Research Article

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Deepfake Technology and the Erosion of Digital Trust: A Multi-Layered Defence Framework

Muhammed Seyrek[a,*], **Merve Yıldırım**[b]

Submitted: 03.07.2026 **Revised:** 28.08.2026 **Accepted:** 28.08.2026 **doi:**10.30855/gmbd.070526A09

a,* Department of Artificial Intelligence, Erzurum Technical University, 25050 – Erzurum, Türkiye, Orcid: 0009-0006-2620-8952
b Department of Software Engineering, Karadeniz Technical University, 61080 – Trabzon, Türkiye, Orcid: 0000-0001-6834-3470
Corresponding author: muhammed.seyrek13@erzurum.edu.tr

Keywords: Deepfake, synthetic media, generative adversarial networks, AI governance, media forensics, cybersecurity, disinformation, media literacy.

ABSTRACT

Deepfake technology, enabled by generative adversarial networks (GANs) and subsequent diffusion-based models, has rapidly evolved from a technical curiosity into a substantial cybersecurity threat. While prior literature has examined detection algorithms, legal regulation, and media literacy as separate avenues of response, this paper argues that a single-layer defence is structurally insufficient against the threat. This study combines a qualitative integrative review and conceptual framework with a proof-of-concept experiment that illustrates the behaviour of the technical layer. Drawing upon documented incidents — including the 2022 Zelensky video, the 2019 UK CEO voice fraud case, and the proliferation of non-consensual synthetic imagery — and supported by reported figures from Sensity AI, Sumsub, and academic surveys of the field, we develop an integrated three-layer defence framework that combines (i) technical detection, (ii) legal and regulatory measures, and (iii) media literacy and individual verification habits. Rather than merely listing these layers, we specify for each layer its objective, instruments, responsible actors, strengths, limitations, and applicability conditions, and we articulate the decision logic that links them. The paper synthesises the existing literature, analyses the threat landscape across four structural dimensions (democratic integrity, institutional trust, corporate security, and individual rights), and discusses implementation challenges. A proof-of-concept experiment on a public benchmark of 140,000 real and GAN-generated faces illustrates the framework's central claim: raising the confidence required of the technical layer improves its accuracy from 95.3% to 99.2%, but only by escalating 19.1% of cases to the remaining layers. The analysis therefore indicates that no single defensive layer is sufficient; rather, the convergence of all three is necessary to mitigate the evolving deepfake threat.

1. Introduction

The question of whether a video can still serve as evidence — in courts, in journalism, in political discourse — has become genuinely uncertain. Since 2017, when the term "deepfake" first appeared on Reddit, advances in deep learning have made the synthesis of fabricated audio-visual content increasingly accessible. What once required substantial computational resources and specialised expertise can today be accomplished using consumer-grade applications. This democratisation has fundamentally altered the threat surface.

The implications extend well beyond technical novelty. Synthetic media has been weaponised in active conflict zones, deployed for high-value financial fraud, and used at scale to target private individuals — predominantly women —

To cite this article: F. M. Seyrek and M. Yıldırım, "Deepfake Technology and the Erosion of Digital Trust: A Multi-Layered Defence Framework", *Gazi Journal of Engineering Sciences,* vol.12, no.2, pp. 285-299, 2026. doi:10.30855/gmbd.070526A09

through non-consensual content. These are not hypothetical scenarios; each represents documented harm with significant human and institutional cost [1,2].

Existing literature has explored three principal lines of defence: (i) algorithmic detection of synthetic content [3], (ii) legal and regulatory frameworks [1,4], and (iii) public education and media literacy [5]. Each strand has been treated largely in isolation, with proponents advocating for one approach as the primary line of mitigation. The central argument of this paper is that this fragmented framing is structurally insufficient. A single-layer defence — whether algorithmic, legal, or educational — leaves systematic gaps that the adversarial nature of the threat will exploit. We propose, instead, an integrated three-layer model in which technical detection, legal regulation, and media literacy operate in concert.

The contribution of this study is not the identification of the three layers individually — each is well established in the literature — but their specification as a single analytical model. Concretely, the study makes four contributions. First, it maps each documented limitation in the literature to the defence layer best able to address it, making explicit which gap motivates which layer. Second, it specifies each layer in terms of its objective, instruments, responsible actors, strengths, limitations, and applicability conditions, rather than describing the layers narratively. Third, it articulates an explicit inter-layer decision logic: a case that the technical layer cannot resolve with confidence is not simply a failure but is escalated to the legal and literacy layers, so that the weakness of each layer is matched to the strength of another. Fourth, it organises the threat itself along four structural dimensions and demonstrates their interdependence, showing why a single-layer response cannot be adequate.

In keeping with this aim, the paper is primarily a conceptual and review-based contribution: it synthesises existing evidence into an analytical framework rather than reporting a new detection system. A focused proof-of-concept experiment is included to illustrate one empirical property of the technical layer, but it is subordinate to the conceptual argument and is not intended as the paper's primary contribution. Unlike existing multi-layered accounts, which typically enumerate the three defences or advocate one as primary, the specific contribution here is an explicit inter-layer decision rule together with a measured parameter — the confidence threshold — that determines when a case passes from the technical layer to the legal and literacy layers.

The paper is organised as follows. Section 2 reviews the relevant literature on deepfake generation, detection, regulation, and societal impact. Section 3 describes the review approach and the criteria used for case and source selection. Section 4 presents the analysis: the technical architecture of GANs, documented real-world cases, reported figures on the scale of the threat, the four-dimensional threat landscape, and the proposed three-layer defence framework specified as an analytical model. Section 5 discusses implementation, stakeholder roles, and limitations. Section 6 concludes.

2. Literature Review

2.1. Generative models and the origins of deepfake

The technical foundation of contemporary deepfake content lies in generative adversarial networks (GANs), first introduced by Goodfellow et al. [6] and subsequently extended through architectures such as StyleGAN [7], CycleGAN, and conditional GANs. Westerlund [8] traced the early emergence of these techniques from research environments to consumer applications, noting that the timeline from academic publication to public accessibility has shortened considerably over the past decade. More recent surveys, particularly Tolosana et al. [3] and Mirsky and Lee [9], have catalogued the rapid diversification of synthesis methods, including face swapping, voice cloning, lip-sync manipulation, and full-body synthesis. The introduction of diffusion-based models from 2022 onwards has further accelerated the realism of generated outputs while reducing computational requirements, a shift documented in recent surveys of the field [10-12] and reflected in the difficulty of detecting the most recent generation of fakes [13,14].

2.2. Detection approaches

Tolosana et al. [3] provide the most comprehensive survey of detection methods, classifying them along the dimensions of physiological inconsistency (e.g., abnormal blinking rates, inconsistent gaze direction), biological artefacts (e.g., unnatural skin texture, pulse extraction), and signal-level anomalies (e.g., compression artefacts, GAN fingerprinting). Verdoliva [15] complements this work by surveying media forensic approaches that examine pixel-level inconsistencies and metadata signatures left by generative pipelines. Korshunov and Marcel [16] further demonstrate that even modern face recognition systems are vulnerable to deepfake-based spoofing attacks. However, these works explicitly note the limitations of purely algorithmic approaches: detection and generation are locked in an adversarial loop in which each improvement in detection provides an implicit training signal for improved generation. This dynamic, which we term the detection–generation arms race, fundamentally constrains the long-term viability of any purely technical defence. Recent large-scale surveys reinforce this conclusion: Masood et al. [10] and Mubarak et al. [11] catalogue the proliferation of generation and detection methods across visual, audio, and textual modalities, while Altuncu et al. [17] highlight the absence of shared definitions and evaluation standards as a structural obstacle to progress. Most tellingly, the Deepfake-Eval-2024 benchmark [13] shows that state-of-the-art detectors trained on academic datasets degrade sharply when confronted with deepfakes actually circulating in real-world settings, confirming that detection performance in controlled conditions overstates real-world robustness.

2.3. Legal and regulatory responses

Chesney and Citron [1] offered an early and influential treatment of the legal dimensions, framing deepfakes as a threat to privacy, democratic processes, and national security. Their analysis anticipated several of the cases that subsequently materialised. The European Union's Artificial Intelligence Act (2024) [4] has introduced mandatory transparency requirements for synthetic media, and several U.S. states (notably California, Texas, and Virginia) have enacted targeted legislation against non-consensual deepfake content. At the federal level, the United States enacted the TAKE IT DOWN Act in May 2025, criminalising the distribution of non-consensual intimate imagery, including AI-generated content, and requiring platforms to remove such material promptly. However, the regulatory landscape remains highly fragmented internationally, and legislative coverage in many jurisdictions is largely underdeveloped — a gap that this paper considers further in Section 5. Perriello [18], in a comparative analysis spanning the European Union, United States, and China, reaches a convergent conclusion: transparency obligations alone are insufficient against malicious actors, cross-border disinformation, and intimate-image abuse, and an adaptive, multi-layered governance model is required — a position that directly supports the framework developed here.

2.4. Disinformation, trust, and media literacy

Vaccari and Chadwick [19] empirically examined the impact of synthetic political video on audience deception, uncertainty, and trust in news media. Their experiments showed that even when deepfakes do not directly deceive viewers, they erode generalised trust in genuine video content. This finding has significant implications: the harm of deepfake technology is not exhausted by successful deception, but extends to the broader degradation of epistemic confidence. Fallis [20] develops this argument philosophically, characterising deepfakes as a structural epistemic threat that undermines our collective capacity to acquire knowledge from recorded media. Citron [21] further argues that the erosion of confidence in visual evidence has particular consequences for the privacy and dignity of those targeted by non-consensual synthetic content. Hwang et al. [22] provide empirical evidence that media literacy education has a measurable protective effect against deepfake-based disinformation, supporting the inclusion of this layer in our integrated framework. A related and comparatively under-examined harm is the so-called liar's dividend: Schiff et al. [23] demonstrate across five survey experiments with over 15,000 participants that public figures can gain support by falsely dismissing genuine evidence as fabricated, a second-order effect that persists even when no deepfake is actually present. Pawelec [24] situates these dynamics within democratic theory, and Twomey et al. [25] provide one of the first real-world analyses of epistemic trust erosion during the 2022 Russia–Ukraine conflict. UNESCO [5] has emphasised the role of digital and media literacy as a complementary intervention, particularly in contexts where regulatory infrastructure is incomplete. Building on this, Halgamuge et al. [26] argue that societal resilience can be actively strengthened through narrative inoculation — pre-emptively exposing audiences to manipulation techniques so that

they are less susceptible when encountering synthetic content — which provides a concrete mechanism for operationalising the media-literacy layer.

3. Review Approach and Framework Development

This study is primarily a conceptual contribution and does not claim to follow a systematic review protocol; it therefore does not report databases, Boolean search strings, or PRISMA-style record counts. Its core method is the integrative review, an established approach for synthesising a diverse body of literature and generating new conceptual frameworks rather than aggregating homogeneous empirical results. Following the stages commonly associated with this method, the study proceeds through (i) problem formulation, (ii) purposive literature and case selection, (iii) critical evaluation and thematic organisation of the evidence, and (iv) synthesis into a conceptual framework. To this we add a fifth, subordinate step: a proof-of-concept experiment (Section 4.6) that illustrates the behaviour of the framework's technical layer. The experiment is not intended to validate the framework as a whole — the legal and literacy layers are not amenable to this kind of measurement — but to test one specific claim on which the model rests. The remainder of this section makes each step transparent so that the basis of the framework can be independently assessed.

Problem formulation. The review was organised around a single guiding question: given that technical, legal, and educational responses to deepfakes are each individually limited, how can they be combined into a coherent defensive model? This question determined both the thematic scope of the literature searched and the criteria against which candidate cases and sources were judged.

Source selection. We prioritised peer-reviewed publications and authoritative institutional reports addressing four themes: deepfake generation, technical detection, legal and regulatory governance, and societal impact including disinformation and media literacy. Priority was given to high-impact and recent contributions (with emphasis on 2020–2025 for this fast-moving field), while foundational earlier works (e.g., the original GAN formulation) were retained for conceptual grounding. Sources were included when they directly informed at least one of the four themes and excluded when they addressed deepfakes only tangentially or lacked verifiable sourcing. Beyond core detection papers, we deliberately incorporated literature on synthetic-media governance, content provenance and watermarking (C2PA), identity fraud, and media-literacy interventions, so that all three defence layers rest on a comparable evidence base rather than the technical layer alone.

Case selection. Three documented cases were selected purposively, using four explicit criteria: (i) each represents a distinct harm vector (political manipulation, financial fraud, individual privacy violation); (ii) each is verifiable through reputable journalistic or research reporting; (iii) each has had a demonstrable real-world impact; and (iv) together they span different geographies and different dimensions of the threat. The aim of this purposive selection is analytical illustration of the threat's breadth, not statistical representativeness; the cases are used to ground, not to prove, the framework.

Use of reported figures. Quantitative statements are drawn from named industry and institutional reports (e.g., Sensity AI, Sumsub, World Economic Forum) and, where possible, cross-referenced against a second source. For each figure we state the source, the measured category, and its limitation. Because the deepfake ecosystem changes rapidly and some 2024–2026 datasets remain proprietary, these figures are treated as indicative baselines rather than precise or current measurements.

Synthesis and framework development. The three-layer framework was derived analytically: each structural threat dimension identified in the literature was mapped to the defensive approach best able to address it, and for each approach we identified the residual gap that the other layers must cover. This explicit mapping — which documented limitation motivates which layer — is summarised in Table 1 and specified in full in Section 4.5, where each layer is characterised by its objective, instruments, responsible actors, strengths, limitations, and applicability conditions.

Proof-of-concept experiment. To examine the framework's claim that the technical layer is necessary but bounded, we conducted a proof-of-concept experiment on the publicly available "140k Real and Fake Faces" benchmark, which pairs authentic photographs with faces synthesised by StyleGAN. The data were used as distributed: 100,000 images for training, 20,000 for validation, and 20,000 held out for testing, balanced between classes. Three architectures were trained under identical conditions and evaluated on the same held-out set: a baseline convolutional network trained from scratch, and two transfer-learning models (MobileNetV2 and EfficientNet-B0) fine-tuned from ImageNet weights, each with its own required input preprocessing. Training used the Adam optimiser with early stopping on validation loss. Performance is reported using accuracy, precision, recall, F1 score, and AUC. We then treated each model's output probability as a confidence signal, defining confidence as the distance of the predicted probability from the decision boundary, in order to measure how the layer behaves when it is required to be more certain. The experiment is deliberately conventional: its purpose is not to propose a novel detector but to instantiate Layer 1 with representative methods so that its limits can be observed.

Table 1. Mapping from documented limitations in the literature to the corresponding defence layer.

Documented limitation in the literature	Corresponding defence layer	Key sources
Detectors lose accuracy as generation improves (detection–generation arms race); performance degrades on unseen, in-the-wild fakes	L1: Technical detection — necessary but bounded; complemented by provenance (C2PA)	[3,13,14,27]
Harmful synthetic media crosses borders faster than any single jurisdiction can respond; accountability is fragmented	L2: Legal and regulatory framework — deterrence, labelling, redress	[1,4]
Harm persists even without deception (erosion of trust, liar's dividend); technical and legal measures cannot reach the individual's judgement	L3: Media literacy and individual verification habits	[5,22,23,25]

4. Analysis and Discussion

4.1. The technical architecture of GAN-based Deepfakes

A generative adversarial network consists of two neural networks trained in opposition: a generator network that produces synthetic candidate samples, and a discriminator network that evaluates whether a given sample is genuine or synthesised. During training, both networks improve iteratively as each adapts to the other's progress. Over thousands of iterations, the generator learns to produce outputs that the discriminator can no longer reliably distinguish from real samples. The structural relationship between these components is illustrated in Figure 1.

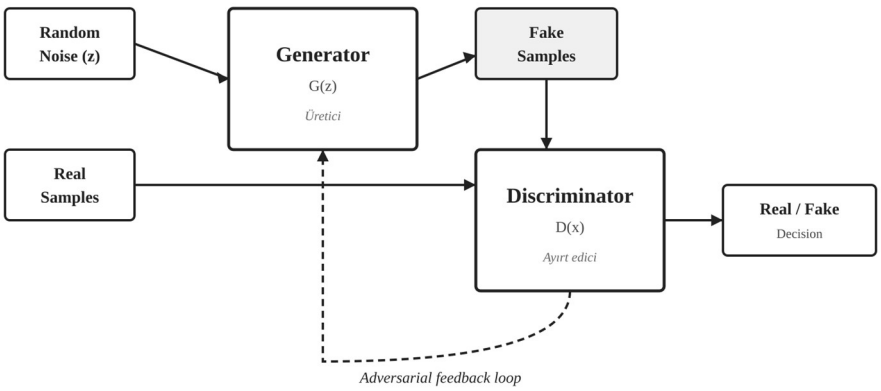


Figure 1. The adversarial training architecture of a Generative Adversarial Network (GAN). The generator produces synthetic samples from random noise; the discriminator evaluates real and synthetic inputs and provides a feedback signal that drives iterative improvement in the generator.

From this base architecture, three principal synthesis categories have emerged. Face swapping replaces the face in an existing video with that of another individual. Voice cloning imitates the vocal characteristics and prosody of a target speaker. Full synthesis generates entirely artificial individuals who have never existed. Each category presents distinct detection challenges and enables distinct categories of harm.

4.2. Documented real-World cases

Three documented cases illustrate the breadth of the threat:

(i) The Zelensky video (2022). In March 2022, during an active phase of the Russia–Ukraine conflict, a fabricated video depicting the Ukrainian President calling on his forces to surrender was broadcast via a hacked news channel and circulated on social media [28]. Although the video was of low quality and debunked within hours, its propagation during a kinetic conflict illustrated the speed at which synthetic content can outpace correction; Twomey et al. [25] provide a detailed analysis of its circulation and its effect on public trust.

(ii) The CEO voice fraud case (2019). A UK-based chief executive authorised a wire transfer of approximately €220,000 following a telephone call from an AI-generated voice imitating his parent company's managing director. The case is significant not for the victim's negligence — none was alleged — but for the demonstration that biometric voice familiarity, long considered a reliable trust signal, no longer offers meaningful authentication. The threat has since escalated in scale: in early 2024 the engineering firm Arup lost approximately US$25 million after an employee joined a video conference in which the chief financial officer and several colleagues were all AI-generated deepfakes, authorising fifteen transfers before the fraud was detected [29].

(iii) Non-consensual synthetic content (2019–ongoing). According to Sensity AI [2], the majority of online deepfake content consists of non-consensual material targeting identifiable individuals, predominantly women. While this category receives less institutional attention than political or financial misuse, it constitutes the most numerically prevalent harm. A recent ten-country study by Umbach et al. [30] documents the prevalence of non-consensual synthetic intimate imagery and the limited public awareness surrounding it, underscoring that the harm is both widespread and under-reported.

4.3. Reported figures on the scale of the threat

Several reported indicators illustrate the scale and trajectory of the threat. Table 2 summarises figures drawn from named, publicly accessible sources; for each indicator the measured category and its source are stated, and each should be read as an indicative figure rather than a precise measurement.

Table 2. Reported indicators of deepfake threat scale and detection difficulty, with source and measurement scope.

Indicator	Value	Context	Source
Non-consensual content	≈ 96%	Share of online deepfake content that is non-consensual	Sensity AI (2019) [2]
Corporate fraud cases	× 10	Year-on-year rise in deepfake-enabled fraud (2022–23)	Sumsub (2023) [31]
Human detection difficulty	high	Humans struggle to reliably distinguish high-quality fakes from authentic media	Tolosana et al. (2020) [3]

The trajectory is further illustrated by the increase in detected deepfakes reported by Sumsub between 2022 and 2023 (Figure 2). These figures refer specifically to the year-on-year change in the number of detected deepfake attempts identified in Sumsub's identity-verification pipeline, aggregated by region; they measure detected cases, not total prevalence. North America and the Asia-Pacific region recorded the steepest increases, but every region measured showed growth of several hundred per cent within a single year [31].

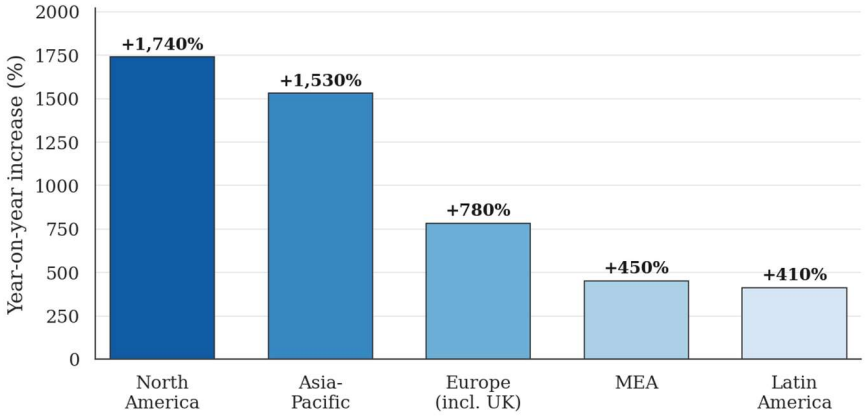


Figure 2. Year-on-year increase in detected deepfakes by region, 2022–2023. Data: Sumsub Identity Fraud Report 2023 [31].

The composition of online deepfake content underscores that the most prevalent harm is directed at individuals rather than institutions. As shown in Figure 3, Sensity AI's landscape analysis found that the overwhelming majority of online deepfake material is non-consensual synthetic content.

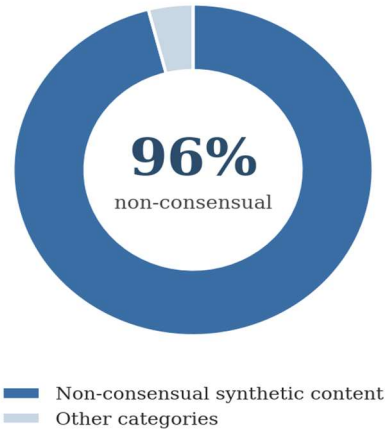


Figure 3. Composition of online deepfake content. Data: Sensity AI (2019) [2].

More recent reporting confirms that these conditions have worsened rather than eased. Sumsub's 2025–2026 reporting documents continued year-on-year growth in detected deepfake-enabled fraud [32], and the World Economic Forum notes that deepfake-enabled fraud is associated with losses exceeding US$200 million in the first quarter of 2025 alone, alongside a reported 1,740% surge in North American cases between 2022 and 2023 [33]. These figures should be read as indicative rather than exhaustive — measurement definitions differ across sources — but the direction of travel is consistent. We therefore treat the figures reported here as conservative baselines rather than upper bounds.

4.4. The four-dimensional threat landscape

Synthesising across the cases and reported indicators, the deepfake threat operates across four structural dimensions, summarised in Table 3.

Table 3. Four-dimensional structure of the deepfake threat with representative cases and references.

Dimension	Nature of Harm	Representative Case	Reference
Democratic integrity	Synthetic political content reaches mass audiences during crises before it can be corrected	Zelensky surrender video, 2022	[25,28]
Institutional trust	Evidentiary value of recordings in courts and journalism is structurally weakened	General epistemic erosion	[1]
Corporate security	Executive impersonation enables unauthorised transactions and system access	UK CEO voice fraud, 2019	[1,34]
Individual rights	Violation of dignity, privacy, and psychological integrity through non-consensual content	Non-consensual synthetic content, ongoing	[2]

These four dimensions are not independent: a successful corporate impersonation [34] simultaneously damages institutional trust [1], while a politically motivated deepfake [19] may also implicate individual rights [2]. The European Union Agency for Cybersecurity has documented similar interdependencies across the broader AI threat landscape [35], indicating that an effective response cannot be limited to any single layer of defence.

4.5. A three-layer integrated defence framework

We propose that effective response requires the integration of three complementary defensive layers, none of which is sufficient in isolation. The structure of the proposed framework is illustrated in Figure 4. What distinguishes this framework from earlier treatments is not the identification of the three layers individually — each is well established — but their specification as a single analytical model with an explicit decision logic. Prior work typically advocates one layer as primary or lists the three side by side; here we define, for each layer, its objective, instruments, responsible actors, strengths, and limitations (Table 4), and we specify how a case moves between layers. The need for such an integrated approach is increasingly recognised: recent work by Gazis et al. [36] independently concludes that an effective response to synthetic media must combine technical tools, platform enforcement, victim support and redress, institutional safeguards, and systematic media literacy, while Sharif et al. [37] propose a governance framework in which technical detection and institutional oversight are treated as interdependent rather than alternative measures. Our contribution is to organise these elements into a single model with an explicit inter-layer decision logic.

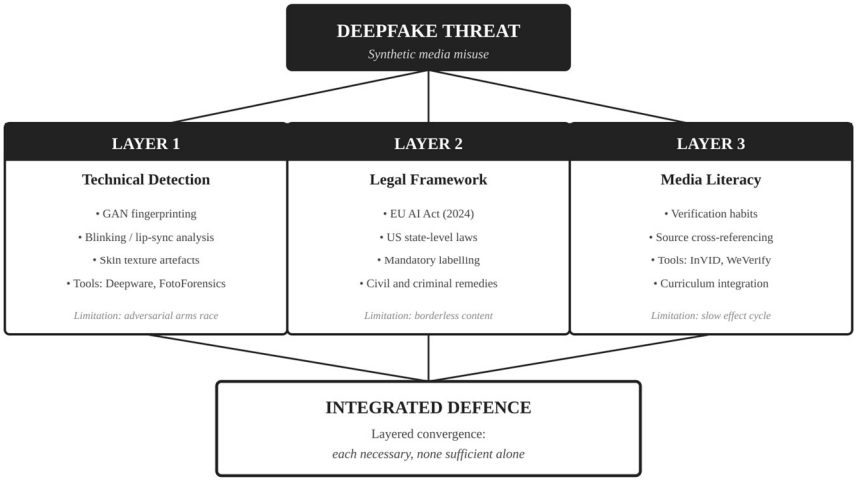


Figure 4. The proposed three-layer integrated defence framework. Technical detection, legal frameworks, and media literacy each address part of the deepfake threat; their convergence produces an integrated defence in which the limitations of each layer are compensated by the strengths of the others.

Table 4. Analytical specification of the three defence layers: objective, actors, strengths, and limitations.

Layer	Objective & Instruments	Responsible Actors	Strengths	Limitations & Applicability
L1: Technical detection	Flag synthetic media via detectors (physiological, signal-level) and provenance/watermarking (C2PA) [3,27]	Platforms, forensic labs, tool developers	Fast, automatable, scalable to high volume	Detection–generation arms race; degrades on unseen fakes; needs continual retraining
L2: Legal & regulatory	Deter and sanction misuse; mandate labelling; enable takedown and victim redress [4]	Legislators, regulators, courts, platforms	Creates accountability and victim remedy	Jurisdictionally fragmented; slow; weak against cross-border, anonymous actors
L3: Media literacy	Build verification habits; teach cross-referencing; support tools (InVID, WeVerify) [5,22]	Educators, media, civil society, individuals	Most scalable and durable; reaches cases tools and law miss	Slow to take effect; uneven reach; hard to measure

Layer 1 — Technical detection. Automated systems analyse physiological, biological, and signal-level indicators of synthesis. Publicly available tools such as Deepware Scanner and FotoForensics provide accessible verification

capability. However, as Tolosana et al. [3] document, the detection–generation arms race precludes any permanent technical advantage; detection therefore remains a necessary but bounded defence. A complementary technical direction shifts the emphasis from detecting falsity to certifying authenticity at the point of capture: the Coalition for Content Provenance and Authenticity (C2PA) standard [27] embeds cryptographically verifiable provenance metadata into media, offering a proactive alternative to reactive detection.

Layer 2 — Legal and regulatory framework. The EU AI Act [4] establishes mandatory transparency and labelling obligations for synthetic media — requiring that AI-generated content be marked as such — and several jurisdictions have introduced targeted legislation against non-consensual deepfake content. These instruments create accountability and deter casual misuse. Their effectiveness, however, is constrained by jurisdictional boundaries, particularly given the borderless nature of online distribution.

Layer 3 — Media literacy and individual verification habits. Tools such as InVID and WeVerify enable individual users to cross-reference suspicious content. More fundamentally, the cultivation of verification habits — pausing before sharing, treating urgency as a warning signal, cross-referencing across trusted sources — addresses the gaps that technical and legal measures cannot reach. As UNESCO [5] notes, this layer is the most scalable and the most sustainable, although it produces effect more slowly than technical or legal interventions.

Decision logic linking the layers. The three layers are not parallel alternatives but a sequence with fallback. A suspect item is first assessed technically (Layer 1); where detection or provenance yields a confident verdict, the case is resolved efficiently and at scale. Where the technical layer is inconclusive — the region in which the arms race bites hardest — the case does not simply fail, but is escalated: legal instruments (Layer 2) provide deterrence and redress for identifiable, in-jurisdiction actors, while media-literacy practices (Layer 3) equip the recipient to withhold judgement and verify independently. In this design the weakness of each layer is deliberately matched to the strength of another, which is precisely why no single layer, deployed alone, is adequate.

To assess whether this specification adds anything beyond existing treatments, Table 6 compares the proposed framework with four representative approaches drawn from the recent literature. The comparison focuses on the two properties that most clearly separate the approaches: how many of the three layers each covers, and whether it specifies an explicit rule for moving between them. The comparison shows convergence on the need for multi-layered defence — which supports the direction taken here — but also a consistent gap: existing approaches either concentrate on one layer or enumerate several without specifying how they interact. The contribution of this paper is located precisely in that gap.

Table 6. Comparison with representative existing approaches, showing that only the proposed framework specifies an explicit rule for moving between layers.

Approach	Layer coverage	Explicit inter-layer decision logic?
Detection-centred work [3,13,14]	L1 only	No
Provenance / watermarking (C2PA) [27]	L1 (proactive)	No
Comparative legal analysis [18]	L2 (argues for multi-layered governance)	No
Governance syntheses [36,37]	L1+L2+L3 (enumerated)	No
This paper	L1+L2+L3 (specified)	Yes: confidence-based escalation from L1 to L2/L3

4.6. Proof-of-concept validation of the technical layer

The framework asserts that technical detection is necessary but structurally bounded, and that its residual uncertainty must be absorbed by the other two layers. This is an empirical claim about Layer 1, and it can be examined directly. Table 7 reports the performance of the three architectures on the held-out test set of 20,000 images.

Table 7. Detection performance of the three architectures on the held-out test set (20,000 images; higher is better).

Model	Accuracy	Precision	Recall	F1	AUC
Baseline CNN (from scratch)	0.9210	0.8851	0.9675	0.9245	0.9801
MobileNetV2 (fine-tuned)	0.9436	0.9340	0.9547	0.9442	0.9877
EfficientNet-B0 (fine-tuned)	0.9532	0.9678	0.9377	0.9525	0.9920

All three models perform well, and the ordering is as expected: fine-tuned transfer-learning architectures outperform the network trained from scratch, with EfficientNet-B0 achieving the highest accuracy (95.32%) and AUC (0.9920). The ROC curves in Figure 5 confirm that the separation between classes is strong across the full operating range for all three models.

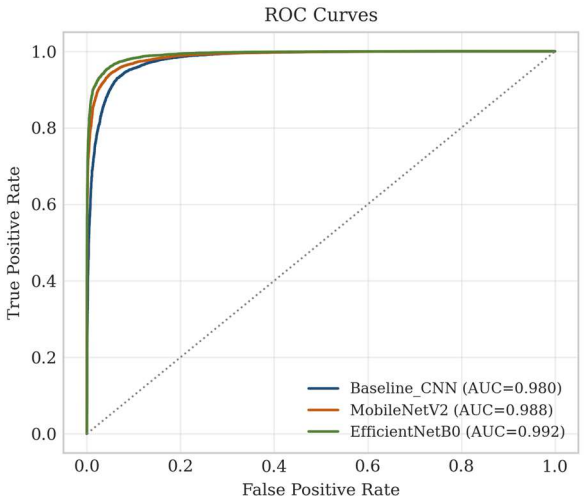


Figure 5. ROC curves for the three architectures on the held-out test set. All models achieve high discrimination, with EfficientNet-B0 performing best (AUC = 0.992).

These figures might appear to suggest that the technical layer is sufficient on its own. Two findings indicate otherwise. First, the errors do not vanish: as Figure 6 shows, the best-performing model still misclassifies 935 of 20,000 images, and its errors are asymmetric — 623 synthetic faces are accepted as authentic, against 312 authentic faces rejected as synthetic. In a deployment context, those 623 cases are precisely the harmful ones, and they pass the technical layer without objection.

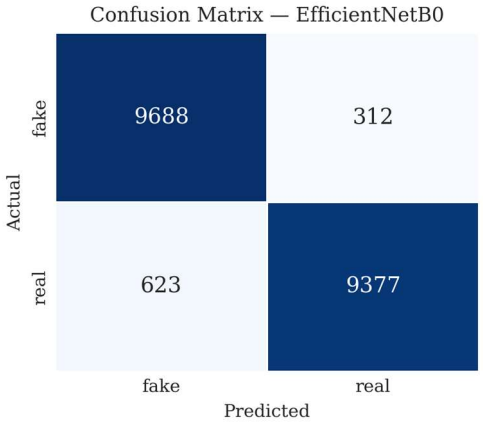


Figure 6. Confusion matrix for EfficientNet-B0 on the held-out test set. Errors are asymmetric: synthetic faces accepted as authentic (623) outnumber authentic faces rejected as synthetic (312).

Second, and more directly relevant to the framework, the accuracy of the technical layer can be raised only by declining to decide. Treating the distance of the model's output from the decision boundary as a confidence signal, Table 8 and Figure 7 report what happens as the layer is required to be more certain. At a threshold of zero — the layer decides everything — accuracy is 95.32%. Raising the required confidence to 0.9 lifts accuracy to 99.24%, an improvement of 3.9 percentage points, but only because 19.11% of all cases are no longer decided automatically. The relationship is monotonic across the range: every gain in reliability is purchased with a larger escalated share. There is no setting at which the technical layer is both highly reliable and complete.

Table 8. Reliability of the technical layer against escalation rate, as a function of the confidence required. Auto-handled and escalated shares sum to 100%.

Confidence threshold	Cases decided automatically (%)	Cases escalated to Layers 2-3 (%)	Accuracy of automated decisions
0.0	100.00	0.00	0.9532
0.1	98.83	1.17	0.9581
0.2	97.69	2.31	0.9633
0.3	96.48	3.52	0.9674
0.4	95.08	4.92	0.9724
0.5	93.61	6.39	0.9767
0.6	91.86	8.13	0.9805
0.7	89.87	10.13	0.9842
0.8	86.65	13.35	0.9885
0.9	**80.89**	**19.11**	**0.9924**

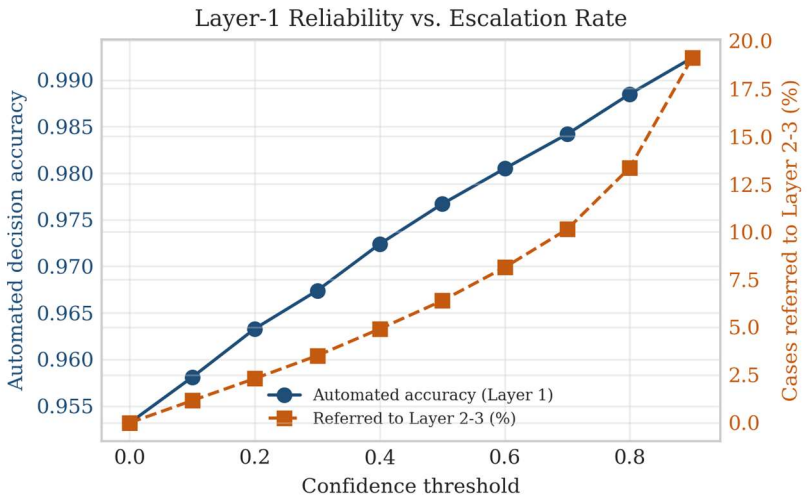


Figure 7. Accuracy of automated decisions (left axis) against the proportion of cases escalated to Layers 2 and 3 (right axis), as a function of the confidence required of the technical layer. Higher reliability is purchased only by escalating more cases.

This is the framework's central claim rendered in numbers. The escalated share is not a defect of the particular models used here; it is the operational form of the detection–generation arms race described in Section 2.2, and it is what the legal and media-literacy layers exist to absorb. The result gives the inter-layer decision logic of Section 4.5 a concrete parameter: the confidence threshold is the point at which a case ceases to be a technical matter and becomes an institutional and human one. Where that threshold should sit is a policy choice rather than a technical one, and it will differ by context — a platform moderating at scale and a court assessing evidence will not tolerate the same error rate. Critically, the escalated cases do not disappear once handed on: their fate depends on what the remaining layers can actually offer, which is the subject of Section 5.

The scope of this experiment should be stated plainly, and it is deliberately narrow. It instantiates Layer 1 only, and only on still facial images generated by GANs drawn from a single public benchmark. It does not address video, audio, or diffusion-generated content, each of which has distinct artefacts and detection dynamics, and performance on a curated benchmark is known to overstate performance on material encountered in the wild [13]. The experiment is therefore offered strictly as an illustration of a structural property — that raising a detector's confidence forces escalation — and not as a claim about detector capability in general or about the framework as a whole. Layers 2 and 3 remain supported conceptually and, for the legal layer, through the comparative analysis of Section 5 rather than through measurement.

5. Implications and Discussion

The framework specified in Section 4.5 carries practical and policy-level implications that can be expressed in terms of who does what. For platforms and tool developers (Layer 1), the implication is that detection should be deployed with an explicit awareness of its limits: confidence should be reported, provenance standards such as C2PA adopted, and

detectors retrained continually rather than treated as a one-off solution. For legislators, regulators, and courts (Layer 2), the priority is cross-border cooperation and clear labelling and redress mechanisms, since a borderless threat is poorly addressed by isolated national statutes. For educators, media, and civil society (Layer 3), the implication is sustained investment in verification habits embedded in formal curricula rather than one-off awareness campaigns.
Two brief scenarios illustrate how the escalation logic would operate in practice. In a newsroom verifying a leaked video before publication, the technical layer would screen the clip; a high-confidence synthetic verdict supports a decision not to publish, whereas a low-confidence result triggers escalation to human verification (Layer 3) and, if the subject is identifiable and harmed, to legal remedy (Layer 2). In a financial-institution setting facing a deepfake voice or video authorising a transfer, the same threshold governs whether an automated check clears the request or routes it to human confirmation and, where jurisdiction permits, to reporting under fraud and personal-data law. In both cases the confidence threshold is the operational hand-off point, and its placement reflects the tolerance for error appropriate to the setting.

The framework also has limitations that should be stated plainly. It is primarily a conceptual model derived from the literature. The proof-of-concept experiment (Section 4.6) supports one specific claim about the technical layer, but the framework as a whole is not empirically validated: the legal and media-literacy layers are argued from the literature rather than measured, and the inter-layer hand-off has been demonstrated only on still images from a single benchmark. The relative weight of each layer is likely to vary by context, threat type, and available institutional capacity, and the model does not prescribe fixed thresholds. Success criteria — for example, reduction in the reach of undetected synthetic content, or improved user verification behaviour — are proposed here conceptually and would require empirical study to operationalise.

Table 5. Comparative overview of selected regulatory instruments applicable to synthetic media.

Jurisdiction	Principal instrument	Deepfake-specific?	Scope and enforcement mechanism
European Union	Regulation (EU) 2024/1689 (AI Act), Art. 50 [4]	Partly (transparency)	Binding disclosure duty: deployers must label AI-generated or manipulated content; exemptions for artistic and satirical works; penalties tied to global turnover
United States (federal)	TAKE IT DOWN Act (2025)	Yes (NCII only)	Criminalises distribution of non-consensual intimate imagery, including synthetic; platforms must operate notice-and-takedown within 48 hours
Kazakhstan	Law No. 230-VIII on Artificial Intelligence (in force 18 Jan 2026) [38]	Partly (labelling)	First dedicated AI statute in Central Asia; synthetic outputs may be distributed only with machine-readable labelling and a user warning; administrative liability for non-disclosure
Uzbekistan	Law No. ZRU-1115 (2026) amending laws for AI regulation [39]	No (indirect)	Introduces a statutory definition of AI and ethical rules for AI use; protection against synthetic media operates indirectly via personal-data and informatisation law
Türkiye	Law No. 6698 (Personal Data); Turkish Penal Code; Law No. 5651	No (indirect)	No deepfake-specific statute in force; protection rests on personal-data, penal and internet-publication provisions applied by analogy; a dedicated bill remains pending

A concrete illustration of why Layer 2 cannot be assumed to be uniformly available is provided by the current regulatory landscape, summarised in Table 5. Even among jurisdictions that have legislated recently, the instruments differ in kind. The European Union imposes a transparency and labelling duty under Article 50 of the AI Act [4]. The United States addresses one harm category through the federal TAKE IT DOWN Act, backed by a 48-hour platform takedown obligation. Kazakhstan, with Law No. 230-VIII in force from January 2026, became the first Central Asian state to enact a dedicated AI statute, permitting distribution of synthetic outputs only where they carry machine-readable labelling and an explicit user warning [38]. Uzbekistan introduced a statutory definition of AI and ethical rules for AI use in 2026, but addresses synthetic media only indirectly [39]. Türkiye, by contrast, has no deepfake-specific statute in force: protection is assembled indirectly from personal-data legislation (Law No. 6698), the Penal Code, and Law No. 5651 on internet publications.

Two implications follow, and the experiment makes them concrete. Section 4.6 showed that a well-performing detector still hands on roughly one case in five when required to be highly reliable. Table 5 shows what those escalated cases

meet when they arrive at the legal layer, and the answer depends entirely on jurisdiction. The same synthetic image, escalated by the same detector at the same threshold, encounters a binding labelling duty in the European Union, a 48-hour takedown obligation in the United States if it constitutes non-consensual intimate imagery, a machine-readable labelling requirement in Kazakhstan, and, in Türkiye, no deepfake-specific provision at all — only personal-data, penal, and internet-publication rules applied by analogy. The first implication is therefore that the legal layer is not a constant but a variable: its strength depends on where the harm occurs, which means the framework must be read as context-dependent rather than universal. The second is that where deepfake-specific legislation is absent or indirect, the burden of the escalated share shifts disproportionately onto the technical and literacy layers. In such settings the inter-layer hand-off does not disappear; it is simply resolved with one fewer instrument, which raises the practical importance of provenance standards and verification habits. This is an argument for building all three layers in parallel rather than sequencing them.

6. Conclusion

Deepfake technology is unlikely to recede; its tools will continue to become cheaper, more accessible, and more convincing. The relevant question is therefore not whether the threat will dissipate, but how it should be governed. This paper has argued that any single-layer defence — algorithmic, legal, or educational — is structurally insufficient, and has specified an integrated three-layer response as an analytical model rather than a mere list.

Technical detection contains the boundary; legal frameworks establish accountability; and media literacy closes the gaps that the other two cannot reach. The contribution of this paper lies in specifying how these three fit together: each layer is characterised by its objective, actors, strengths, and limitations, and an explicit decision logic governs how a case moves between them when the technical layer cannot resolve it with confidence. The proof-of-concept experiment in Section 4.6 gives this argument an empirical anchor: the technical layer could be made 99.24% accurate only by escalating 19.11% of cases to the layers behind it. The analysis therefore suggests that the durability of a society's defence against synthetic media depends less on any single layer than on the way the three are combined. That recent independent work [18,36] converges on similarly multi-layered conclusions suggests the framework is well aligned with the direction of the field. Future research should extend the empirical evaluation to video, audio, and diffusion-generated deepfake content and to material encountered in the wild, and beyond the technical layer, by conducting country-level policy analysis, or by developing institutional verification protocols that operationalise the three layers.

Conflict of Interest Statement

The authors declare that there is no conflict of interest

Authorship Contribution Statement

M.S.: Conceptualization, Methodology**,** Software, Supervision; **M.D:** Validation, Investigation, Writing – Editing & Original Draft; **M.Y.:** Formal analysis, Visualization, Project administration, Writing – Review

References

[1] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 6, pp. 1753-1820, 2019.

[2] Sensity AI, "The State of Deepfakes: Landscape, Threats and Impact," Amsterdam: Sensity AI Report, 2019.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131-148, 2020.

[4] European Parliament, "Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act)," *Official Journal of the European Union*, Brussels, 2024.

[5] UNESCO, "Guidance for Generative AI in Education and Research," Paris: UNESCO Publishing, ED-2023/WS/56, 2023.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672-2680.

[7] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401-4410.

[8] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39-52, 2019.

[9] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, Article 7, 1–41, 2021.

[10] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974-4026, 2023.

[11] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access, 11, 14pp. 4497-144529*, 2023.

[12] B. Liu, B. Liu, T. Zhu, and M. Ding, "A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models," *International Journal of Intelligent Systems*, vol. 2025, no. 1, Article 9987535, 2025. doi: 10.1155/int/9987535

[13] N. A. Chandra, R. Murtfeldt, L. Qiu, et al., "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," *arXiv:2503.02857*, 2025.

[14] Q. Xu, D. Chen, J. Chen, S. Lyu, and C. Wang, "Recent Advances on Generalizable Diffusion-Generated Image Detection," *arXiv preprint arXiv:2502.19716*, 2025.

[15] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910-932, 2020.

[16] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv:1812.08685*, 2018.

[17] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: Definitions, Performance Metrics and Standards, Datasets, and a Meta-Review," *Frontiers in Big Data, 7*, 1400024, 2024.

[18] L. E. Perriello, "Blurred Realities: Legal Strategies for the Deepfake Era," *Maastricht Journal of European and Comparative Law*, 2026. doi: 10.1177/1023263X261433380

[19] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, vol. 6, no. 1, 2020.

[20] D. Fallis, "The Epistemic Threat of Deepfakes," *Philosophy & Technology*, vol. 34, no. 4, pp. 623-643, 2021.

[21] D. K. Citron, "The Fight for Privacy: Protecting Dignity, Identity, and Love in the Digital Age," *New York: W. W. Norton & Company*, 2022.

[22] Y. Hwang, J. Y. Ryu, and S. H. Jeong, "Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education," *Cyberpsychology*, Behavior, and Social Networking, vol. 24, no. 3, pp. 188-193, 2021.

[23] K. J. Schiff, D. S. Schiff, and N. S. Bueno, "The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?," *American Political Science Review*, vol. 119, no. 1, pp. 71-90, 2025.

[24] M. Pawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society*, vol. 1, no. 2, Article 19, 2022.

[25] J. Twomey, D. Ching, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy, "Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets that Discuss Deepfakes in the Russian Invasion of Ukraine," *PLOS ONE*, vol. 18, no. 10, e0291668, 2023.

[26] M. N. Halgamuge, T. Moulahi, and P. Lorenz, "Societal Resilience: Inoculating People Against Deepfake Manipulation," in *Mitigating the Risks of AI Deepfakes, Boca Raton: CRC Press / Taylor & Francis*, 2026.

[27] Coalition for Content Provenance and Authenticity (C2PA), "Content Credentials: Technical Specification," *C2PA*, 2024.

[28] B. Allyn, "Deepfake Video of Zelenskyy Could Be Tip of the Iceberg in Info War, Experts Warn," NPR, 16 March 2022.

[29] T. C. Helmus, "Artificial Intelligence, Deepfakes, and Disinformation: A Primer," *Santa Monica: RAND Corporation, PE-A1043-1*, 2022.

[30] R. Umbach, N. Henry, G. Beard, and C. Berryessa, "Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries," *Proceedings of the CHI Conference on Human Factors in Computing Systems, Article 743, 1–20*, 2024.

[31] Sumsub, "Identity Fraud Report 2023," London: Sumsub Research, 2023.

[32] Sumsub, "Identity Fraud Report pp. 2025-2026," London: Sumsub Research, 2025.

[33] World Economic Forum, "Detecting Dangerous AI Is Essential in the Deepfake Era," Geneva: World Economic Forum, 2025.

[34] Europol, "Facing Reality? Law Enforcement and the Challenge of Deepfakes," *The Hague: Europol Innovation Lab*, 2022.

[35] European Union Agency for Cybersecurity (ENISA), "AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence," *Heraklion: ENISA*, 2020.

[36] A. Gazis, E. Karypidis, K. Santamouri, T. Vavouras, N. E. Mastorakis, and S. Pappas, "Deepfakes and Synthetic Media: Generation, Detection, and Governance," *Encyclopedia*, vol. 6, no. 8, Article 165, 2026. doi: 10.3390/encyclopedia6080165

[37] H. Sharif, A. Atif, and A. A. Nagra, "Deepfake-Style AI Tutors in Higher Education: A Mixed-Methods Review and Governance Framework for Sustainable Digital Education," *Sustainability*, vol. 17, no. 1, Article 234, 2025. doi: 10.3390/su17010234

[38] Republic of Kazakhstan, "Law No. 230-VIII on Artificial Intelligence," signed 17 November 2025, in force 18 January 2026, Astana, 2026.

[39] Republic of Uzbekistan, "Law No. ZRU-1115 on Amendments and Supplements to Certain Laws in Connection with Artificial Intelligence Regulation," adopted 21 January 2026, Tashkent, 2026.