however, their analysis remains general and does not address audience interaction with artificial characters. Campbell C. et al. [8] analyse the potential impact of deepfake advertising on consumer behaviour, highlighting opportunities for personalisation and enhanced engagement, yet their work is limited to conjecture and lacks empirical validation.

The second cluster addresses technical approaches to the generation, protection, and detection of deepfake content. Perov I. et al. [6] present DeepFaceLab, a modular and extensible face-swap framework, whereas Boutadjine A. et al. [7] conduct a comprehensive study of multimedia deepfake algorithms, comparing GAN architectures and attribution methods. Mubarak R. et al. [4] provide an extensive survey of synthetic-content detection across visual, audio, and textual formats, classifying detectors by their use of spatiotemporal features and deep-learning techniques. Kirchenbauer J. et al. [3] focus on watermarking in large language models, proposing cryptographic methods for embedding trace information in generated text—an approach potentially applicable to audio and video as well.

The third cluster explores legal and regulatory aspects. Chawki M. [2] analyses U.S. case law and legislative gaps

Citation: Samoilenko Vladyslava, "AI and Deepface Influencers: The Challenge of Authenticity in the Online Space", Universal Library of Engineering Technology, 2025; 2(2): 21-26. DOI: https://doi.org/10.70315/uloap.ulete.2025.0202004.

concerning deepfake media, revealing the need to distinguish clearly among copyright, image rights, and platform liability, while proposing enhanced enforcement through specialised provisions. Yadlin-Segal A. and Oppenheim Y. [10] expand this discussion to global social-media regulation, addressing dilemmas of censorship, freedom of expression, and the commercialisation of content control.

The fourth cluster examines political and ethical dimensions of authenticity. Vaccari C. and Chadwick A. [5] investigate how synthetic political videos affect news credibility, demonstrating increased uncertainty and reduced trust with even minor distortions. Gregory S. [9] advances the debate on “front-line” and “remote” witnessing in journalism, emphasising that deepfake technology undermines authenticity infrastructures and necessitates new protocols for verification and fact-checking.

Collectively, these studies highlight the multifaceted nature of the problem, from engineering solutions for detecting and labelling synthetic content to analyses of audience impact, legal mechanisms, and socio-political consequences. Nevertheless, notable contradictions persist: technical research often prioritises detection and protection without considering the socio-psychological effects of interacting with deepfake influencers, whereas marketing studies view them primarily as tools of enhanced persuasion while overlooking risks and legal complexities. Despite extensive surveys and theoretical frameworks, empirical investigations of user perceptions and the long-term effects of synthetic personas remain insufficient, as does interdisciplinary integration of technical, social, and legal approaches. These gaps create opportunities for future work on comprehensive models that assess the risks and benefits of deploying deepfake influencers and on ethically and legally grounded practices for their use.

The **objective** of the present study is to identify and systematise the key technological, social, and legal aspects of employing synthetic influencers based on DeepFake technologies and to assess their impact on audience trust in the online environment.

The **study’s novelty** lies in the development of an integrated theoretical and methodological framework for systematising research on AI and DeepFake influencers. This framework includes a unified classification of their media formats, identification of structural challenges to online authenticity, and justification of an integrated suite of multimodal synthetic-content detectors alongside conceptual legal mechanisms for counteraction.

The **author’s hypothesis** posits that, in the absence of clear labelling mechanisms and reliable real-time detection systems, synthetic influencers will accelerate the erosion of audience trust in all forms of media content, potentially leading to an “information collapse” in digital society.

The **methodological** foundation of the study is a comparative analysis of existing research, through which the challenges to online authenticity posed by AI advancement and the emergence of deepfakes have been identified.

Technological Basis and Classification of Deepfake Influencers

Methods for generating and synthesizing multimedia content using deep neural networks underpin every synthetic influencer. The key architectures and models employed to create visual, audio, and textual DeepFakes are outlined below, followed by their systematization by format and purpose.

Generative models (GANs, Generative Adversarial Networks) comprise two neural networks—a generator and a discriminator—trained in an adversarial manner [1, 3]. The generator attempts to produce synthetic examples, while the discriminator distinguishes them from real data. This iterative process gradually enables the generator to output increasingly realistic images and videos. GAN architectures form the basis of most modern face-swapping and face-generation systems [4].

Variational Autoencoders (VAEs) adopt a different approach by encoding input data into a parametric latent space and then reconstructing it with added stochastic noise. Owing to their probabilistic nature, VAEs can generate new data by sampling various points in latent space. Although their results are less sharp than those of GANs, VAEs are widely used for preliminary analysis and dimensionality reduction of samples [2].

Diffusion models iteratively add and then remove noise while learning the reverse process. Their training stability and synthesis quality have made them integral to contemporary text-to-image systems (e.g., DALL·E 2, Midjourney) and increasingly relevant for high-precision face generation [1].

Transformers have revolutionized natural-language processing; the self-attention mechanism allows a model to capture global dependencies in sequences of any type. Modern large language models employ this architecture for text generation, dialogue management, and real-time control of virtual characters’ skeletal animation [7, 8].

Voice transformation (TTS and voice conversion):

- WaveNet and Tacotron establish the baseline for neural speech synthesis, generating the waveform and mel-spectrogram, respectively.

- Voice conversion—implemented, for example, with autoencoder-based or cyclic GAN models—retunes the prosody and timbre of one speaker’s utterance to match another’s voice while preserving linguistic content [1, 2].

Table 1 below presents a systematization of the main types of synthetic influencers by content format, generation technology, and application examples.

Table 1. Systematization of the main types of synthetic influencers by content format, generation technologies, and application examples (compiled by the author based on the analysis: [1, 2, 7, 8]).

Category	**Briefdescription**	**Key technologies**	**Examples**
Visual	Characters that exist solely as generated faces and movements and may lack an audio or textual component.	GAN (StyleGAN 2), VAE, diffusion, face-swap (DeepFaceLab)	LilMiquela, Shudu
Audio influencers	Synthetic voices that read scripts and appear in podcasts or audio advertising, without a visual "shell."	WaveNet, Tacotron, voice conversion(AutoVC,StarGAN-VC)	"TomCruise" voicegeneration
Textual (chatbots)	Virtual "personalities" that exist only in written communication and can conduct dialogue or generate articles and posts.	GPT-3/4, LSTM, transformers	ChatGPT persona, Replika
Multimodal	Full virtual influencers that integrate visual, audio, and text and interact with users simultaneously through video, audio, and text channels.	Combinations of GAN + TTS + LLM and real-time engines (Unity/Unreal)	FN Meka, Blawko

Three principal types of synthetic influencers differ in the method of image generation, the extent of real-data involvement, and the manner of audience interaction. Fully virtual avatars are produced from scratch by generative neural networks such as StyleGAN 2/3 and are equipped with synthesized voices and reactions created by large language models. Because they are untethered to any real individual or likeness, they avoid obligations concerning rights to original material; nonetheless, these characters frequently suffer from the uncanny-valley effect, as subtle inconsistencies in facial expressions or intonation can elicit a sense of artificiality in viewers.

Deepfake clones of real individuals obtain their bodies and voices through face-swap and voice-conversion algorithms that superimpose a recognizable celebrity onto an animated or three-dimensional model. The resulting recognizability and perceived trustworthiness effectively capture attention and foster loyalty, yet such solutions entail significant legal risks related to the right of publicity and potential copyright infringement.

Hybrid characters combine real and fictional elements, merging distinctive traits of a specific individual with features of an artificially created avatar while retaining a unique brand identity. Implementation commonly employs face-generation technologies alongside precise three-dimensional skeletons and GPT-controlled dialogue. Virtual consultants in e-commerce that stream and answer user inquiries in real time illustrate this approach [5, 6].

The range of architectures—from GANs and VAEs to transformers and diffusion models—provides the flexibility to create diverse forms of deepfake influencers. Classifying them by media format and degree of "reality" supports targeted application in marketing, entertainment, and education, while simultaneously underscoring the need for detection mechanisms and legal regulation.

Impact on Authenticity, Trust and the Market

Deepfake influencer technologies exert a multilayered impact on the perception of online content, altering users' conceptions of authenticity, eroding trust, and creating new economic risks and opportunities. The emergence of increasingly realistic deepfake videos and images has introduced the phenomenon known as the "liar's dividend," whereby audiences begin to doubt the genuineness of any media, including authentic material [4]. Controlled experiments show that participants previously exposed to deepfake content are, on average, 30 % less likely to trust recordings featuring politicians and journalists [5]. This widespread paranoia complicates the work of news organizations and undermines information hygiene. Figure 1 below illustrates a face-swap example in which Jim Carrey's face is substituted onto video of Alison Brie using DeepFaceLab.

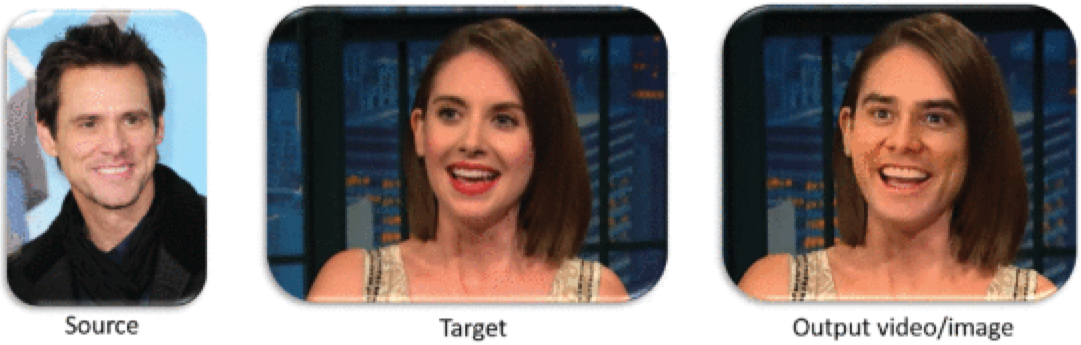

Figure1. An example of face replacement, in which Jim Carrey's face was replaced by Alison Brie's video using Deep Face Lab [1].

More specifically regarding face synthesis, this process has beneficial applications in digital art and video-game character creation, enabling artists to design unique and diverse characters without relying on real-life references. Well-known publicly available face-generation tools include "This Person Does Not Exist" (TPDNE), which employs generative adversarial networks (GANs) to create realistic human faces that do not correspond to real individuals. In addition, advanced models such as Midjourney and DALL-E 2 are readily accessible online, allowing faces to be generated instantaneously from simple text prompts[1]. For example, in DALL-E 2 the prompt "create the face of a 30-year-old Asian man with a moustache" produces a highly realistic visage within seconds, as illustrated in Figure 2.

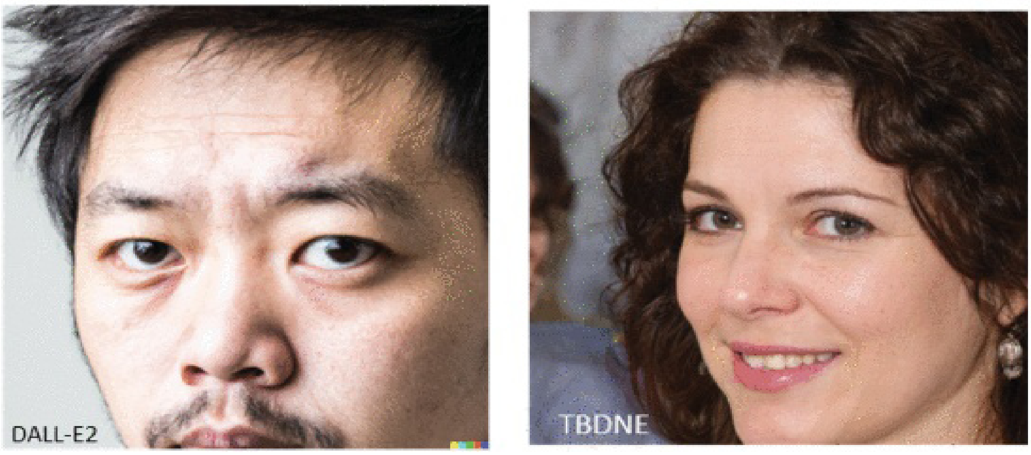

Figure 2. The result of creating a face using DALE 2 and TBDE [1].

Social bots that disseminate synthetic content can create the illusion of a viral public consensus, thereby amplifying the echo-chamber effect and fostering group radicalization [1].

Synthetic influencers offer brands new promotional channels: virtual personas scale effortlessly, require no logistics, and can operate continuously. The most popular virtual influencers generated more than US $10 million in sponsorship contracts within a single year. At the same time, deepfake tools are exploited for financial fraud: in a notable case, attackers mimicked a CEO's voice and persuaded staff to transfer €220 000 to a fraudulent supplier account [4].

Table 2 summarizes the principal consequences of employing deepfake influencers.

Table 2. Key consequences of using deepfake influencers(compiled by the author based on the analysis:[4, 5]).

Parameter	**Type ofimpact**	**Mechanism**	**Illustration**
Authenticity	Doubts about the "real" nature of any video	*Liar's dividend*: any recording can be claimed as fake	30 % of users mistrust political addresses after encountering deepfakes
Trust	Diminished confidence in media and institutional sources	Imitation of official channels, substitution of faces and voices	Circulation of deepfake videos containing fabricated political statements
Economy	Opportunities—large-scale partnerships; risks—financial fraud and reputational loss	Virtual influencers secure sponsorships; fraudulent CEO voice calls	US $10 million in revenue for leading virtual influencers; €220 000 transferred after a deepfake call

Thus, deepfake influencers radically reshape the landscape of digital trust: they stimulate innovative marketing strategies while simultaneously undermining fundamental assumptions about the authenticity of online content and generating risks for both the media and business sectors.

Detection Methods and Legal Regulation

The DeepFake-influencer problem calls for a combination of technical tools for synthetic-content detection and well-designed legal responses. The main approaches in both domains are outlined below.

Research distinguishes two broad classes of methods for visual DeepFake detection:

- Feature-engineered techniques: analysis of inconsistencies in shadows, optical flow, blinking patterns and facial anatomy (e.g., SURF, physiological signals, discrepancies in background illumination).

- Deep-feature techniques: convolutional neural networks (CNN), capsule networks (CapsuleNet), recurrent–convolutional hybrids (CNN + LSTM), attention mechanisms, 3-D convolutions and neural autoencoders (MesoNet, XceptionNet, DPNet).

Approaches to synthetic-speech detection likewise fall into handcrafted and deep-learning categories:

- Handcrafted techniques examine spectral and bifurcation properties of speech (log-Mel spectrograms, LCNN, bispectral analysis, the "breathing–speech–silence" phenomenon).

- Deep-learning techniques rely on convolutional and recurrent architectures (ResNet, RawNet 2, EfficientCNN, Attentive Filtering Networks) to extract neural “fingerprints” of authentic speech and locate activation anomalies (DeepSonar, RW-ResNet).

Methods for identifying AI-generated text include:

1. Simple classifiers (SVM, logistic regression) based on lexical–statistical features such as TF-IDF and frequency patterns.
2. Zero-shot techniques that employ the generative models themselves (GPT-2, Grover) to assess log-probabilities and next-token curvature (DetectGPT, GLTR) [2].
3. Fine-tuned large language models (RoBERTa, BERT) for the binary task of “human vs. machine text” [9].
4. Logit-level watermarks that embed an easily detectable signature in every generated word or token [4, 10].

Table 3 presents the legal instruments for DeepFake regulation.

Table 3. Comparative characteristics of legal instruments on Deepface regulation (compiled by the author based on the analysis: [1, 2, 4, 9, 10]).

Jurisdiction	**Key actsandnorms**	**Scope**	**Limitations**
United States	• Deepfake Report Act (2019) – annual DHS report on DeepFake threats • Section 230 Communications Decency Act (CDA) – platform immunity for UGC • State-level statutes on defamation, right of publicity and non-consensual pornography • Criminal provisions on fraud and cyber-crime (Title 18 U.S.C.)	Reporting, civil remedies, criminal enforcement	Section 230 hinders suits against platforms; state laws are fragmented; proving “intent” and causation for fakes is difficult
European Union	• Digital Services Act (Reg. EU 2022/2065) – platform liability for moderating illicit content • AI Act proposal (COM/2021/206) – AI-risk classification and transparency requirements	Platforms, high-risk AI applications	Proposal still pending; focus on institutional AI users rather than UGC DeepFakes
United Kingdom	• Online Safety Bill – duties of social networks to remove illegal content; specific amendments on DeepFake pornography	Platform obligations, user rights protection	Bill not yet in force; ongoing debate on free-speech boundaries
China	• Cybersecurity Law – oversight of online information • Draft Regulations on Deep Synthesis (2023) – mandatory labelling of synthetic content	Online-media regulation, compulsory identification of synthesis	Strict state censorship; reduced trust in any online information

To ensure responsible use of DeepFake technology, compulsory digital labelling of all synthetic video content and avatars is recommended, via watermarks or metadata, allowing clear differentiation between authentic and generated material. Maximum algorithmic transparency should be pursued through open registries of models that record architecture, release date and potential risk profiles, thereby strengthening user trust and facilitating independent audits. Reliability of DeepFake-detection systems can be enhanced by developing unified international certification standards under ISO/IEC, providing consistent testing and validation procedures. A key element in combating disinformation is improved media literacy within society and business communities through dedicated educational programmes, regular “red-flag” reviews for influencer campaigns and the dissemination of practical guidelines for critical evaluation of multimedia content. In cases of systemic failures in moderating DeepFake material, limiting the scope of Section 230 CDA for major intermediary platforms is proposed, creating an economic incentive to refine moderation algorithms and increasing legal accountability for distributing harmful content [2].

Only a comprehensive implementation of technical and legal measures lays the foundation for a balanced approach, preserving the innovative potential of AI-based influencers while minimising risks to trust and security within the digital ecosystem.

CONCLUSION

The conducted study indicates that deepfake influencers, leveraging contemporary generative models—including GANs, VAEs, diffusion models, and transformers—are emerging as a prominent component of the digital ecosystem, evoking mixed reactions among audiences that range from admiration for their realism to profound unease. The keyconsequencesare:

1. Erosion of trust. The “liar’s dividend” phenomenon diminishes users’ receptiveness to any media content and complicates the work of both journalists and marketers.
2. Socio-political risks. Synthetic content is employed for disinformation, increased polarization, and interference in public discourse.

3. Economic opportunities and threats. Virtual influencers generate multimillion-dollar revenues through sponsorship projects, yet simultaneously endanger the financial security of companies and individuals through deepfake-enabled fraud.

Future efforts should focus on:

• Adaptive multimodal frameworks for real-time detection of deepfake influencers;

• Harmonization of legal norms at the international level and a reassessment of platform immunity for user-generated content;

• Studies of the perception of synthetic personas across diverse cultural and age groups.

A balanced approach that combines advanced technical tools with mature legal mechanisms will preserve the potential of AI-driven influencers as an innovative marketing and educational phenomenon while minimizing risks and maintaining users' trust in the digital environment.

REFERENCES

1. Alsharairi N. A., Li L. Social marketing targeting healthy eating and physical activity in young adult university students: A scoping review //Heliyon. - 2024.- Vol. 10 (11).-pp. 144497 – 144529. DOI: 10.1109/ACCESS.2023.3344653.
2. Chawki M. Navigating legal challenges of deepfakes in the American context: a call to action //Cogent Engineering. - 2024. - Vol. 11 (1).- pp. 1-13. DOI: 10.1080/23311916.2024.2320971.
3. Kirchenbauer J. et al. A watermark for large language models //International Conference on Machine Learning. PMLR. - 2023. - pp. 17061-17084.
4. Mubarak R. et al. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats //Ieee Access.- 2023. - Vol. 11. - pp. 144497-144529.
5. Vaccari C., Chadwick A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news // Social media+ society. - 2020.- Vol. 6 (1). - pp. 1-8. DOI:10.1177/20563051209034.
6. Perov I. et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework //arXiv preprint arXiv:2005.05535. - 2020. - pp. 1-8. DOI: 10.48550/arXiv.2005.05535.
7. Boutadjine A. et al. A comprehensive study on multimedia DeepFakes //2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS). IEEE. - 2023.- pp. 1-6. DOI: 10.1109/ICAECCS56710.2023.10104814.
8. Campbell C. et al. How deepfakes and artificial intelligence could reshape the advertising industry: The coming reality of AI fakes and their potential impact on consumer behavior //Journal of Advertising Research. - 2022. - Vol. 62 (3). -pp. 241-251. DOI: 10.2501/JAR-2022-017.
9. Gregory S. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism //Journalism. - 2022.- Vol. 23 (3). - pp. 708-729.DOI: 10.1177/14648849211060644.
10. Yadlin-Segal A., Oppenheim Y. Whose dystopia is it anyway? Deepfakes and social media regulation // Convergence. - 2021.- Vol. 27 (1). - pp. 36-51. DOI: 10.1177/13548565209239