



**ҚЫЛМЫСТЫҚ ҚҰҚЫҚ. ҚЫЛМЫСТЫҚ процесс / Criminal law.
Criminal procedure/ Уголовное право. Уголовный процесс**

Scientific article
IRSTI 10.85.31

<https://doi.org/10.32523/2616-6844-2026-156-3-171-187>

Assessment of the potential of artificial intelligence technologies and other modern methods for detecting deepfake materials in fraud and other unlawful acts

K.M. Beaver*^{}

College of Criminology and Criminal Justice, Florida State University, USA, Florida

(e-mail: kevinmichaelbeaver@gmail.com)

Abstract. The article assesses how far artificial intelligence and other modern methods can detect deepfake material used in fraud and other unlawful acts. It describes the main forms of synthetic media: face swapping, facial reenactment, voice cloning, full-face synthesis and forged biometric credentials. It then compares the available detection tools: convolutional neural networks (XceptionNet), vision transformers, frequency-domain analysis, biological signal analysis and cryptographic provenance (C2PA). These tools perform almost flawlessly on benchmarks, but their accuracy falls on real-world material. The causes are lossy compression, poor generalisation to unfamiliar data, and vulnerability to adversarial attacks. Recent benchmark evaluations demonstrate that Vision Transformers exhibit superior cross-dataset generalization (11.33% drop) compared to CNNs (>15% drop), while classical machine learning models like Random Forest achieve 99.64% accuracy with 2 ms inference time. The procedural analysis focuses on the law of the Republic of Kazakhstan, covering the admissibility of electronic evidence and the distinct roles of experts and specialists. The article concludes that a detector's probabilistic output cannot stand as independent evidence. It becomes usable only with expert interpretation, explainable AI (XAI) and a verified chain of custody. The proposals include a qualifying feature in Article 190 of the Criminal Code (or an aggravating circumstance in Article 54), mandatory labelling of synthetic content, validated forensic methodologies, an interagency reference database and anti-deepfake safeguards in banking.

Keywords: deepfake, synthetic media, fraud, artificial intelligence, detection, electronic evidence, forensic examination.

Received: 18.09.2026. Accepted: 23.09.2026. Available online: 30.09.2026.

К.М. Бивер*

*Флорида штатының университетінің Криминология және қылмыстық сот төрелігі
колледжі, АҚШ, Флорида
(e-mail: kevinmichaelbeaver@gmail.com)*

Алаяқтық және өзге де құқыққа қайшы әрекеттер жасау кезінде дипфейк материалдарын анықтау үшін жасанды интеллект технологиялары мен басқа да заманауи әдістерді қолдану мүмкіндіктерін бағалау

Аңдатпа. Мақалада алаяқтық пен өзге де құқыққа қайшы әрекеттерде пайдаланылатын дипфейк-материалдарды анықтау үшін жасанды интеллект технологиялары мен заманауи әдістердің мүмкіндіктері бағаланады. Синтетикалық медианың негізгі түрлері сипатталады: бетті ауыстыру, бет-әлпет реэнактменті (мимиканы көшіру), дауысты клондау, толық синтезделген бейнелер және жалған биометриялық құжаттар. Анықтау әдістері салыстырылады: үйірткілі нейрондық желілер (XceptionNet), визуалды трансформерлер, жиіліктік талдау, биологиялық сигналдарды талдау және контенттің шығу тегін криптографиялық растау (C2PA). Зертханалық жағдайда бұл жүйелердің дәлдігі жоғары, бірақ нақты материалдарда ол төмендейді. Себептері – файлдарды шығынмен сығу, жаңа деректерге жалпылау қабілетінің төмендігі және қарсыластық шабуылдарға осалдық. Процестік бөлім Қазақстан Республикасының заңнамасына негізделген. Онда электрондық дәлелдемелердің жол берілетіндігі мен сарапшы және маман рөлдерінің аражігі қарастырылады. Автор автоматты детектордың ықтималдық нәтижесі өз бетінше дәлелдеме бола алмайды деген қорытындыға келеді. Ол үшін сарапшылық түсіндіру, түсіндірілетін ЖИ (ХАИ) және дәлелдемелерді сақтаудың расталған тізбегі қажет. ҚР ҚК 190-бабын саралаушы белгімен (немесе 54-бапты ауырлататын мән-жаймен) толықтыру, синтетикалық контентті міндетті таңбалау, жаңа сот-сараптамалық әдістемелер, ведомствоаралық үлгілер базасы және банк секторындағы қорғаныс шаралары ұсынылады.

Түйін сөздер: дипфейк, синтетикалық медиа, алаяқтық, жасанды интеллект, анықтау, электрондық дәлелдемелер, сот сараптамасы.

Бивер К.М.*

*Колледж криминологии и уголовного правосудия Университета штата Флорида,
США, Флорида
(e-mail: kevinmichaelbeaver@gmail.com)*

Оценка возможностей применения технологий искусственного интеллекта и иных современных методов для выявления дипфейк-материалов при совершении мошенничества и иных противоправных деяний

Аннотация. Статья оценивает, насколько технологии искусственного интеллекта и другие современные методы способны выявлять дипфейк-материалы, используемые для мошенничества и иных противоправных деяний. Описаны основные разновидности синтетических медиа: подмена лица, лицевой реэнактмент (перенос мимики), клони-

рование голоса, полностью синтетические лица и поддельные биометрические документы. Сопоставлены методы детекции: сверточные нейронные сети (XceptionNet), визуальные трансформеры, частотный анализ, анализ биологических сигналов, а также криптографическое подтверждение происхождения контента (C2PA). В лабораторных условиях эти системы почти безошибочны, однако на реальных материалах их точность падает. Причины – сжатие файлов с потерями, слабая переносимость на незнакомые данные и уязвимость к состязательным атакам. Процессуальная часть опирается на законодательство Республики Казахстан. В ней рассмотрены допустимость электронных доказательств и разграничение ролей эксперта и специалиста. Автор приходит к выводу, что вероятностный результат автоматического детектора не может служить самостоятельным доказательством. Для этого нужны экспертная интерпретация, объяснимый ИИ (XAI) и подтвержденная цепочка хранения доказательств. Предложено дополнение ст. 190 УК РК квалифицирующим признаком (либо ст. 54 – отягчающим обстоятельством), а также обязательная маркировка синтетического контента, новые судебно-экспертные методики, межведомственная база образцов и меры защиты в банковском секторе.

Ключевые слова: дипфейк, синтетические медиа, мошенничество, искусственный интеллект, детекция, электронные доказательства, судебная экспертиза.

Introduction

Synthetic media has changed faster in the past few years than almost any other field of applied computing. Generative Adversarial Networks, diffusion models, and neural vocoders were built for film post-production, dubbing, game design, and similar lawful work. The same tools can now fabricate a face, a voice or a whole scene with enough realism that an ordinary viewer or listener has little chance of noticing. Content of this kind is usually called a deepfake, and it has moved well beyond entertainment. It now features in fraud and in a growing range of other unlawful conduct.

For offenders, the appeal is practical. A scheme that once needed skilled forgers or insiders can now be run by someone with a laptop and an open-source model. Voice cloning, built on text-to-speech synthesis and voice conversion, lets a caller sound like a specific person: a company director authorising a payment, or a son phoning his mother after a supposed accident. That is the basis for Business Email Compromise, chief executive officer fraud and the «relative in distress» scams. Video is exploited in a similar way. Real-time face-swapping and motion-transfer tools allow an impostor to appear on a video call as a manager, a friend, or a state official. Generative tools are also used against institutions themselves. They produce forged identity documents and synthetic biometric traits aimed at Know-Your-Customer procedures and Automatic Speaker Verification systems in banks and other parts of critical digital infrastructure. These attacks combine convincing sound and image with remote, automated delivery and no physical contact. Their target is wider than individual victims: they erode public administrative order and the everyday trust on which social and economic exchange depends [1].

The figures bear this out. The Federal Bureau of Investigation Internet Crime Complaint Center, in its Annual Report for 2025, recorded more than 22,000 complaints that referred

to artificial intelligence, with adjusted losses above \$893 million. Investment fraud using AI-generated media accounted for most of that sum, over \$632 million. AI-enabled Business Email Compromise caused losses exceeding \$30 million. Confidence, romance, and emergency scams built on cloned voices added more than \$24 million [2]. The United States Financial Crimes Enforcement Network has separately warned financial institutions about a rise in suspicious activity reports in which generative deepfakes were used to get past identity checks. Europol reports describe a comparable spread across Europe, most visibly in corporate impersonation, extortion and the tampering of evidence [3].

The Kazakhstani picture is harder to draw. The source base for this article contains no specialised statistics on how often deepfake-enabled fraud occurs in the Republic of Kazakhstan or how much damage it causes. Neither the Committee on Prosecutorial Statistics and Special Records of the General Prosecutor's Office nor the Ministry of Internal Affairs provides such data in the available materials. This absence says more about how offences are recorded than about the scale of the problem. Generative tools are available everywhere, and the criminal patterns seen abroad are unlikely to stop at national borders. Domestic law enforcement bodies will face them, if they do not already.

The research literature on the subject divides into two largely separate fields. Computer science and digital forensics have produced a large body of work on detection. Early methods looked for spatial and frequency-domain artifacts; later ones rely on Convolutional Neural Networks, Vision Transformers and models that catch mismatches between the audio and video tracks. The results are impressive on paper, but systematic reviews repeatedly point to the same weaknesses, noting severe generalisation limits across unseen datasets [4]. They can be deceived by adversarial manipulation. Their accuracy drops sharply once a file has been compressed, which is exactly what happens to most media shared through messengers and social networks.

Legal scholarship approaches the subject from the other side, and mostly in theoretical terms. Its main concerns are how electronic evidence is authenticated under existing rules, such as Rule 901 of the Federal Rules of Evidence, who bears the burden of proof, how jurors react when manipulation is alleged, and the "liar's dividend", where a genuine recording is dismissed as fake. Kazakhstani legal science has done little to connect the two fields. It has not yet examined systematically how automated detection software, forensic tools and cryptographic provenance mechanisms could enter a pretrial investigation lawfully, or in what form their results could become admissible electronic evidence or part of an expert opinion. Several related issues also remain open: the procedural weight of a specialist's assessment, the certification of AI-based forensic software, and the tactics for investigating fraud committed with generative models. This gap between what the technology can detect and what criminal procedure is prepared to accept is the subject of this article.

Research methods and materials

The study draws on four groups of sources. The first is international cybercrime statistics: the annual reports of the Federal Bureau of Investigation Internet Crime Complaint Center, alerts issued by the Financial Crimes Enforcement Network, and Europol reports. The second is legislation and case law from the Republic of Kazakhstan, the United States, the European Union, the People's Republic of China and the United Kingdom. The third is technical authenticity standards, mainly those of the National Institute of Standards and Technology and

the C2PA specification. The fourth is published results from deepfake detection benchmarks (FaceForensics, the DeepFake Detection Challenge and Celeb-DF).

No single discipline could handle this material alone. The research combines digital forensics, computer science and criminal procedure law. The technical findings are read from a legal standpoint: they matter here to the extent that they affect evidence and its assessment.

Comparative legal and normative analysis was used to set foreign regulation against Kazakhstani law and identify gaps. Forensic and technical analysis covered the detection methods themselves: CNN architectures, Vision Transformers, spectral analysis, biological signal analysis, and XAI tools. The final stage was a critical evaluation. It asked how reliable the algorithms are once they leave benchmark conditions, and whether their output can be admitted in court as electronic evidence.

Findings

3.1 Concept, Types and Unlawful Uses of Deepfakes

Legal and forensic writing has settled on a fairly stable understanding of the deepfake. It is synthetic media produced or altered by specific classes of AI models: generative adversarial networks, autoencoders, diffusion models, and neural vocoders. What sets it apart from older forms of editing is realism. The resulting sound, image or combined recording can be so close to a genuine one that the difference is hard to catch without special tools. Regulators have adopted their own definitions. Under the European Union AI Act, a deepfake is image, audio or video content that has been generated or manipulated by AI and resembles real persons, objects, places, entities or events, so that a reasonable observer would wrongly take it to be authentic or truthful. The National Institute of Standards and Technology takes a broader view. Its frameworks treat as synthetic any content substantially generated or altered by algorithms, and they stress what this means in practice: new difficulties for digital forensics, for information integrity and for verifying electronic evidence.

Technically, deepfakes are not one phenomenon but several. They can be grouped into five modalities.

Face swapping, sometimes called identity swapping, is the best known. An algorithm takes a source image or video and replaces one person's facial features and head geometry with those of another.

Facial reenactment works differently. Under the labels of lip-syncing and motion transfer, the system tracks a source actor's micro-expressions, eye movements and lip movements in real time. It then transfers them onto the target's image or video stream, so that the target's mouth moves in step with a synthesised soundtrack.

Voice cloning relies on text-to-speech synthesis and voice conversion. The model studies the acoustic features of a particular voiceprint and reproduces its pitch, tempo, intonation and prosody closely enough to deceive listeners who know the speaker well.

Full-face synthesis goes a step further. The system does not copy anyone: deep learning models create photorealistic faces, and sometimes whole visual profiles, for people who do not exist.

The fifth group is less spectacular but matters a great deal for fraud. It covers forged identity documents, synthetic biometric credentials, and altered images designed to deceive automated processing systems rather than human reviewers.

Fraud built on these tools relies on two things: the victim knows less than the offender, and the victim can be manipulated psychologically. Business Email Compromise and chief executive officer fraud show both clearly. The offender uses a cloned voice, or a face swapped in real time, to pose as a director, an executive, or a finance officer. An employee then receives an urgent instruction: move company money to an offshore account, or approve a transaction that should never pass. Pressure on employees is only one variant. The same approach works on families. In «relative in distress» or «grandparent» scams, a cloned voice or fabricated proof-of-life footage stages a kidnapping, a road accident or a sudden legal emergency. The victim, frightened and short of time, pays a ransom or sends «emergency» funds.

A second group of schemes needs no personal contact at all. Here the targets are markets and financial institutions. One common technique is fake investment advertising: AI-generated videos and cloned voices of celebrities, officials or well-known financial commentators promote fraudulent crypto platforms and high-yield schemes. More technically demanding are cognitive attacks on banks' own defences. During remote identity verification, the offender presents synthetic facial imagery or a cloned voiceprint and gets past Know-Your-Customer onboarding, liveness detection and Automatic Speaker Verification. The result is a fraudulent account opened in someone else's name, or an existing account taken over, often without triggering the usual forensic screening.

Not every harmful use of deepfakes is about money. Many target a person's image, dignity, and psychological well-being, and more broadly, public trust. The most widespread example is the creation and distribution of non-consensual intimate or sexually explicit deepfakes. A real person's face is placed onto obscene material, which then becomes a tool for blackmail, extortion or sustained harassment. Deepfakes are also used for defamation, corporate sabotage and political discreditation: fabricated statements and staged compromising situations can damage both private individuals and institutions.

The damage goes beyond the immediate victims. Once fabricated audio and video are widely known to exist, suspicion falls on genuine recordings as well. Courts, investigators, and the public begin to doubt electronic evidence as a category. This weakens one of the foundations on which modern evidentiary practice rests.

3.2 Technological Methods for Detecting Deepfakes

AI-based detection is not a single technique but a set of architectures, each looking for a different kind of anomaly. Some examine the picture as a spatial object. Others track changes over time, move the signal into the frequency domain, or look for biological traces that a living person leaves and a generator does not.

Convolutional neural networks were the first workhorse of the field. Standard models such as XceptionNet, ResNet, EfficientNet and 3D-CNNs build hierarchies of spatial features. These let them spot local traces of forgery: the seam where a swapped face was blended in, or colour that shifts slightly from one frame to the next. Their weakness is a limited receptive field, since a convolution sees only a small neighbourhood at a time. Vision Transformers and purpose-built anti-deepfake transformer frameworks address this with self-attention. They treat the image as a sequence of patch embeddings and model how distant regions relate to each other. This makes global inconsistencies visible, the kind a local filter would miss.

A separate line of detectors ignores generic features and looks at how people actually behave on camera. Blinking that is too rare or too regular, micro-expressions that do not fit together, lighting that falls wrongly across the face, jerky or unnatural motion: each of these can reveal

a synthetic video. Biological signal analysis goes further. Remote photoplethysmography extracts faint changes in skin colour caused by the pulse. After a Fast Fourier Transform, these are rendered as matrix heatmaps. Current generative models do not reproduce such cardiovascular variations, which makes their absence a meaningful sign.

Spatial analysis has a counterpart in the frequency domain, where generators leave mathematical fingerprints that are hard to remove. The Discrete Cosine Transform, the Discrete Fourier Transform, and the Discrete Wavelet Transform all reveal what the eye cannot see: excess high-frequency noise, checkerboard patterns left by upsampling, breaks in the spectrum that appear during synthesis.

Audio detection has its own architecture, usually split into two stages: frontend feature extraction and backend classification [5]. The classic handcrafted options are Log Power Spectrum, Linear Filter Bank, Constant-Q Transform, Mel-Frequency Cepstral Coefficients and Linear Frequency Cepstral Coefficients. They combine these with graph attention networks or capsule networks, which can model complex spectro-temporal relationships. Such systems are designed to recognise the traces of voice cloning, text-to-speech synthesis and neural vocoders.

All of the methods above are passive: they inspect a finished file and try to decide whether it is fake. Active methods work the other way round. They attach proof of origin at the moment content is created, so that later tampering becomes detectable.

Digital watermarking embeds visible or machine-readable signals in visual or audio content, allowing automated systems to verify its origin and detect unauthorised modification. In synthetic speech, inaudible acoustic watermarks may also help identify or deter voice cloning.

C2PA and the Content Credentials framework use digital signatures to bind provenance information to a media file, including its creation and editing history. Although platforms may remove metadata, cryptographically signed provenance records can support subsequent authentication and chain-of-custody verification.

Classical digital forensics still carries most of the weight in court, and expert testimony depends on it. Its main methods are the following:

- Metadata analysis looks at Exchangeable Image File Format headers, container structure and the history of compression. Error Level Analysis, for example, shows areas where the compression rate differs from the rest of the image, which often marks an edited region.
- Phonoscopic and forensic audio examination measures acoustic parameters: the spectral envelope, formant distribution and the contour of the fundamental frequency. It compares speakers under realistic conditions, where channel mismatch, babble noise and clipping or saturation of the signal have to be taken into account.
- Portrait and video-technical examination studies sensor noise patterns such as Photo-Response Non-Uniformity. It also looks for differences in optical flow fields that betray motion inconsistent from frame to frame.

Under the silent witness doctrine, examinations of this kind turn electronic material into objective demonstrative proof. That holds only if the chain of custody is unbroken and forensic integrity has been preserved throughout.

Technology alone does not close the gap. Financial institutions, telecommunications networks and law enforcement agencies also rely on organisational and operational safeguards.

In remote Know-Your-Customer onboarding, banks deploy biometric liveness detection with both active and passive challenge-response checks, aimed at identity spoofing and synthetic facial presentation attacks. For large transfers, operational risk rules require multi-factor and

two-channel verification. An instruction received on a video call with an «executive», or in a voice message, is confirmed through a separate, out-of-band channel before any money moves.

A further layer is automated anti-fraud monitoring. It processes transaction streams and behavioural telemetry in real time, flags unusual account activity and blocks transactions that depend on synthetic identity documents or cloned biometric credentials.

3.3 Assessment of Effectiveness and Limitations

Published accuracy figures give a flattering picture. On controlled benchmark datasets, modern detectors look almost flawless. In real forensic work, where the material is messy and nobody knows in advance how it was produced, the numbers fall.

FaceForensics shows the problem clearly. Convolutional Neural Networks, Vision Transformers and frequency-domain models all reach near-perfect classification on it, with Area Under the Curve scores above 97% to 99%. Systematic evaluations suggest that this says more about the dataset than about the detectors. FaceForensics is saturated. Its manipulations were produced under controlled conditions and its compression artifacts are uniform, so it can no longer separate strong models from weak ones, least of all when current generators are involved.

Harder, less constrained collections tell a different story. On the DeepFake Detection Challenge and Celeb-DF datasets, accuracy varies widely and often drops to between 65% and 91%. The reasons are those of ordinary life: uneven lighting, faces at arbitrary angles, low-resolution capture, lossy compression.

Complexity is no guarantee of quality either. As demonstrated in comparative benchmarks [6], a Random Forest classifier trained on handcrafted features reached 99.64% accuracy on the DeepFake Detection Challenge with an inference time of just 2 milliseconds, whereas multimodal deep architectures required up to 60 milliseconds per frame while achieving lower overall accuracy.

The most serious technical obstacle is generalisation across datasets. Across methodological families, studies record an average drop of 10% to 15% once a detector meets data from an unfamiliar distribution or a synthesis method it has not seen. Architecture makes some difference: Vision Transformers lose 11.33% on average, while pure Convolutional Neural Networks lose more than 15%. Even so, the overall decline points to a basic flaw. The classifiers have mostly learned the quirks of their training data, such as its compression signatures, blending seams and local colour-space artifacts. They have not learned what makes media synthetic in general.

This leaves detection permanently behind generation; in an arms race, it is structurally placed to lose. The move from Generative Adversarial Networks to diffusion models and flow-matching architectures has changed the spectral fingerprints that detectors relied on and removed the spatial boundaries they looked for. Tools trained on the older generation of fakes stop working.

Adversarial attacks make matters worse. Gaussian noise too faint to see, hidden backdoor triggers or adaptive adversarial examples can push a synthetic file past the classifier while the fake looks exactly as convincing as before.

Moving a detector from the laboratory into an investigation or a courtroom raises further difficulties, and they are practical rather than algorithmic.

The first is the state of the evidence itself. Audio and video rarely reach an expert in original form. They have usually passed through messaging platforms with lossy compression, and

the signal may be heavily clipped or overlaid with babble noise. Under such conditions, false acceptance and false rejection rates rise sharply.

The second is bias. Synthetic speech detectors produce more false positives for certain age groups, for speakers with non-native accents and for people with speech impairments. In a criminal case, this is not a minor statistical footnote.

The third problem is procedural and probably the hardest to solve. A deep learning detector is an opaque system. It outputs a confidence score, often reduced to a yes-or-no answer, with no reasoning a person can follow. An expert cannot explain to the court why the system reached its conclusion, and the output does not meet judicial standards of authentication.

If such results are ever to count as admissible electronic evidence, forensic practice will need explainable artificial intelligence (XAI). Tools such as SHapley Additive exPlanations (SHAP), attention map visualisations or structural anchors can turn a bare probability into proof that can be demonstrated and checked. This would lower the risk of wrongful accusation and limit the wider liar's dividend in court.

Table 1 – Comparative assessment of deepfake detection methods

Detection Method	Working Principle	Key Advantages	Operational Limitations	Evidentiary Value in Court
Convolutional Neural Networks (CNNs / XceptionNet)	Learns hierarchies of spatial features; finds local colour mismatches and blending seams within frames	Very high accuracy on familiar data (>98%); light computation, fast inference	Loses more than 15% on unfamiliar datasets; sensitive to lossy compression and adversarial noise	Moderate: useful for initial technical screening, but its decisions are hard to explain in expert testimony
Vision Transformers (ViT / ADT)	Uses self-attention to relate distant image regions and capture global context	Better generalisation (average drop of 11.33%); detects global structural inconsistencies	3–5× more parameters and latency; demanding for real-time use on edge devices	Moderate to High: attention maps can show the court which areas were manipulated
Multimodal Audio-Visual Models (e.g., AVoiD-DF)	Encodes audio and video together to find desynchronisation, phoneme-viseme mismatches and prosodic inconsistencies	Accurate against sophisticated fakes; relies on physical limits of how sound and image align	Slow (~60 ms); useless for muted videos, still images or partial audio	High: the mismatch between sound and image can be shown directly to a jury
Classical Machine Learning (Random Forest / SVM)	Classifies using handcrafted acoustic and visual features (MFCCs, LBP, HOG, spatial texture descriptors)	Very fast (2 ms); small memory footprint; competitive accuracy within defined conditions	Features must be engineered by hand; performance collapses against unseen cloning algorithms	High: the feature space is transparent and fits forensic expert standards

Frequency-Domain Analysis (DFT / DCT / Wavelets)	Decomposes the signal spectrally to reveal high-frequency noise, upsampling checkerboard patterns and GAN fingerprints	Robust to spatial blurring; helps attribute content to a model and catch subtle generative traces	Multi-scale transforms are computationally heavy; strong lossy compression (JPEG, MP3) weakens results	Very High: spectral graphs serve as objective demonstrative evidence
Cryptographic Provenance & Watermarking (C2PA / Digital Signatures)	Embeds hidden perturbations at generation, or binds a cryptographically signed asset manifest to the file	Verifies authenticity deterministically; with public key infrastructure or blockchain, the chain of custody cannot be altered	Social media platforms may strip metadata; adversarial re-noising can remove hidden watermarks	Highest: conclusive proof of origin and chain of custody under electronic evidence rules

3.4. Legal Regulation and Procedural Aspects in Kazakhstan

The Criminal Procedure Code of the Republic of Kazakhstan treats as evidence factual data that were obtained lawfully and that establish circumstances relevant to the correct resolution of a criminal case [7]. A digital media file may constitute evidence if it is obtained, preserved, examined and submitted in accordance with criminal procedural requirements. The output of an automated detector may have evidentiary relevance only when it is verified and procedurally documented by an authorised specialist or forensic expert.

Article 125 of the Code requires every piece of evidence to be assessed according to inner conviction, formed through a comprehensive, full and objective examination of all the circumstances of the case taken together. A number produced by software cannot replace that assessment. The Law of the Republic of Kazakhstan «On Forensic Examination Activity» adds a further requirement: the scientific and technical means and methods used in forensic research must be scientifically sound, safe and applied in the established procedure [8].

It follows that a raw confidence score, an automatically generated analytical log, or any unverified output of a standalone detection program cannot enter the case as self-sufficient electronic evidence. The route to admissibility runs through a human expert. The expert has to interpret what the algorithm found, check the features on which that finding rests, and present the result in procedural form, expert conclusion и specialist conclusion.

A further statutory basis has recently appeared. The Law of the Republic of Kazakhstan «On Artificial Intelligence» No. 230-VIII of November 17, 2025 establishes a normative framework for synthetic content and makes its labelling mandatory [9]. For proceedings, this matters because a label (or its absence) gives the investigation a legal point of reference when it needs to establish where a recording came from. The subordinate measures for executing this legislative mandate were formally established by Decree of the Prime Minister of the Republic of Kazakhstan dated January 14, 2026, No. 2-r [10]. As analyzed in legal literature (Oralbayev, 2025), the Law introduces foundational statutory definitions for synthetic results, a risk-based classification model, and mandatory labeling requirements [11].

The Code distinguishes two participants who may bring technical knowledge into a case: the expert (Article 79) and the specialist (Article 80). Their roles differ, and with synthetic media the difference has practical weight.

The specialist works alongside the investigator. He or she helps find, seize and preserve electronic evidence and carries out its preliminary technical examination. The expert acts independently: the expert examines the objects submitted, applies validated methodologies and signs the opinion in his or her own name.

Under Order of the Minister of Justice No. 306, audiovisual material currently falls mainly within two types of examination: forensic portrait examination, and forensic video and sound recording examination [12]. Both were designed for a world in which faces and voices could be compared on the assumption that the recording itself was genuine. Generative adversarial networks, diffusion models and neural vocoders now produce multimodal content realistic enough to undermine that assumption. Traditional facial identification and acoustic comparison run into limits that are both technical and methodological.

The system needs to respond on two fronts. First, the Center of Forensic Examinations of the Ministry of Justice should develop and validate specialised methodologies and enter them in the State Register of Forensic Examination Methodologies, as Law No. 44-VI requires. Second, the official list of forensic expert specialties should include a separate specialty for examining synthetic media and the artifacts of artificial intelligence. Without that, the task will keep falling between existing specialties, none of which fully covers it.

In substantive criminal law, deepfake-enabled offences are mostly qualified as fraud under Article 190 of the Criminal Code of the Republic of Kazakhstan [13]. This classification is correct but incomplete. A deepfake scheme attacks several interests at once. It takes property, it misuses a person's image and identity, and it undermines social trust.

Generative AI also changes how such crimes are organised. The technical threshold for entry falls. Criminal schemes break down into interchangeable modules and spread faster. As demonstrated by Songyang Sai and Zifan Wang (2026) [14], deepfake-enabled fraud exhibits a modularized criminal chain, low entry barriers, and precise targeting methods. Attacks can be aimed precisely at particular citizens, financial institutions or public bodies.

Article 190 does not currently reflect any of this: it has no qualifying feature for offences committed with artificial intelligence technologies or synthetic media. Investigators and courts therefore rely on the general provisions, or on cumulative qualification with offences in the sphere of informatization and communications (Articles 205-213) or with violations of privacy and personal data legislation (Article 147).

Legal scholarship suggests two ways forward. One is to add a qualifying feature to Article 190. The other is to recognise the malicious use of generative AI as an aggravating circumstance under Article 54 of the Criminal Code. Either change would give the law a direct answer to the problem. Sentences would become more proportionate, the heightened social danger of synthetic media fraud would be expressly recognised, and judicial practice would gain the legal certainty it currently lacks.

3.5. Foreign Experience

The United States has no single deepfake statute. Instead, it responds through several layers at once: executive policy, targeted federal legislation, alerts from financial regulators, and gradual adjustment of the rules of evidence.

On the policy side, the key technical document is National Institute of Standards and Technology Report AI 100-4 [15]. It was prepared under Executive Order 14110, which was revoked in January 2025, but the report itself remains a reference point. It puts transparency of digital content first and describes the available technical approaches: digital watermarking, metadata recording, and automated detection of synthetic content.

Legislation has so far addressed specific harms. The TAKE IT DOWN Act (Public Law 119-12) [16] prohibits non-consensual intimate imagery, including synthetic content, and requires its prompt removal. The Federal Communications Commission has also ruled that AI-generated voices in robocalls fall within the restrictions of the Telephone Consumer Protection Act.

Enforcement agencies track the problem closely. The Financial Crimes Enforcement Network [17] and the Federal Bureau of Investigation Internet Crime Complaint Center both monitor deepfake-enabled cybercrime. In 2025, the latter logged more than 22,000 complaints, with adjusted losses above \$893 million.

In federal courts, electronic evidence created or altered using artificial intelligence is authenticated under Rule 901 of the Federal Rules of Evidence [18]. The proposed Rule 901(c) would require a party alleging fabrication to provide a sufficient factual basis before the proponent is required to demonstrate that the evidence is more likely than not authentic.

The draft is designed to block empty allegations of forgery without making genuine challenges impossible. Trial courts are already working through the same issues under existing standards. In *State v. Puloka*, a Washington court held a Frye hearing and excluded AI-enhanced video evidence. In *Huang v. Tesla, Inc.*, a California court refused to accept unsupported claims that recorded public statements might be deepfakes. The two cases address opposite risks: false evidence admitted, and genuine evidence rejected.

Research funding completes the picture. The Defense Advanced Research Projects Agency's Semantic Forensics programme combines multimodal detection with attribution of content to its source and characterisation of the intent behind it.

The European Union has adopted a centralised approach based on Regulation (EU) 2024/1689, the Artificial Intelligence Act [19]. Article 3(60) defines a deepfake as AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and falsely appears authentic or truthful.

The operative rules are found in Article 50, read together with Recital 133. Providers and deployers of generative AI systems must meet transparency obligations, including machine-readable marking of synthetic output through watermarks, provenance metadata or similar techniques. Where a deployer publishes AI-generated text meant to inform the public on matters of public interest, its artificial origin must be disclosed. There are exceptions for law enforcement purposes and for text that has undergone human review or editorial control.

Regulation is one side of the question; operational reality is the other. Reports from the Europol Innovation Lab show that police across the Union face growing exposure, especially to executive impersonation, extortion and possible tampering with electronic evidence in criminal cases.

China approaches synthetic media mainly as a matter for administrative supervision of generative AI and deep synthesis technologies. The foundation was laid by the Provisions on the Administration of Deep Synthesis of Internet-based Information Services (2022). On that basis, the Measures for Labeling Artificial Intelligence-Generated and Synthetic Content introduced dual labelling for service providers and distribution platforms.

The first label is explicit: it must tell the user plainly that the content is synthetic. The second is implicit: it is embedded in the file's metadata so that the content can be traced to its source and audited later. Platforms bear active duties. They must check that labels are present, run automated identification mechanisms, and act against unlabelled synthetic media that threatens public order or national security.

The regulatory model sits alongside a strong research base. Comparative bibliometric studies show a heavy concentration of deepfake detection research in Chinese universities and research centres, much of it built on Convolutional Neural Networks, Vision Transformers and spatial-frequency feature extraction.

The United Kingdom has addressed deepfakes mainly through online safety and criminal law. The Online Safety Act 2023 makes sharing non-consensual intimate deepfakes a criminal offence.

The courts have already felt the problem. In one family court case, a party submitted manipulated voice audio to discredit the other side, and the fabrication came to light only through forensic examination of the recording. The case illustrates how easily such material can enter proceedings and how much depends on expert scrutiny.

Some responses come from outside the law. Public broadcasters now use visual and audio disclosures whenever a programme contains face swapping or voice synthesis.

Beyond national systems, a shared technical infrastructure is emerging through multi-stakeholder initiatives, chiefly the Coalition for Content Provenance and Authenticity (C2PA) and the Content Credentials framework. The C2PA specification is an open standard that works across platforms. It cryptographically binds asset manifests, creation metadata and editing histories to media files. Provenance can then be checked, and the chain of custody stays intact as content moves through international distribution networks.

Discussion

Taken together, the technical and legal findings point to one conclusion: detection technology is moving faster than criminal justice systems can procedurally absorb it.

Computer science research presents a mixed record. Deep learning detectors, especially Vision Transformers and specialised frequency-domain models, perform very well on controlled benchmarks. Once they meet unfamiliar data distributions, compression artifacts or real acoustic noise, they lose a consistent 10% to 15% of their performance. The work of Univaso and San Segundo makes this concrete for voice: leading detectors degrade sharply under babble noise and signal clipping. These are ordinary conditions in real criminal investigations, and forensic voice comparison suffers accordingly. Univaso and San Segundo [20] demonstrate that leading synthetic-speech detectors deteriorate significantly under babble noise and signal clipping, both of which are common in real criminal investigations.

Legal scholarship approaches the problem from the side of harm. Deepfake-enabled fraud does not fit neatly into the model of a traditional property offence, because it damages several interests together: personality rights, digital identity and the credibility of institutions. Songyang Sai and Zifan Wang connect this to broader social change. In their view, deepfakes operate at a distance, spread through many channels and look real, and together these features push society toward a post-truth condition in which people lose basic confidence in what they see and hear.

The liar's dividend, as analysed by Livingston, makes that loss of confidence worse. Because the public knows that generative AI exists, a dishonest party can dismiss authentic electronic evidence as fake, and the claim sounds plausible.

The sharpest disagreement in criminal procedure concerns the status of automated detection. One view holds that its output could stand as electronic evidence in its own right. The other would confine it to auxiliary screening. The technical evidence favours the second position.

Current detectors are opaque. They return a probability, not an explanation. Cao and Gong add a more troubling finding. Adversarial noise and backdoor exploits can be engineered to move a detector's classification boundary while leaving the image visually unchanged. A tool that can be steered this way, and whose reasoning cannot be examined, is a poor candidate for independent proof. Admitting its output on that basis would put core evidentiary standards at risk and invite judicial error.

Foreign law has reached a similar position. In the United States, Rule 901 and the proposed Rule 901(c) require authenticity to be demonstrated rather than presumed. The European Union Artificial Intelligence Act classifies certain AI systems used in the administration of justice as high-risk. Detection output therefore requires expert interpretation, explainable AI and a verified chain of custody. Even then, it should form part of an expert opinion rather than operate as independent evidence.

Forensic voice comparison protocols take the same line. Automated detectors are for preliminary screening, to be followed by full expert evaluation. The probability a machine reports must never take the place of fact-finding by the court.

Fighting AI-driven crime has costs that the law must weigh against it. Automated detection and surveillance tools inevitably affect privacy, the protection of biometric data and the presumption of innocence. Blanket screening of digital communications would create a standing apparatus of surveillance. Some practices have already been ruled out: Regulation (EU) 2024/1689 expressly prohibits untargeted scraping of facial images to build facial recognition databases.

A workable balance could rest on two principles. The first concerns data collection. Informed consent should be the default rule for gathering biometric data, with narrow exceptions for criminal investigations, set out in statute and subject to judicial or prosecutorial oversight.

The second concerns the courtroom. A party who claims that a recording is a deepfake should have to point to specific factual grounds, and the claim should then be tested by expert examination. This would keep unsupported allegations from derailing proceedings. It would also leave the burden of proof where it belongs, rather than shifting it onto the accused.

The gaps identified here are normative, procedural and operational, and they cannot be closed one at a time. What is needed is a coordinated set of legislative and organisational measures.

First, Kazakhstani legislation needs a normative definition of synthetic (deepfake) content, aligned with Article 3(60) of the EU AI Act and with the terminology of the National Institute of Standards and Technology. A definition alone will do little. It should come with statutory duties for providers of generative models and for online platforms: machine-readable provenance metadata, digital watermarks, and clearly visible or audible disclosure labels on AI-generated output.

Second, substantive criminal law should reflect how serious AI-facilitated offences are. One option is to add a qualifying feature to Article 190 of the Criminal Code of the Republic of Kazakhstan covering fraud committed with artificial intelligence technologies or synthetic media. The other is to add a corresponding aggravating circumstance to Article 54. Either way, the law would recognise what distinguishes these attacks: more convincing deception, more precise targeting, and greater financial and social harm.

Third, the Center of Forensic Examinations of the Ministry of Justice of the Republic of Kazakhstan should develop standard methodologies for authenticating synthetic media. They should set out how spatial, temporal, frequency and acoustic artifacts are to be evaluated. They should also introduce mandatory certification and explainability requirements for any AI detection software used in pretrial investigations.

Fourth, banks, payment system operators and telecommunications providers should be legally required to protect against deepfakes. The minimum should include multimodal liveness detection in remote Know-Your-Customer onboarding, out-of-band confirmation of transactions through a second channel, and real-time anti-fraud monitoring against biometric spoofing and voice cloning.

Fifth, law enforcement bodies should build a centralised interagency reference database of deepfake samples, generative model fingerprints and synthetic voiceprints to support forensic investigations. Infrastructure without people who can use it is of limited value, so the database should be paired with specialised training for investigators, prosecutors and judges. Better technical literacy would help courts examine electronic evidence with proper rigour.

Conclusion

Generative artificial intelligence has become powerful and widely accessible faster than criminal justice systems have adapted their procedures. Deepfakes have moved from the media industry into the toolkit of cybercriminals, where they serve financial fraud and the circumvention of biometric protection. The damage extends beyond individual victims. Confidence in electronic evidence as a whole is being eroded.

The technical and legal analysis supports three conclusions.

1. Passive detection has clear limits. Neural models (CNNs, Vision Transformers) and spectral methods perform very well on benchmark datasets. Under real investigative conditions, their performance falls by 10-15% or more. Messenger compression, channel noise and adversarial attacks all contribute, and no current architecture is immune to them;

2. Opaque algorithms cannot serve as evidence on their own. A detector reports a confidence score without any explanation a person can follow. Its logs and outputs therefore cannot be admitted as self-sufficient evidence. They can enter a case only through a forensic expert, supported by explainable AI (XAI) and an unbroken chain of custody;

3. Legislative and institutional reform has to be systemic. The Kazakhstani framework, read against foreign practice, shows that isolated amendments will not be enough. The following measures are needed together:

- a statutory definition of synthetic content, with mandatory labelling by providers and platforms;

- recognition of generative AI use either as a qualifying feature of fraud under Article 190 of the Criminal Code of the Republic of Kazakhstan or as an aggravating circumstance under Article 54;

- specialised methodologies developed by the Center of Forensic Examinations of the Ministry of Justice, and a separate forensic specialty for AI-generated material;

- a legal obligation for financial institutions to apply secure multi-factor KYC verification and anti-fraud monitoring, together with a centralised interagency database of deepfake samples.

Detection after the fact will always lag behind generation. It needs a counterpart at the moment content is created, in the form of end-to-end cryptographic provenance standards such as C2PA and digital watermarking [C2PA Technical Specifications (Version 2.4) / Coalition for Content Provenance and Authenticity. – 2024. – [Electronic resource] – URL: <https://spec.c2pa.org/> (Accessed: 15.05.2026)]. It also needs criminal procedure law that is adapted to evidence which can no longer be trusted on sight.

Author's contribution

K.M. Beaver conceived the study, conducted the research, and prepared the manuscript.

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26103625 «Online fraud using deepfake-technologies and social engineering: problems of criminal law counteraction, prospects for legislative regulation»).

References

1. Sandoval, V. et al. (2024) 'Threats deepfakes pose to the criminal justice system: A systematic review', *Crime Science*, 13, Article 41. doi: <https://doi.org/10.1186/s40163-024-00239-1>.
2. Federal Bureau of Investigation, Internet Crime Complaint Center (2026) FBI IC3 2025 Annual Report. Washington, D.C.: Federal Bureau of Investigation. Available at: https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf (Accessed: 10 June 2026).
3. Europol Innovation Lab (2022) Facing Reality? Law Enforcement and the Challenge of Deepfakes. Updated 2024. The Hague: European Union Agency for Law Enforcement Cooperation. doi: <https://doi.org/10.2813/158794>.
4. Heidari, A., Jafari Navimipour, N., Dag, H. and Unal, M. (2024) 'Deepfake detection using deep learning methods: A systematic and comprehensive review', *WIREs Data Mining and Knowledge Discovery*, 14(2), Article e1520. doi: <https://doi.org/10.1002/widm.1520>.
5. Zhang, B., Cui, H., Nguyen, V. and Whitty, M. (2025) 'Audio deepfake detection: What has been achieved and what lies ahead', *Sensors*, 25(7), Article 1989. doi: <https://doi.org/10.3390/s25071989>.
6. Raza, A., Basit, A., Amin, A., Arfeen, Z.A., Masud, M.I., Fayyaz, U. and Jumani, T.A. (2026) 'A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures', *AI*, 7(2), Article 68. doi: <https://doi.org/10.3390/ai7020068>.
7. Republic of Kazakhstan (2014a) Criminal Procedure Code of the Republic of Kazakhstan dated 4 July 2014 No. 231-V, with amendments and additions as of 12 August 2026. Available at: <https://prg.kz/> (Accessed: 15 September 2026).
8. Republic of Kazakhstan (2017a) Law of the Republic of Kazakhstan dated 10 February 2017 No. 44-VI "On Forensic Examination Activity", with amendments and additions as of 1 July 2026. Available at: <https://prg.kz/> (Accessed: 15 September 2026).
9. Republic of Kazakhstan (2025) Law of the Republic of Kazakhstan dated 17 November 2025 No. 230-VIII ZRK "On Artificial Intelligence". Available at: <https://prg.kz/> (Accessed: 15 September 2026).
10. Prime Minister of the Republic of Kazakhstan (2026) On Measures for the Implementation of the Laws of the Republic of Kazakhstan dated 17 November 2025 "On Artificial Intelligence" and "On Amendments and Additions to Certain Legislative Acts of the Republic of Kazakhstan on Artificial Intelligence and Digitalization": Decree No. 2-r dated 14 January 2026. Available at: <https://prg.kz/> (Accessed: 20 June 2026).
11. Oralbayev, Ch. (2025) 'Key aspects of the Law of the Republic of Kazakhstan dated 17 November 2025 No. 230-VIII ZRK "On Artificial Intelligence"', PARAGRAPH Information System. Available at: <https://prg.kz/> (Accessed: 18 June 2026).
12. Ministry of Justice of the Republic of Kazakhstan (2017) On Approval of the List of Types of Forensic Examinations Conducted by Forensic Examination Bodies and Expert Specialties for Which Qualifications Are Awarded by the Ministry of Justice of the Republic of Kazakhstan: Order of the

Minister of Justice of the Republic of Kazakhstan No. 306 dated 27 March 2017, with amendments as of 3 July 2026. Available at: <https://prg.kz/> (Accessed: 15 September 2026).

13. Republic of Kazakhstan (2014b) Criminal Code of the Republic of Kazakhstan dated 3 July 2014 No. 226-V, with amendments and additions as of 20 August 2026. Available at: <https://prg.kz/> (Accessed: 15 September 2026).

14. Sai, S. and Wang, Z. (2026) 'Criminal regulatory approaches to deepfake-related offenses: Focusing on the crime of fraud', International Journal of Asian Social Science Research, 3(1), pp. 20–31.

15. National Institute of Standards and Technology (2024) Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4. Gaithersburg, MD: U.S. Department of Commerce. doi: <https://doi.org/10.6028/NIST.AI.100-4>.

16. United States Congress (2025) TAKE IT DOWN Act (Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act). Public Law 119-12, 119th Congress, 19 May 2025, 139 Stat. 55.

17. Financial Crimes Enforcement Network (2024) FinCEN Issues Alert on Fraud Schemes Involving Deepfake Media Targeting Financial Institutions. Available at: <https://www.fincen.gov/> (Accessed: 22 May 2026).

18. Legal Information Institute (n.d.) 'Rule 901: Authenticating or identifying evidence', Federal Rules of Evidence. Cornell Law School. Available at: https://www.law.cornell.edu/rules/fre/rule_901 (Accessed: 2 June 2026).

19. European Parliament and Council of the European Union (2024) 'Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)', Official Journal of the European Union, L 2024/1689. Available at: <http://data.europa.eu/eli/reg/2024/1689/oj> (Accessed: 5 June 2026).

20. Univaso, P. and San Segundo, E. (2025) 'Detection of cloned voices in realistic forensic voice comparison scenarios', Frontiers in Artificial Intelligence, 8, Article 1678043. doi: <https://doi.org/10.3389/frai.2025.1678043>.

Information about the author

Beaver K.M. – corresponding author, PhD, Judith Rich Harris Professor of Criminology, College of Criminology and Criminal Justice, Florida State University, 222 S. Copeland Street, Tallahassee, FL 32306, USA

Бивер К.М. – хат-хабар авторы, PhD, Флорида мемлекеттік университетінің Криминология және қылмыстық сот төрелігі колледжіндегі Джудит Рич Харрис атындағы криминология профессоры, 222 S. Copeland Street, Таллахасси, FL 32306, АҚШ

Бивер К.М. – автор для корреспонденции, PhD, профессор криминологии имени Джудит Рич Харрис, Колледж криминологии и уголовного правосудия Университета штата Флорида, 222 S. Copeland Street, Таллахасси, FL 32306, США



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).