Preprints.org

Review

Not peer-reviewed version

Deepfakes and Synthetic Media: Generation, Detection, and Governance

Alexandros Gazis *, Efstathios Karypidis, Kleanthi Santamouri, Theodoros Vavouras, Nikos E. Mastorakis, Stylianos Pappas

Posted Date: 11 June 2026

doi: 10.20944/preprints202606.0925.v1

Keywords: deepfake detection; synthetic media; generative adversarial networks; diffusion models; multimedia forensics; content provenance; face reenactment; military deepfake threats; military content verification; AI governance

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Deepfakes and Synthetic Media: Generation, Detection, and Governance

Alexandros Gazis [1,2,*], Efstathios Karypidis [3], Kleanthi Santamouri [4,5], Theodoros Vavouras [6,7], Nikos E. Mastorakis [8,9] and Stylianos Pappas [9]

[1] Edinburgh Business School, Heriot-Watt University, Edinburgh EH14 4AS, UK
[2] Department of Electrical and Computer Engineering, School of Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
[3] School of Engineering, National and Technical University of Athens, 15780, Greece
[4] English Language Teacher, Private Primary Education, Athens, Greece
[5] Salvezza Energy Systems Ltd, Sirakouson 4, Paphos 8016, Greece
[6] Department of Humanities, School of Humanities, Hellenic Open University, 26335 Patras, Greece
[7] Department of Philosophy, School of Italian Language and Literature, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[8] English Language Faculty of Engineering, Technical University of Sofia, 1756 Sofia, Bulgaria
[9] Electrical Engineering and Computer Science, Hellenic Naval Academy, Terma Chatzikyriakou, 18539 Piraeus, Greece
* Correspondence: agazis@teemail.gr

Abstract

Deepfakes, synthetic audiovisual content produced by deep generative models, have escalated into a critical threat across civilian and military domains, enabling identity fraud, disinformation campaigns, and evidence fabrication. In military and high-security contexts specifically, the stakes extend to operational deception, compromised intelligence, and erosion of command trust. This entry explores how modern visual intelligence and computer vision techniques are used to detect deepfakes. It outlines key deepfake generation models, such as GANs, autoencoders, neural rendering, and diffusion systems, while also explaining how adversarial methods enhance realism and challenge existing detectors. The overview highlights visual artifacts, digital patterns, and physiological cues commonly leveraged in detection, and reviews major CNN, transformers, and frequency-based approaches. It also summarizes evaluation practices, and the difficulty of achieving strong generalization. Finally, it identifies emerging directions, including modern intelligence techniques for civilian and military content verification. This survey covers generation architectures (GANs, latent diffusion, neural rendering, video synthesis), the spatial, temporal, frequency-domain, and physiological artifacts they produce, and the detector families that exploit them. We examine evaluation benchmarks and protocols, highlighting cross-generator generalization as the field's central open challenge. Beyond detection, we discuss cryptographic provenance standards, watermarking, and regulatory frameworks (EU AI Act, DSA, GDPR). We conclude that effective deepfake governance requires defense-in-depth integrating forensic detection, verifiable provenance, and institutional accountability.

Keywords: deepfake detection; synthetic media; generative adversarial networks; diffusion models; multimedia forensics; content provenance; face reenactment; military deepfake threats; military content verification; AI governance

1. Definition, Classification and Why Deepfake is Important

The term deepfake commonly refers to synthetic or heavily manipulated media, most notably images, videos, and speech; whose realism is enabled by modern machine learning, particularly deep

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

generative models. In practice, the boundary between "deepfakes" and other manipulated media is increasingly blurred: contemporary pipelines combine generative components (e.g., diffusion or GAN-based synthesis), classical editing (compositing, retouching), and most notably, post-processing optimized for distribution environments (compression, resizing, platform-specific transcoding). This issue is really important as it may affect our everyday lives from civilian to military applications. As a result, an operational definition for researchers and practitioners is functional. A deepfake property refers to media whose perceptual plausibility is sufficiently high to create a credible false impression about who did what, when, or where, and whose creation is materially facilitated by AI-based generative or reenactment methods [1–3]. This is of great importance as for example, in military contexts, this can affect situational awareness, command communication, and operational trust.

Based on the functional definition, one can classify deepfakes into the following broad categories based on their usage, properties and functionality [4]:

1. Modality deepfakes
    a. Visual: face swaps, reenactment, lip-sync, attribute editing (age, expression), full-body synthesis, scene relighting.
    b. Audio: voice cloning (speaker identity conversion), speech-to-speech conversion, text-to-speech impersonation.
    c. Multimodal: synchronized audio–video generation, avatar systems, "talking head" models with cloned voice.
2. Manipulation intent deepfakes
    a. Identity substitution (impersonation, fraud, non-consensual sexual content).
    b. Event fabrication (false evidence, fake statements, fake presence).
    c. Contextual distortion (true footage reframed via synthetic overlays, selective edits, or deceptive narration).
3. Generation regime deepfakes
    a. Closed-world generation (trained on a specific target identity).
    b. Open-world generation (foundation models enabling broad, low-friction synthesis).

The above taxonomy, even though generic and abstract showcases real value as it suggests that detection and governance depend on which deepfake type is targeted. This distinction is also relevant for military risk assessment, where impersonation, event fabrication, and contextual distortion may produce different operational consequences. Lastly, it is noted that many failures detections methods in real cases deployments stem from treating "deepfake detection" as a single, uniform classification problem.

1.1. Why Are Deepfakes Important?

Deepfakes are more than a computer-vision issue; they are a socio-technical risk shaped by actors, incentives, and distribution channels. Analytically, in a typical threat model, an adversary has either access to a target's public media (photos, interviews, livestreams, military briefings, or open-source operational footage), commodity tools to generate or edit content, or platforms to distribute at scale [5–7]. This means that the harm is amplified by platform dynamics (virality, recommender systems) and by asymmetric verification costs (it is cheaper to generate than to thoroughly authenticate) [8,9].

A practical implication is that risk management should be framed at system level, covering creation, publishing, detection, and incident response, rather than relying on "a model that flags deepfakes" as a standalone solution. This is especially important for military and security organizations, where false positives and false negatives may influence operational decisions. Moreover, it aligns with governance-oriented frameworks that treat AI risk across the full lifecycle and operational context [10].

As such, deepfakes combine high realism with personalization [11]: they can be tailored to a specific individual and context, which increases plausibility and emotional impact. For example, in

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

military settings, personalization can be used to imitate officers, soldiers, or official communication channels. As such, even when a given deepfake is debunked, the broader informational ecosystem can still be damaged via two mechanisms:

- Evidentiary erosion: over time, authentic recordings become easier to dismiss as fake (e.g. someone might state: "this is not genuine, it could be AI").
- The liar's dividend: public figures can strategically exploit uncertainty about synthetic media to evade accountability, deny authentic evidence, or muddy public understanding [12,13].

Therefore, the deepfake problem is not only "fake media exists" [1], but "the credibility of media as evidence is destabilized". This is important both in terms of journalism or even informed civil decision but also, in regards to military communication, where it can weaken the trust in genuine battlefield footage, official announcements, or intelligence material.

1.2. Why Detection Is Necessary But Not Always Sufficient?

State-of-the-art detection has progressed rapidly, but deepfake defense is an arms race: generation improves, artifacts shift, and distribution transforms signals (compression, re-encoding) [16,17]. Deep convolutional neural networks have consistently outperformed classical feature-based approaches for visual recognition tasks [55], motivating their central role in forensic detection. For military users, this defense-in-depth model should also include forensic review, provenance checks, and command-level verification procedures. As such, for this reason, current best practice is defense-in-depth, combining [18]:

1. Forensic detection (model-based classifiers, artifact analysis, physiological/temporal cues) [1–3,19].
2. Provenance and authenticity mechanisms (cryptographic manifests, signed metadata, content credentials, watermark recovery workflows), which aim to answer to common questions such as: "where did this come from?" rather than the usual generic: "does it look fake?" [20,21].
3. Policy and compliance controls, including transparency duties for certain AI outputs and platform obligations for risk mitigation and accountability. In the EU context, deepfakes intersect directly with:
    a. the AI Act (risk-based obligations; transparency requirements for certain synthetic or manipulated content contexts) [22]
    b. the Digital Services Act (systemic risk management and transparency duties for platforms) [23], and
    c. GDPR (identity, biometric data, lawful basis, data subject rights) [24,25].

These instruments do not replace technical measures, but they shape operational requirements, documentation, and response procedures.

2. Creating Deepfakes: Types of Fakes, Standard Pipelines, and Generative Models

On a technical level, deepfakes are generated by generative modeling and reenactment/editing systems, which transform audiovisual information to produce a more convincing but inauthentic result. This is equally relevant to civilian media and military information operations. In practice, implementations combine several methods such as synthesis models (GANs/diffusion/neural rendering), classic processing stages (compositing/blending), and most recently post-processing adapted to platform environments (compression, re-encoding, scaling). As such, a more in-depth technical understanding of "how deepfakes are produced" is necessary so one can accurately map the expected artifacts, generalization failures, and appropriate evaluation protocols.

2.1. Common Deepfake Types

Based on the above, the most common functional typologies, i.e. types of deepfakes that are detected in industry and academia alike are the following:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

- Face swapping: the target's face is replaced with the source's face, with the goal of preserving the pose, lighting, and background. Historically, this was based on GAN families and evolved into high-fidelity models (e.g., the StyleGAN line) [26–28]. Such techniques are especially problematic for military briefings or crisis communication videos.
- Face reenactment (expression/motion reenactment): the target's identity is preserved, but the expression is "driven" by a source (video/live stream), often with 3D parameterization and realistic rendering [29,30].
- Lip-sync / talking-head editing: primarily the mouth region is modified to match a new audio signal; Wav2Lip is a key benchmark for synchronization under "uncontrolled" conditions [31,32].
- Audio deepfakes (voice cloning/conversion): speech is generated or transformed to mimic a specific speaker (voice conversion / TTS), often as part of a full audiovisual deepfake [33–36]. For example, in military contexts, this may support voice-based impersonation or false command messages.
- Text-to-video / full scene synthesis: video sequences (and not just "doctored" faces) are generated using diffusion-based video generation and text-to-video approaches that expand the threat from "evidence tampering" to "event fabrication" [37,38]. This is crucial as it creates a risk of fabricated military events, scenes, or operational incidents.

2.2. Typical Production Pipeline

Although the software tools and methods used vary, most production workflows typically share some common stages [40]:

1. Data collection/selection: sufficient variety of poses, expressions, and lighting (especially for identity-specific models).
2. Localization/normalization: face detection, landmarks, alignment, cropping, and photometric normalization.
3. Model training or adaptation: either general-purpose (foundation-style) or tailored to a specific individual/target.
4. Inference and temporal consistency: especially in video, temporal consistency is crucial for perceptual plausibility.
5. Post-processing: blending, color matching, denoising/oversampling, and final re-encoding, which often "hides" obvious traces and shifts detection to more subtle statistical cues.

2.3. Major Families of Generative Models

It is worth noticing that GANs, can generally be categorized into three generic categories based on the generative models used. The first is in regards to the photorealistic facial synthesis & Image-to-image translation and attribute editing. The second applies for 3D reenactment, neural rendering and latent diffusion models. The third one is connected with image to video: spatiotemporal generative modeling and 3D-aware renderings. Lastly, the fourth focuses on audiovisual deepfakes, i.e. voice cloning and lip-syncing and Implications for detection and evaluation. The aim of this section it to briefly present these families and pinpoint their unique characteristics.

As for the first family of GANS, this occurred when GANs initially introduced the competitive generator–discriminator framework and laid the foundation for realistic image synthesis [26]. The evolution toward high-fidelity architectures, such as [27] drastically increased texture detail and the stability of the generative process as GANs started to have a direct consequence to "visual plausibility" no longer being a sufficient criterion for authenticity. Moreover, unpaired image translation models, such as CycleGAN and StarGAN, have enabled feature/domain transformations without paired data and are used as building blocks for targeted modifications (e.g., expression, style, facial features) [41–43]. In such scenarios, the "identity" may remain intact, while semantically critical properties are altered.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

Secondly, as for the family of models related with 3D reenactment, neural rendering and latent diffusion models, the reenactment line was enhanced by 3D approaches: Face2Face demonstrated real-time expression transfer to RGB video through 3D parameterization and realistic rendering [29,30]. Meanwhile, Deferred Neural Rendering introduced neural textures and a learnable rendering pipeline, reducing the need for "manual" blending techniques but introducing new statistical "signatures" into the output signal [44–46]. In addition, Diffusion models (DDPM) treat synthesis as a gradual denoising process and, in many cases, offer stable training and high-quality samples [14,39,47,48]. DDIMs accelerated sampling using implicit procedures [49,50]. Practical scaling to high resolutions was achieved with latent diffusion, where generation occurs in a compressed latent space, drastically reducing computational cost [51,52]. For deepfakes, this broadens the scope from "face-centric" to generalized high-fidelity synthesis/editing.

Thirdly, in regards to the transformation from image to video, whether that is spatiotemporal generative modeling or something more complex, video generation requires modeling spatiotemporal continuity. This means that video Diffusion Models extend diffusion architectures to video, treating the temporal dimension as part of the generative process [37,38]. Furthermore, text-to-video approaches such as [38] demonstrate that video generation can rely on text–image data and unsupervised motion knowledge, reducing dependence on large text–video pairs. In practical terms, this increases the likelihood of "fully synthetic" visual evidence. Lastly, Neural Radiance Fields (NeRF) introduced continuous 3D scene representation from 2D images, enabling new viewpoints with high consistency [54]. Although not a "deepfake tool" in and of itself, 3D-aware compositing supports avatars and view-consistent renderings that reduce failures in angles/lighting, making detection based solely on obvious geometric errors more difficult.

Lastly, the fourth major family of generative deepfake models is in regards to audiovisual deepfakes, from voice cloning and lip-syncing to implications for detection and evaluation methods. Specifically, full impersonation often requires synchronizing voice and lip movements. As such, in order to transfer a speaker's tone and style, recent studies such as [33] proposes zero-shot voice style transfer using a lossy autoencoder. Similarly, for high-fidelity waveform synthesis, vocoders such as [34] enable efficient and realistic performance, while other research such as [35] serves as a classic foundation for generative modeling of raw audio signals. On the visual side, [31] significantly improves speech-to-lip synchronization in real videos. As a result, these projects pave the way for detection that leverages cross-modal consistency (sound–lips) rather than just spatial artifacts.

As a conclusion, we may note that the evolution of generative models is shifting the focus from "artifact hunting" to multi-level defense. This is evident as newer generators reduce visible traces, post-processing and platform compression alter the signals that detectors "see," and most notably the rise of text-to-video increases the need for out-of-distribution evaluation and for methods that do not rely exclusively on specific artifacts. For military applications, this also means that detection tools must be tested against both face-centric impersonation and broader synthetic event fabrication. Lastly, the proliferation of audiovisual deepfakes reinforces the importance of multimodal checks (audio–visual alignment) in detection.

3. Deepfake Properties and Characteristics

The term "deepfake" is used to describe synthetic and/or deliberately altered audiovisual content (image, video, audio), which is produced or modified using deep learning techniques in such a way as to convincingly attribute a person's identity, appearance, or voice to events that did not occur or were not recorded in that manner [56]. Although the term has become established in the public sphere, there is no absolutely uniform or universal standard that definitively defines what it does and does not encompass; in practice, deepfakes exist on a continuum between "editing" and "synthesis," where the distinction is not always clear (e.g., when a video undergoes multiple, small but systematic alterations) [56].

Specifically, for a more operational definition one may adopt the following: "content whose authenticity/ plausibility is achieved primarily through deep learning models that learn patterns of

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

human faces/movements/voices and reproduce or replace them, with the aim of misleading the viewer as to "who" appears/speaks or "what" happened [56,57].

3.1. Deepfake Categories for Audio and Visual Objects

Recent literature proposes a practical classification of facial manipulations into four main categories, which is particularly helpful for both the analysis and mapping of detection research [56–58]:

1. Entire face synthesis: the creation of an entirely new face (or even an entire image/scene) that does not correspond to a real, recorded person.
2. Identity swap / face swap: replacing one subject's face with another's, typically while preserving the "host's" pose, lighting, and motion.
3. Attribute manipulation: changes to specific characteristics (age, gender/expression, morphological features), without necessarily changing the identity.
4. Expression swap / reenactment: we "transfer" the expressions/movements of a target from a source so that the mouth, eyebrows, and micro-expressions match another video or a live source. A classic (predating modern deep models but fundamental) family of reenactment techniques is exemplified in projects such as [29].

3.2. Characteristics That "Give away" Deepfakes: Spatial, Temporal, Frequency, and Physiological Traces

Deepfakes, even when visually convincing, often leave detectable traces. These traces are not "fixed" (they depend on the model, training quality, and post-processing), but they form a central basis for the logic of visual intelligence and digital forensics. These traces of operations are the following:

1. Spatial/morphological artifacts: Examples include: imperfect blending at the face–skin/hair boundaries, unrealistic geometry in teeth/lips, inconsistencies in shading/lighting, or a "mismatched" background in high-frequency details. Such indicators are systematized in approaches that analyze visual artifacts as production "signatures" [59,60].
2. Temporal inconsistencies: In video, realism is strongly judged by frame-to-frame consistency: blink rate, facial micro-movements, lip-to-voice synchronization, head movement in relation to lighting/shading. A classic example of utilizing such a signal is the detection of unnatural blinking [61,62].
3. Frequency-domain traces: Many generative models leave traces in the frequency spectrum due to resampling and up sampling (e.g., "regularities" not systematically found in natural images). Frequency-domain analysis has been proposed as a complementary "channel" of evidence, especially when spatial cues are attenuated by compression [63,64].
4. Physiological cues: A more recent line of thinking capitalizes on the fact that real-time facial video incorporates subtle physiological changes (e.g., remote photoplethysmography based on micro-changes in skin color). Deviations in such signals can serve as an indication of synthetic origin [65,66].
5. Device/sensor origin traces (forensic fingerprints): In digital forensics, camera model fingerprints help determine whether the content bears consistent traces of physical capture or has undergone alterations that destroy or replace them [67,68].

Lastly, alongside visual deepfakes, voice cloning/synthesis (audio deepfakes) has advanced rapidly. In research, evaluation datasets and protocols such as the one presented in [69] serve as "benchmarks" for detecting synthetic/spoofed speech and for comparing metrics across methods. In realistic scenarios, the threat is often multimodal: visual and audio signals may be simultaneously altered or "work together" to increase plausibility; this also affects how we define "deepfake characteristics" in practice [56].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

4. Detection and Evaluation of Deepfakes: Methodologies, Data, and Benchmarks

Deepfake detection is a central subfield of multimedia forensics and content integrity. Unlike "closed" classification problems, deepfake detection is shaped by the continuous evolution of generative models, changes in distribution channels (platform compression, transcoding, screen recording), and the variety of operational scenarios (large-scale moderation versus forensic verification). In military use, the scenario may also involve rapid verification of battlefield media or command-related footage. As a result, evaluation cannot be limited to intra-dataset accuracy, but must systematically examine generalization, robustness, and decision calibration under "real-world" conditions [70–79].

As such, detection levels for what it means to detect a deepfake means that three functional detection objectives are identified in the literature:

1. authenticity classification (real vs. fake),
2. spatial/temporal localization of alterations, and
3. classification of forgery types (e.g., swap, reenactment, synthesis).

This distinction is important because different datasets and benchmarks support different objectives: e.g., approaches such as [80] focus on detecting a forged face, while the other line of works extends the scope to a more "universal" analysis of forgery involving multiple tasks (classification and localization) [76–78].

4.1. Cues and "Signatures" of Forgery

Detection methods utilize cues that can be grouped into four complementary categories:

1. Spatial artifacts: Cues resulting from blending, facial distortion, texture/lighting inconsistencies, or imperfect rendering of details. Works such as [59] make an effort to systematize such visual anomalies as practical detection cues, while the detection of warping artifacts has been proposed as a more specific category of cues in face manipulation [81,82].
2. Temporal consistency: In videos, deepfakes may exhibit inconsistencies in the progression of facial expressions, micro-motions, or the stability of facial features from frame to frame. A classic line of research leverages "natural" temporal regularities, such as blink rate, to reveal synthetic content [61,62].
3. Frequency-domain/spectral anomalies (frequency-domain cues): Generative pipelines often introduce patterns that are more prominent in the frequency domain, particularly due to up sampling and resampling. Frequency-domain analysis has been proposed as a complementary approach when purely visual artifacts are attenuated by compression [63].
4. Biological/physiological signals (physiological cues): The use of signals such as rPPG (micro-changes in skin color related to pulsatile blood flow) aims to provide a more "generalizable" criterion, as such signals are not always reproduced with natural consistency in synthetic videos. A prime example of this approach is [65].

Lastly, multimedia forensics examines "device/camera traces," such as camera model fingerprints as [67] that proposes a CNN-based fingerprint that can be used as an indicator of consistency/inconsistency with natural capture.

4.2. Detector Families: From Frame-Level to Multimodal

Based on how cues are utilized, detectors are broadly categorized as follows:

- Frame-level detectors (image-based detection): They focus on individual frames and offer advantages in terms of computational cost and are used in large-scale screening, but they are vulnerable to the selection of "good" frames or to deepfakes that optimize spatial artifacts [84].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

- Video-level detectors (spatio-temporal models): They incorporate temporal information (e.g., 3D CNN, temporal aggregation, transformers) and tend to improve detection in cases where the forgery "escapes" spatially but remains temporally inconsistent [85].
- Compact / mesoscopic architectures: for example, [86] presented a compact architecture for face forgery detection, targeting meso-level features that remain useful under standard compression.
- Multimodal detection (audio–video): As deepfake production shifts toward full audiovisual synthesis, the integration of audio and video becomes particularly important. Datasets such as the ones used in [79] were designed specifically for evaluating multimodal scenarios (face + voice), while anti-spoofing benchmarks for speech support systematic evaluation of synthetic/transformed speech were also developed at times [69].

4.3. Datasets and Benchmarks: What We Measure and Why It Matters

Progress in detection depends largely on the available data. Datasets vary in various terms such as the types of forgery, levels of production "realism," abundance of sources/usage rights, compression/transformations rations, and evaluation objectives. As such, it is worth noticing the following datasets for benchmarking:

- DFDC Dataset (DeepFake Detection Challenge Dataset): large-scale data for benchmarking and systematic comparison of detection approaches [71,72].
- Celeb-DF: designed as a more challenging dataset to reduce the "ease" of detection via simplistic artifacts and push for generalization [73].
- DeeperForensics-1.0: focuses on conditions that closely resemble real-world pipelines and highlights the implications of cross-dataset evaluation [74].
- WildDeepfake: compiles an "in-the-wild" collection, highlighting the drop in performance when detectors are applied to real-world internet conditions [75].
- ForgeryNet and Challenge: serves as a flexible benchmark for multiple forgery tasks (classification and localization) and as a challenge platform, enhancing the comparability of methods [76–78].
- FaceForensics++: serves as a classic evaluation dataset for manipulated facial images and is widely used in experimental comparisons [80].

4.4. Evaluation Protocols and Metrics: From AUROC to Operational Reliability

The literature agrees that evaluation must be multi-level:

1. Intra-dataset vs. cross-dataset evaluation: Cross-dataset performance better captures domain shift and is closer to real-world conditions, where the generator or compression channel is not known in advance [73–75,83].
2. Classification metrics and class imbalance: In addition to accuracy, AUROC and AUPRC are used (especially in cases of class imbalance), while in anti-spoofing scenarios, metrics such as EER are standard [69].
3. Robustness tests: Tests under re-encoding, resolution changes, cropping, filtering, and platform transformations are considered essential, as these steps often "neutralize" surface artifacts [71–75,79].
4. Calibration and decision thresholds: In operational scenarios, the output is not merely a "label," but a risk score that leads to action (e.g., human review, takedown, posting ban). Therefore, calibration and threshold selection are part of the evaluation.

Lastly, it is worth mentioning that the detection of deepfakes is evolving into an arms race. For this reason, complementary strategies are being developed that shift the focus from "assessing the likelihood of forgery" to "provenance." In this context, the C2PA specification proposes a technical framework for statements of provenance and content integrity [87], while organizations such as the CAI promote interoperability and adoption practices [88]. The combined approach (detection +

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

provenance) aims to reduce both false positives and the "strategic doubt" that arises when society is unable to distinguish between genuine and synthetic content [70,87,88].

5. Mitigation, Governance, and Regulatory Compliance for Synthetic Content

The approach to addressing deepfakes and synthetic media is shifting from "purely detection-based" approaches toward a multi-layered risk mitigation architecture (defense-in-depth). In this framework, detection functions as a single subsystem within a broader framework that includes content transparency, provenance, governance controls, and regulatory transparency/accountability requirements. The necessity of this hybrid model stems from two consistent findings in the literature:

- a detector's performance on controlled data does not guarantee operational reliability under varying dissemination channels, and
- the "trade-off" between generation and detection is dynamic, thus requiring continuous evaluation and revision of controls [89,90].

5.1. Threat Identification, Risk Modeling, and Transparency Disclosure Controls

Threat modeling and classification into risk categories into either information integrity risk (misinformation/manipulation), or identity integrity risk (impersonation, voice cloning, remote authentication), or personal harm/harassment (targeting, non-consensual material), or evidentiary risk (evidence reliability, chain of custody). As such, this taxonomy allows for the mapping of controls at each stage of the lifecycle (create–edit–publish–moderate–archive), as well as the definition of auditability and traceability requirements (e.g., who viewed, who evaluated, who decided, based on what evidence) [89,91].

This risk-based classification also provides the basis for deciding when transparency and disclosure measures should be activated as operational controls. As such, at the regulatory level, there is a growing emphasis on requirements mandating disclosure that content is artificially generated or modified when it has the potential to be misleading (i.e. with deceptive potential). Such transparency obligations serve as a "first line of defense" because they reduce the scope for deception regardless of the performance of detectors [92]. For practical implementation, the disclosure must be: salient, persistent upon redistribution, and unambiguous regarding what it denotes (i.e. in terms of semantic clarity: "AI-generated," "AI-edited," "composite," etc.), so as to avoid false impressions or over-labeling that lead to "trust fatigue".

5.2. Risk Governance, Management Systems, and Technical Content Transparency

For compliance to be operationally feasible, a structured management system with clear roles, procedures, and documentation is required. ISO/IEC 42001:2023 (AIMS) supports the establishment of policies, objectives, and controls for organizations that develop or use IT systems, while ISO/IEC 23894:2023 provides guidance on risk management specifically in the context of IT systems [93,94].

In the field of deepfakes, the implementation of these frameworks involves:

- TEVV (testing–evaluation–verification–validation) for detection/labeling/provenance tools,
- assurance case (documented argumentation that the system is "sufficiently secure/reliable" for a specific use),
- change management (what changes when the codec, model provider, or platform changes),
- monitoring & drift management (monitoring of performance degradation/increase in errors).

The essence is that the technical solution is to be transformed into a controlled system with measurable requirements and accountability mechanisms.

Within this governance structure, technical content transparency functions emerge as the evidence layer that supports verification, accountability, and auditability. As such, technical transparency (content transparency) includes approaches such as metadata-based provenance, watermarking, and tamper-evident mechanisms. One of the main concerns one must take under consideration is the distinction between:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

- detection signals (probabilistic evidence from ML detectors) and
- verifiable evidence that can be verified cryptographically or through structured manifests.

NIST AI 100-4 ([89]) summarizes technical guidelines for digital content transparency, emphasizing that no single technique is sufficient on its own; value arises from clear threat models, resilience testing, and integration into processes [84]. In practice, provenance logic also requires key management (key/signature management), key revocation/rotation policies, and mechanisms for "handling provenance gaps" (e.g., legacy material or metadata loss).

5.3. Watermarking Resilience, Operational Integration, and Forensics-Grade Documentation

Watermarking addresses the problem of persistent marking under channel distortions: compression, recoding, scaling, cropping, and redistribution. The ATSC standards for audio/video watermark embedding and for content recovery in redistribution reflect mature technical logic: clear specifications for signal embedding, recovery conditions, and operational scenarios under channel noise [95–97]. For synthetic media, the analogy is useful: a watermark is only valuable if it is accompanied by measurable robustness guarantees, a clear definition of failure modes, and evaluation procedures against removal/tampering attacks.

These robustness requirements connect watermarking to operational and forensic workflows, where technical signals must be documented, interpreted, and validated within controlled decision processes. As a result, a comprehensive approach to deepfakes requires linking detection techniques with decision workflows: when to perform automatic flagging, when human-in-the-loop intervention is required, how to define thresholds, and how to document the decision. In high-stakes environments (e.g., evidentiary use), chain-of-custody procedures, transformation logging, and reproducibility of the evaluation methodology are essential. Initiatives to evaluate analytical systems for AI-generated deepfakes reinforce precisely this dimension: standardized tests, clear limits on conclusions, and a distinction between "what a tool proves" versus "what it assumes" [90].

At the same time, ethical frameworks (UNESCO, OECD) serve as a horizontal design principle: proportionality of measures, avoidance of disproportionate side effects, accountability, and appeal/redress mechanisms, so that technical enforcement does not become a source of injustice [98,99]. Overall, effectiveness is judged as a system property: disclosure/transparency [92], risk governance [93,94], technical transparency [89], and forensics-grade documentation [90], with a legal understanding of the risks to democracy/security/privacy [91].

6. Implications and Scenarios of Abuse: A Case-Based Analysis in Key Application Areas

Deepfakes and, more broadly, synthetic media pose a challenge that transcends the "true/false" dichotomy. The critical dimension is the reconfiguration of trust mechanisms: which narrative is deemed credible, how is accountability assigned, how is the authenticity of an audiovisual document verified, and what is the organizational cost of verification. The international literature points out that the consequences are systemic, as they are linked to the speed of dissemination, personalization/targeting, and the potential for escalating disinformation at low production costs [100–104]. At the same time, the impacts vary significantly by sector, depending on whether the primary concern is the public sphere (information integrity), identity verification (identity proofing), forensics/chain-of-custody, or the protection of vulnerable groups (targeting, non-consensual material) [100–105].

6.1. Case A - News, Journalism, and Fact-Checking (Newsroom Verification)

In the news environment, deepfakes act as a catalyst for "evidentiary erosion": the increased likelihood of synthetic content amplifies uncertainty and enables the strategic denial of authentic evidence (liar's dividend), thereby undermining accountability [103,104]. In practice, verification in a newsroom is not a purely technical assessment (detector score), but a combination of several factors

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

such as dissemination and source analysis (first post, reproduction networks, signs of coordination), source/metadata verification (where available), and multimodal consistency (audio–lip sync–lighting–temporal coherence) and cross-referencing with independent contexts [105,106].

In this context, provenance approaches (e.g., Content Credentials/C2PA) are viewed as a complementary "line of defense" against exclusive reliance on detection, as they shift the focus from "does it look fake" to "what provenance history does the file carry" [106]. Lastly, the European trend toward transparency requirements for certain uses/outputs of digital content reinforces the value of documented labeling and disclosure procedures, particularly when the content has the potential to be misleading [108,109].

6.2. Case B - Financial Fraud, Corporate Security, and Remote Identity Verification (KYC/Remote Onboarding)

In financial and corporate settings, the threat shifts from the "public narrative" to operational identity fraud (impersonation fraud), where voice cloning, synthetic video calls, and biometric abuse can influence high-value decisions (e.g., payment orders, data changes, remote registration) [100–102,110]. The technical challenge is not limited to detection but rather includes the design of risk-based identity proofing and "adaptive friction" for critical actions: combining biometric and non-biometric signals, documenting controls, and establishing clear out-of-band verification channels for high-risk decisions [110,111].

From a compliance perspective, the processing of biometric data may be subject to heightened data protection requirements; therefore, organizational maturity requires documentation of the legal basis, purpose, proportionality, and security measures [112]. The literature notes that in these scenarios, resilience is achieved primarily through a combination of procedural controls and technical measures, not through a single detection tool [100–102,110].

6.3. Case C - Public Sector, Law Enforcement, and Forensic Use (Forensics, Chain of Custody)

In law enforcement and the justice system, deepfakes affect both the production of falsified evidence and the reliability of authentic records. Specific analyses for law enforcement emphasize that the response must be based on standardized chain-of-custody procedures, secure collection and storage, recording of transformations, and traceability of the verification process [113].

In this context, technical detection is part of a broader "evidence package" that includes: documented collection, hashing, logging, assessment of method limitations, and, where possible, provenance verification. This need is reinforced by the fact that modern synthesis models and platform-specific transformations (re-encoding, down sampling) can alter or eliminate traces, requiring caution in drawing conclusions and ensuring the reproducibility of the process [89,113].

6.4. Case D - Education, Academic Integrity, and the Protection of Students

In the educational setting, the implications unfold along two main lines. Firstly, the production of synthetic "evidence" (images/videos/audio) can undermine academic assessment and increase the scope of deception beyond traditional plagiarism [102,105]. Secondly, targeting with deepfakes (particularly non-consensual synthetic material and defamatory videos) has serious implications for the safety and psychosocial well-being of learners [101,102].

At the level of policy and pedagogical practice, the literature supports the integration of "evidence hygiene" procedures: mandatory documentation of material sources in assignments, the development of media literacy skills, and clear incident management workflows (reporting–recording–support–removal of reposts) [102,114]. The UNESCO ethical framework highlights the need for a human-centered approach, particularly when surveillance/detection measures may produce disproportionate side effects [114].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

6.5. Synthetic Assessment: Effectiveness as a System Property

A comparative analysis of the above cases leads to a consistent conclusion: the countermeasures against deepfakes are effective when approached as a system property rather than as a standalone tool. The international technical literature on transparency/synthetic content and provenance standards supports a multi-layered model: detection (statistical evidence), provenance (cryptographically verifiable history), and organizational risk management with documentation and continuous assessments [89,107].

In the European context, transparency obligations and accountability/due diligence requirements for platforms act as a catalyst for the institutional integration of these practices [108,109]. Similarly, guidelines on digital identity and risk-based proofing propose combinations of controls and adaptive measures that mitigate harm even when detection is not "perfect" [110,111].

7. Conclusions

The current literature shows that deepfakes have moved from an isolated forgery technique to a wider synthetic content ecosystem, where creation, dissemination, and persuasiveness interact with platforms, social networks, and real-world information inequalities [70,101]. The next phase of this domain is expected to be shaped by higher video and audio fidelity, stronger temporal consistency, and near real-time operation in avatars, teleconferences, customer service systems, and live-style broadcasts. At the same time, the convergence of image, audio, and text generation means that deepfakes are no longer only visual imitations. They can now become complete synthetic personas, with convincing speech, language, facial expression, and paralinguistic behavior. This creates serious risks for corporate fraud, political disinformation, identity abuse, and targeted harm against individuals. However, the same technologies can also support legitimate uses in film production, accessibility, education, and creative media, which makes it necessary to assess deepfakes according to use, context, intent, and potential harm [100,101].

The threat landscape is also changing as attacks are moving away from single fake videos or single-medium manipulations and toward more complete fraud workflows. These may include target data collection, voice cloning, face-swapped video production, fake documents or profiles, and social engineering. This means that the effectiveness of a deepfake does not depend only on technical quality. It also depends on the social context, the trust relationship between sender and receiver, the distribution channel, and existing inequalities or vulnerabilities [70]. In education and public communication, repeated exposure to synthetic content can increase source confusion and make it harder for users to distinguish reliable from unreliable information. For this reason, media literacy and information literacy are not secondary issues; they are part of the wider response to deepfakes [102,105].

From a technical standpoint, the main challenge is no longer only to improve detection accuracy on known datasets. The more important challenge is operational reliability under domain shift, unknown generators, new compression methods, different distribution channels, and multimodal forgeries that combine face and voice. Detection systems therefore need to be evaluated not only in controlled research settings, but also under realistic conditions. This requires unified and reproducible evaluation protocols across multiple datasets, with clear reporting of false positives, false negatives, calibration, robustness, and practical trade-offs. Benchmarks and evaluation tools such as DeepfakeBench are important because they bring together multiple datasets, protocols, and detection methods, helping to connect research performance with operational deployment [115]. At the same time, systematic surveys remain necessary because they organize the field by creation methods, detection categories, and open research gaps, including robustness, interpretability, and calibration [116].

As such, a purely detection-based approach is not enough, because detectors can fail when the content distribution changes or when new generation methods appear. For this reason, the defense architecture is increasingly moving from detection alone to provenance documentation, platform

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

transparency, and organizational risk management. Provenance standards such as the C2PA family of specifications are important because they provide a technical way to document the origin and transformation history of digital content, including generative content [112]. At the organizational level, standards such as ISO/IEC 42001 for AI management systems and ISO/IEC 23894 for AI risk management guidance provide a useful framework for policies, controls, auditing, lifecycle documentation, and continuous improvement [111,112]. In practice, this means that content integrity should be treated as a process, not only as a detector output.

Watermarking is another important part of this wider strategy, especially for diffusion and generative models. Recent approaches move watermarking closer to the production stage, so that generated content carries a detectable signature without obvious visible distortion. Stable Signature, for example, embeds a watermark into latent diffusion models by adapting the decoder, with the aim of supporting robust signature recovery after common transformations [117]. Tree-Ring Watermarking uses a fingerprint in the initial noise vector and structures it in Fourier space, aiming to remain detectable after transformations such as cropping or rotation [118]. These methods show that watermarking can support tracking and accountability. However, watermarking should not be treated as a complete solution. The literature also shows that watermarks can be weakened, removed, or attacked through fine-tuning and other techniques [119,120]. Therefore, watermarking needs continuous evaluation, clear robustness guarantees, and integration with provenance standards, platform-level measures, and governance processes [89,117–120].

Overall, the most realistic response to deepfakes is a multi-level strategy as regulation is also becoming more mature. In the European context, the trend is moving from general recommendations toward clearer transparency, accountability, and due diligence obligations. The Digital Services Act requires risk management and transparency measures, especially for very large online platforms, with the aim of reducing systemic risks linked to misinformation and harmful content distribution [109]. However, legal rules alone cannot solve the problem. Work on public discourse and democratic resilience shows that deepfake governance must combine technical tools, platform enforcement, victim support and redress mechanisms, institutional safeguards, and systematic media literacy. As such, it is worth noticing that future work should focus on evaluation under real-world conditions, robustness against adaptive attacks, clearer documentation of failure modes, and stronger links between technical standards, governance frameworks, and social resilience.

Author Contributions: All authors contributed equally to this manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. Available online: https://doi.org/10.1109/JSTSP.2020.3002101 .
2. Akhtar, Z. Deepfakes generation and detection: A short survey. *Journal of Imaging*. 2023, 9(1), 18. https://doi.org/10.3390/jimaging9010018 .
3. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* **2020**, *64*, 131–148. Available online: https://doi.org/10.1016/j.inffus.2020.06.014 .
4. Maxmudjanov, S.; Naimov, A. DEEPFAKE CONTENT TYPES AND THEIR GENERATION METHODS. *Потомки Аль-Фаргани* **2026**, *1*, 26–31. Available online: https://doi.org/10.5281/zenodo.18708717 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

5. Rainie, L.; Anderson, J.; Vogels, E.A. Experts doubt ethical AI design will be broadly adopted as the norm within the next decade. *Pew Research Center* **2021**, *16*, 121–154. Available online: https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/ (accessed on 1 June 2026).
6. Mahr, D. Sexualized deepfakes as a socio-technical continuation of gendered power. *AI & SOCIETY* **2026**, 1–13. Available online: https://doi.org/10.1007/s00146-026-03080-z .
7. Twomey, J. Socially (de) constructing deepfakes: understanding socio-technical interests and concerns in deepfake media. Ph.D. Thesis, University College Cork, Cork, Ireland, 2025. Available online: https://hdl.handle.net/10468/18856 (accessed on 1 June 2026).
8. Perriello, L.E. Blurred realities: Legal strategies for the deepfake era. *Maastricht Journal of European and Comparative Law* **2026**, 1023263X261433380. Available online: https://doi.org/10.1177/1023263X261433380 .
9. Zhang, B. Governing deepfake realities: The truth immunity framework for telecommunications policy. *Telecommunications Policy* **2026**, 103168. Available online: https://doi.org/10.1016/j.telpol.2026.103168 .
10. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. Available online: https://doi.org/10.6028/NIST.AI.100-1 .
11. Soto-Sanfiel, M.T.; Wu, Q. How audiences make sense of deepfake resurrections: A multilevel analysis of realism, ethics, and cultural meaning. Comput. Hum. Behav. **2025**, 108822. Available online: https://doi.org/10.1016/j.chb.2025.108822 .
12. Putra, A.B. The Legal Standing of Deepfake Digital Evidence in Criminal Proceedings: Challenges of Evidentiary Integrity in the AI Era. *Fox Justi: Jurnal Ilmu Hukum* **2026**, *16*, 283–290. Available online: https://ejournal.seaninstitute.or.id/index.php/Justi/article/view/8328 (accessed on 1 June 2026).
13. Cheng, E.K. Deepfakes, Photographs, and Trust in Evidence. *Virginia Law Review Online* **2025**. Available online: https://dx.doi.org/10.2139/ssrn.5771825 .
14. W. Peebles and S. Xie, ""Scalable Diffusion Models with Transformers," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 4172-4182, doi: https://doi.org/10.1109/ICCV51070.2023.00387 .
15. Barari, S.; Lucas, C.; Munger, K. Political deepfakes are as credible as other fake media and (sometimes) real media. *J. Polit.* **2025**, *87*, 510–526. Available online: https://doi.org/10.1086/732990 .
16. Soundarya, B.C.; Gururaj, H.L. Deepfake detection: critical review of state-of-the-art approaches and future perspectives. *Discov. Appl. Sci.* **2026**. Available online: https://doi.org/10.1007/s42452-025-08174-9 .
17. Sharma, S.; Selwal, A. Potential of artificial intelligence in deepfake media: From generation to detection mechanisms, state-of-the-art, and challenges. *Comput. Sci. Rev.* **2026**, *60*, 100866. Available online: https://doi.org/10.1016/j.cosrev.2025.100866 .
18. Erokhin, D.; Komendantova, N. A Review of Tools and Technologies to Combat Deepfakes. *Information* **2026**, *17*, 347. Available online: https://doi.org/10.3390/info17040347 .
19. Dhanapal, R.; Ps, A. Defending Against Adaptive Deepfake Generation and Detection Evasion Attacks. In Proceedings of the 2026 9th International Conference on Inventive Computation Technologies (ICICT), April 2026; pp. 1619–1624. Available online: https://doi.org/10.1109/ICICT68280.2026.11511164 .
20. Modi, K. The Deepfake Conundrum: Assessing Generative AI's Threat to Digital Reality and Proposing a Multi-Layered Defense Framework. *Int. J. Comput. Trends Technol.* **2025**. Available online: https://doi.org/10.14445/22312803/IJCTT-V73I6P112 .
21. Coalition for Content Provenance and Authenticity (C2PA). Content Credentials: C2PA Technical Specification, Version 2.2; C2PA, 2025. Available online: https://spec.c2pa.org/specifications/specifications/2.2/specs/_attachments/C2PA_Specification.pdf (accessed on 1 June 2026).
22. European Parliament; Council of the European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 1 June 2026).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

23. European Parliament; Council of the European Union. Regulation (EU) 2022/2065 (Digital Services Act). *Off. J. Eur. Union* **2022**, *L 277*, 1–102. Available online: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng (accessed on 1 June 2026).
24. European Parliament; Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *L 119*, 1–88. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 1 June 2026).
25. Romero Moreno, F. Generative AI and deepfakes: a human rights approach to tackling harmful content. *Int. Rev. Law Comput. Technol.* **2024**, *38*, 297–326. Available online: https://doi.org/10.1080/13600869.2024.2324540 .
26. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *arXiv* **2014**, arXiv:1406.2661. Available online: https://doi.org/10.48550/arXiv.1406.2661 .
27. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119. Available online: https://doi.org/10.1109/CVPR42600.2020.00813 .
28. Waseem, S.; Bakar, S. A. R. S. A.; Ahmed, B. A.; Omar, Z.; Eisa, T. A. E.; Dalam, M. E. E. DeepFake on face and expression swap: A review. *IEEE Access*. **2023**, *11*, 117865-117906. https://doi.org/10.1109/ACCESS.2023.3324403 .
29. Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395. Available online: https://doi.org/10.1109/CVPR.2016.262 .
30. Dhanyalakshmi, R.; Popirlan, C.I.; Hemanth, D.J. A survey on deep learning based reenactment methods for deepfake applications. *IET Image Process.* **2024**, *18*, 4433–4460. Available online: https://doi.org/10.1049/ipr2.13201 .
31. Prajwal, K.R.; Mukhopadhyay, R.; Namboodiri, V.; Jawahar, C.V. A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild. In Proceedings of the 28th ACM International Conference on Multimedia (ACM MM 2020), Virtual Event, 12–16 October 2020; pp. 484–492. Available online: https://doi.org/10.1145/3394171.3413532 .
32. Xiong, X.; Patel, P.; Fan, Q.; Wadhwa, A.; Selvam, S.; Guo, X.; et al. Talkingheadbench: A multi-modal benchmark & analysis of talking-head deepfake detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2026; pp. 4139–4149. Available online: https://doi.ieeecomputersociety.org/10.1109/WACV61042.2026.00403 (accessed on 1 June 2026).
33. Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; Hasegawa-Johnson, M. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. *arXiv* **2019**, arXiv:1905.05879. Available online: https://doi.org/10.48550/arXiv.1905.05879 .
34. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv* **2020**, arXiv:2010.05646. Available online: https://doi.org/10.48550/arXiv.2010.05646 .
35. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499. Available online: https://doi.org/10.48550/arXiv.1609.03499 .
36. Alnaqbi, M.; Ikuesan, R.A. A systematic review of audio deepfake detection techniques for digital investigation. *Discov. Comput.* **2026**, *29*, 202. Available online: https://doi.org/10.1007/s10791-026-10077-1 .
37. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video Diffusion Models. *arXiv* **2022**, arXiv:2204.03458. Available online: https://doi.org/10.48550/arXiv.2204.03458 .
38. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; Taigman, Y. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv* **2022**, arXiv:2209.14792. Available online: https://doi.org/10.48550/arXiv.2209.14792 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

39. Vu, P. T. A Fusion Model for Precipitation Nowcasting from Radar and Satellite. In *Multi-disciplinary Trends in Artificial Intelligence: 18th International Conference, MIWAI 2025, Ho Chi Minh City, Vietnam, December 3–5, **2025**, Proceedings, Part I, 16353*, 430. Springer Nature. https://doi.org/10.1007/978-981-95-4957-3_35 .
40. Li, Q.; Wang, W.; Du, S.; Peng, B.; Dong, J.; Wang, K.; et al. Towards High Fidelity Face Swapping: A Comprehensive Survey and New Benchmark. *arXiv* **2026**, arXiv:2605.00883. Available online: https://doi.org/10.48550/arXiv.2605.00883 .
41. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2242–2251. Available online: https://doi.org/10.1109/ICCV.2017.244 .
42. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797. Available online: https://doi.org/10.1109/CVPR.2018.00916 .
43. Liu, X.; Xu, R.; Chen, Y. Securing Digital Media Integrity: A Survey of Watermarking and Manipulation Detection for Image Authentication. *Authorea Preprints* **2025**. Available online: https://doi.org/10.36227/techrxiv.176186513.32946978/v1 .
44. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Trans. Graph.* **2019**, *38*, 66. Available online: https://doi.org/10.1145/3306346.3323035 .
45. He, Z.; Henderson, P.; Pugeault, N. Beyond Reconstruction: A Physics Based Neural Deferred Shader for Photo-Realistic Rendering. In Proceedings of the International Conference on Artificial Neural Networks, September 2025; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 378–389. Available online: https://doi.org/10.1007/978-3-032-04555-3_31 .
46. Svitov, D.; Dahaghin, M. NBAvatar: Neural Billboards Avatars with Realistic Hand-Face Interaction. *arXiv* **2026**, arXiv:2603.12063. Available online: https://doi.org/10.48550/arXiv.2603.12063 .
47. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239. Available online: https://doi.org/10.48550/arXiv.2006.11239 .
48. Zhang, X.; Chen, C. Parameter-efficient quantum denoising diffusion probabilistic models with temporal encoding. *Future Gener. Comput. Syst.* **2026**, *174*, 107981. Available online: https://doi.org/10.1016/j.future.2025.107981 .
49. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. *arXiv* **2020**, arXiv:2010.02502. Available online: https://doi.org/10.48550/arXiv.2010.02502 .
50. Zhang, Q.; Tao, M.; Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv* **2022**, arXiv:2206.05564. Available online: https://doi.org/10.48550/arXiv.2206.05564 .
51. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695. Available online: https://doi.org/10.1109/CVPR52688.2022.01042 .
52. Lin, Z.; Wang, X.; Zhang, S.; Wang, T.; Guo, Z.; Liu, T. TSCM: Efficient Image Synthesis Using Latent Diffusion. *Pattern Recognit. Lett.* **2026**. Available online: https://doi.org/10.1016/j.patrec.2026.02.005 .
53. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *arXiv* **2020**, arXiv:2003.08934. Available online: https://doi.org/10.48550/arXiv.2003.08934 .
54. Wang, K.; Wei, K.; Li, S.Y. Dynamic view synthesis with topologically-varying neural radiance fields from sparse input views. *Neurocomputing* **2026**, 132942. Available online: https://doi.org/10.1016/j.neucom.2026.132942 .
55. Efstathios Karypidis, Stylianos G. Mouslech, Kassiani Skoulariki, Alexandros Gazis, "Comparison Analysis of Traditional Machine Learning and Deep Learning Techniques for Data and Image Classification," WSEAS Transactions on Mathematics, vol. 21, pp. 122-130, 2022, DOI: https://doi.org/10.37394/23206.2022.21.19 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

56. Sun, X.; Xu, Z.; Li, X.; et al. Deepfake: definitions, performance metrics and standards, datasets and future directions. *Front. Big Data* **2024**, *7*, 1400024. Available online: https://doi.org/10.3389/fdata.2024.1400024 .
57. Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C.; Wu, Y.; et al. Deepfake generation and detection: A benchmark and survey. *ACM Comput. Surv.* **2026**, *58*, 1–41. Available online: https://doi.org/10.1145/3801962 .
58. Hossain, S.; Sudarsan, D.; Zaffar, H.; Ahmad, N. Social and psychological impact of deepfakes: a comprehensive bibliometric review. *Glob. Knowl. Mem. Commun.* **2026**, 1–20. Available online: https://doi.org/10.1108/GKMC-11-2024-0734 .
59. Matern, F.; Riess, C.; Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019; pp. 83–92. Available online: https://doi.org/10.1109/WACVW.2019.00020 .
60. Syed Abu Bakar, S.A.R.; Waseem, S.; Omar, Z.; Bilalashfaqahmed. Exploring the Advancements and Challenges of Deepfake Face-swap: A Survey. *Multimed. Tools Appl.* **2026**, *85*, 14. Available online: https://doi.org/10.1007/s11042-026-21285-8 .
61. Li, Y.; Chang, M.-C.; Lyu, S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 2018. Available online: https://doi.org/10.1109/WIFS.2018.8630787 .
62. Sar, A.; Roy, S.; Choudhury, T.; Abraham, A. Zero-shot visual deepfake detection: can ai predict and prevent fake content before it is created? *Found. Trends Signal Process.* **2025**, *19(3)*, 212–361. Available online: https://doi.org/10.1561/2000000136 .
63. Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; Holz, T. Leveraging Frequency Analysis for Deep Fake Image Recognition. *arXiv* **2020**, arXiv:2003.08685. Available online: https://doi.org/10.48550/arXiv.2003.08685 .
64. Hamadene, A.; Allili, M.S. Cross-Model Deepfake Detection Through Contourlet-Based Inter-Channel Spectral Analysis. *IEEE Trans. Biom. Behav. Identity Sci.* **2026**. Available online: https://doi.org/10.1109/TBIOM.2026.3687469 .
65. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. Available online: https://doi.org/10.1109/TPAMI.2020.3009287 .
66. Sharma, D.; Garg, G.; Chawla, J. Fusion of EfficientNetB0 and ResNet50 for Enhanced Deepfake Video Detection. In Proceedings of the 2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON), March 2026; pp. 1–6. Available online: https://doi.org/10.1109/I3CTCON68242.2026.11507467 .
67. Cozzolino, D.; Verdoliva, L. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 144–159. Available online: https://doi.org/10.1109/TIFS.2019.2916364 .
68. Siddiqui, N.; Islam, S. An Efficient Feature-Based Framework for Camera Model Identification. *IEEE Access* **2026**. Available online: https://doi.org/10.1109/ACCESS.2026.3690702 .
69. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech. *Comput. Speech Lang.* **2021**, *64*, 101114. Available online: https://doi.org/10.1016/j.csl.2020.101114 .
70. Langmia, K. *Black Communication in the Age of Disinformation: DeepFakes and Synthetic Media*; Springer Nature: Cham, Switzerland, 2023. Available online: https://doi.org/10.1007/978-3-031-27696-5 .
71. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Canton Ferrer, C. The DeepFake Detection Challenge (DFDC) Dataset. arXiv **2020**, arXiv:2006.07397. Available online: https://doi.org/10.48550/arXiv.2006.07397 .
72. Meta AI. Deepfake Detection Challenge Dataset (DFDC). Available online: https://ai.meta.com/datasets/dfdc/ (accessed on 1 June 2026).
73. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216. Available online: https://doi.org/10.1109/CVPR42600.2020.00327 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

74. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, 13–19 June 2020; pp. 2889–2898. Available online: https://doi.org/10.1109/CVPR42600.2020.00296 .
75. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.-G. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), Virtual Event, 12–16 October 2020; pp. 2382–2390. Available online: https://doi.org/10.1145/3394171.3413769 .
76. He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; Liu, Z. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, 20–25 June 2021; pp. 4360–4369. Available online: https://doi.org/10.1109/CVPR46437.2021.00434 .
77. He, Y.; Sheng, L.; Shao, J.; Liu, Z.; et al. ForgeryNet—Face Forgery Analysis Challenge 2021: Methods and Results. *arXiv* **2021**, arXiv:2112.08325. Available online: https://doi.org/10.48550/arXiv.2112.08325 .
78. CodaLab. ForgeryNet—Face Forgery Analysis Challenge 2021 (Competition Page). Available online: https://competitions.codalab.org/competitions/33386 (accessed on 1 June 2026).
79. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. *arXiv* **2021**, arXiv:2108.05080. Available online: https://doi.org/10.48550/arXiv.2108.05080 .
80. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. Available online: https://doi.org/10.1109/ICCV.2019.00009 .
81. Li, Y.; Lyu, S. Exposing DeepFake Videos by Detecting Face Warping Artifacts. *arXiv* **2018**, arXiv:1811.00656. Available online: https://doi.org/10.48550/arXiv.1811.00656 .
82. Bai, W.; Liu, Y.; Zhang, A.; Wang, Y.; Li, B.; Hu, W.; Zhang, Z. Deepfake Detection via Exploring Degradation Inconsistency. *IEEE Trans. Inf. Forensics Secur.* **2026**. Available online: https://doi.org/10.1109/TIFS.2026.3695433 .
83. Z. Yan, Y. Luo, S. Lyu, Q. Liu and B. Wu, Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 8984-8994, doi: https://doi.org/10.1109/CVPR52733.2024.00858 .
84. Tariq, R.; Heo, M.; Tariq, S.; Woo, S. Through the Lens: Benchmarking Deepfake Detectors Against Moiré-Induced Distortions. *Adv. Neural Inf. Process. Syst.* **2025**, *38*. Available online: https://proceedings.neurips.cc/paper_files/paper/2025/hash/75c4f24bf8a51f0b0870f0f9bceea4ea-Abstract-Datasets_and_Benchmarks_Track.html (accessed on 1 June 2026).
85. Alanazi, S.; Asif, S. VIDS-Guard: A Novel Forensics-Aware Multi-Stream Transformer Framework for Robust Deepfake Video Detection. *Int. Syst. Appl.* **2026**, *200664*. Available online: https://doi.org/10.1016/j.iswa.2026.200664 .
86. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: A Compact Facial Video Forgery Detection Network. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018. Available online: https://doi.org/10.1109/WIFS.2018.8630761 .
87. Mesa-Simón, M., Escobar-Molero, A., Sáez-Mingorance, B., Morales, D. P., Álvarez-Bermejo, J. A., & Romero, F. J. Enabling live video provenance and authenticity: A C2PA-based system with TPM-based security for livestreaming platforms. *IEEE Transactions on Multimedia,* **2026**, 1-12. https://doi.org/10.1109/TMM.2026.3679049 .
88. Content Authenticity Initiative (CAI). Video Provenance and the Ethics of Deepfakes (Blog post). 2021. Available online: https://contentauthenticity.org/ (accessed on 1 June 2026).
89. Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. Evaluating trustworthiness in AI: Risks, metrics, and applications across industries. *Electronics*, **2025**,*14*(13), 2717. https://doi.org/10.3390/electronics14132717 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

90. Guan, H.; Horan, J.; Zhang, A. *Guardians of Forensic Evidence: Evaluating Analytic Systems Against AI-Generated Deepfakes*; Forensics@NIST 2025; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2025. Available online: https://www.nist.gov/publications/guardians-forensic-evidence-evaluating-analytic-systems-against-ai-generated-deepfakes (accessed on 1 June 2026).
91. Chesney, R.; Citron, D. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. Law Rev.* **2019**, *107*, 1753–1819. Available online: https://doi.org/10.2139/ssrn.3213954 .
92. European Commission. *AI Act Service Desk: Article 50—Transparency Obligations for Providers and Deployers of Certain AI Systems*; European Commission: Brussels, Belgium, 2025. Available online: https://artificialintelligenceact.eu/article/50/ (accessed on 1 June 2026).
93. International Organization for Standardization (ISO); International Electrotechnical Commission (IEC). *ISO/IEC 42001:2023—Information Technology—Artificial Intelligence—Management System*; ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en (accessed on 1 June 2026).
94. International Organization for Standardization (ISO); International Electrotechnical Commission (IEC). *ISO/IEC 23894:2023—Information Technology—Artificial Intelligence—Guidance on Risk Management*; ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/77304.html (accessed on 1 June 2026).
95. Advanced Television Systems Committee (ATSC). *A/334—Audio Watermark Emission*; ATSC: Washington, DC, USA, 2025. Available online: https://www.atsc.org/atsc-documents/a3342016-audio-watermark-emission/ (accessed on 1 June 2026).
96. Advanced Television Systems Committee (ATSC). *A/335—Video Watermark Emission*; ATSC: Washington, DC, USA, 2025. Available online: https://www.atsc.org/atsc-documents/a3352016-video-watermark-emission/ (accessed on 1 June 2026).
97. Advanced Television Systems Committee (ATSC). *A/336—Content Recovery in Redistribution Scenarios*; ATSC: Washington, DC, USA, 2024. Available online: https://www.atsc.org/atsc-documents/a3362017-content-recovery-redistribution-scenarios/ (accessed on 1 June 2026).
98. Ganbaatar, U. Do Ethics in AI Still Matter? A Review of the 2021 UNESCO Recommendation on the Ethics of AI. *The Review of Faith & International Affairs*, **2025**, *23*(3), 26-33. https://doi.org/10.1080/15570274.2025.2531638 .
99. OECD. *Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449)*; OECD: Paris, France, 2019. Available online: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 (accessed on 1 June 2026).
100. Gaur, L., Ed. *DeepFakes: Creation, Detection, and Impact*; CRC Press: Boca Raton, FL, USA, 2022. Available online: https://doi.org/10.1201/9781003231493 .
101. Schick, N. Deepfakes: The Coming Infocalypse; Twelve (Hachette Book Group): New York, NY, USA, 2020; ISBN 978-1538754313. Available online: https://www.hachettebookgroup.com/titles/nina-schick/deepfakes/9781538754313/?lens=twelve (accessed on 1 June 2026).
102. Mooney, C. *AI and Deception: Plagiarism, Deep Fakes, and More*; ReferencePoint Press: San Diego, CA, USA, 2025; ISBN 978-1-6782-1066-3. Available online: https://catalog.pcpls.org/Record/23319106 (accessed on 1 June 2026).
103. Kietzmann, J.; Lee, L.W.; McCarthy, I.P.; Kietzmann, T.C. Deepfakes: Trick or treat? *Bus. Horiz.* **2020**, *63*, 135–146. Available online: https://doi.org/10.1016/j.bushor.2019.11.006 .
104. Vaccari, C.; Chadwick, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Soc. Media Soc.* **2020**, *6*, 2056305120903408. Available online: https://doi.org/10.1177/2056305120903408 .
105. Gregory, J. *The Trouble with Deepfakes*; Cherry Lake Publishing: Ann Arbor, MI, USA, 2024; ISBN 978-1668946992. Available online: https://cherrylakepublishing.com/shop/show/53868 (accessed on 1 June 2026).
106. Chandra, B.; Dunietz, J.; Roberts, K. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency* (NIST AI 100-4); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024. Available online: https://doi.org/10.6028/NIST.AI.100-4 .

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2026 doi:10.20944/preprints202606.0925.v1

107. Coalition for Content Provenance and Authenticity (C2PA). *C2PA Technical Specification, Version 1.3*; C2PA, 2023. Available online: https://spec.c2pa.org/specifications/specifications/1.3/specs/_attachments/C2PA_Specification.pdf (accessed on 1 June 2026).
108. European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act)—Transparency obligations (Art. 50). Available online: https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-50 (accessed on 1 June 2026).
109. de Araujo Meirelles Magalhães, F., & Calle, I. M. Consumer Protection in the Digital Age: A Reflection on Regulation 2022/2065 (Digital Services Act—DSA) and Agenda 2030. In *International Conference a Digital Europe for Citizens, Data governance, Digital Markets, Digital Services*, **2026**, 171-188. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-02500-5_10 .
110. ENISA. *Remote ID Proofing Good Practices*; European Union Agency for Cybersecurity: Athens/Heraklion, Greece, 2024. Available online: https://www.enisa.europa.eu/sites/default/files/2024-11/Remote%20ID%20Proofing%20Good%20Practices_en_0.pdf (accessed on 1 June 2026).
111. NIST. SP 800-63A: Digital Identity Guidelines—Identity Proofing and Enrollment; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2017. Available online: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63a.pdf (accessed on 1 June 2026).
112. European Union. Regulation (EU) 2016/679 (General Data Protection Regulation)—Art. 9 (special categories of personal data). Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 1 June 2026).
113. Europol Innovation Lab. *Facing Reality? Law Enforcement and the Challenge of Deepfakes*; Europol: The Hague, The Netherlands, 2022. Available online: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf (accessed on 1 June 2026).
114. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (accessed on 1 June 2026).
115. Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; Wu, B. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. *arXiv* **2023**, arXiv:2307.01426. Available online: https://doi.org/10.48550/arXiv.2307.01426 .
116. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–41. Available online: https://doi.org/10.1145/3425780 .
117. Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; Furon, T. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023. Available online: https://doi.org/10.48550/arXiv.2303.15435 .
118. Wen, Y.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *arXiv* **2023**, arXiv:2305.20030. Available online: https://doi.org/10.48550/arXiv.2305.20030 .
119. Hu, Y.; Jiang, Z.; Guo, M.; Gong, N. Stable Signature is Unstable: Removing Image Watermark from Diffusion Models. *arXiv* **2024**, arXiv:2405.07145. Available online: https://doi.org/10.48550/arXiv.2405.07145.
120. Ci, H.; et al. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-Key Identification. *arXiv* **2024**, arXiv:2404.14055. Available online: https://doi.org/10.48550/arXiv.2404.14055 .

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.