EVIWRITE FORMAL REPORT

Digital Identity and Deepfake Harm Evidence Monitor

Why deepfake harm is now a proof-chain problem: removal notices, voice identity, platform response records, and victim-safe evidence preservation

An EviWrite proof-landscape report on deepfake abuse, voice cloning, impersonation and likeness harms, focused on the Removal–Proof Paradox, victim-safe preservation, platform response evidence and the records needed before harmful content disappears or spreads.

REPORT NUMBER	DIDH-2026-01
VERSION	1.1
PUBLICATION DATE	2026-05-27
STATUS	Draft / publication-ready source
DOCUMENT CLASS	Public evidential trend report
REFERENCE	DIDH-2026-01

REPORT POSITION

Proof landscape, not threat landscape.

EviWrite reports identify where public digital claims become evidentially weak, what stronger records would have required, and how similar claims should be evidenced before pressure arrives.

Document control

Source file	digital-identity-deepfake-harm-evidence-monitor-2026.md
Status	Draft / publication-ready source
Version	1.1
Period	Public signals reviewed to 27 May 2026 (2024-01-01 to 2026-05-27)
Prepared by	EviWrite
Report hash	not-issued
PDF hash	not-issued
Receipt	not-issued

Exclusions

- Definitive global incidence of deepfake abuse.
- Forensic determination of any individual media item.
- Legal advice or liability findings.
- Private platform logs not publicly available.
- Political persuasion analysis or party/candidate assessment.
- Operational instructions for creating deepfakes or evading detection.

Proof limits

- That any named party committed wrongdoing.
- That any individual case would fail in court or before a regulator.
- That private evidence does not exist.
- That a cited source is complete, final, uncontested, or legally determinative.
- That EviWrite has independently audited the underlying private systems, logs, files, datasets, or forensic records.
- That selected public signal counts represent global incidence.
- That forecast scores represent probability, incidence, legal outcome, or market size.
- That selected international breakdowns represent global prevalence.

Digital Identity and Deepfake Harm Evidence Monitor

Executive summary

Deepfake harm is no longer mainly a question of whether content looks fake. It is becoming a question of whether the harmed person can prove the right facts quickly enough: what was created, who was depicted, whether consent existed, where it appeared, when notice was given, what copies existed, what the platform did, and what evidence survives after removal.

This report's central finding is the Removal–Proof Paradox: victims need harmful material removed quickly, but later redress often depends on proof that can survive the removal. The evidence process therefore has to do two things at once: minimise harm and preserve enough controlled proof to support takedown, recurrence tracking, safeguarding, legal advice, platform escalation, employer/school response, account recovery, financial dispute or regulatory complaint.

This is not a threat-landscape report, legal update, platform compliance audit, cyber telemetry report or global incidence study. It is a proof-landscape report: it examines where public deepfake, voice-cloning, impersonation and likeness claims become evidentially weak, and what records would make those claims more defensible.

The selected public signals reviewed here show five pressure points: victim takedown packets, duplicate-removal evidence, voice/channel authentication, detector reproducibility, and consent/likeness lineage. The hard lesson is precise: deepfake response fails when proof is assembled only after panic begins.

Source basis for this report

This report combines official regulatory and government sources, technical standards, public safety alerts, academic benchmarks and bounded public-signal journalism. Repeatable sources provide continuity; signal-specific sources ground the individual evidence-pressure points. EviWrite classification is applied to identify proof weakness. Counts in this report are not global incidence.

This report deliberately weights official regulator, government and public-safety material because the evidential pressure here is procedural: notice, removal, safeguarding, reporting, platform response and institutional handling. Journalism is used mainly to identify visible public controversy or cross-border pressure, not to establish legal findings or private platform facts.

Source group	Used for	Not used for
FTC, Ofcom, GOV.UK, European Commission, eSafety, FBI and ICO material	Duties, warnings, public reporting pathways, technical and regulatory direction	Proving any named party's liability or private evidence absence
Legal texts and official explanatory notes	Statutory context and evidence-pressure timing	Legal advice or complete jurisdictional analysis
NIST and government technical reports	Provenance, detection, watermarking and evidence-boundary analysis	Certifying any detector or proving media authenticity
Academic benchmarks and legal/technical analysis	Practical limits, classification pressure and emerging architecture gaps	Binding legal interpretation or forensic conclusion
Journalism and public reporting	Public signal discovery and case-study context	Final findings, complete facts, or court-quality proof

Source boundary: Public sources can show visible evidence pressure. They cannot show all private records, all cases, all victims, all platform actions, or all enforcement outcomes.

The EviWrite evidencing lens

Every signal in this report is assessed through six questions:

Lens	Question
Source	Where did the record, content, dataset, action, or decision come from?
Timing	When did the relevant event, creation, access, disclosure, decision, or knowledge occur?
Control	Who controlled the file, system, account, platform, dataset, log, or workflow?
Sequence	What happened before, during, and after the claim being made?
Verification	Can the claim be checked without simply trusting the claimant or the platform?
Limits	What does the record not prove?

The EviWrite Evidence Failure Stack

Layer	Failure question	Common weakness
Event	Did the thing happen?	The abuse, impersonation, generation or notice is asserted but not reconstructed.
Record	Was it recorded at the time?	The first useful record is created after content has spread or money has moved.
Context	Does the record explain meaning, authority and limits?	Screenshot, label, detector score or platform message exists without enough context.
Custody	Can control and movement be shown?	The strongest record sits with the platform, model provider, account holder or scam channel.
Trust Boundary	Did the record leave the system or party that may later be questioned?	Evidence remains inside a platform or provider environment the victim cannot independently verify.
Verification	Can the claim be checked without belief?	Public claim depends on private logs, a familiar voice, a bare detector screenshot or unsupported assertion.
Permanence	Will the proof survive time and challenge?	URLs, labels, metadata, dashboard records and platform tickets may disappear or change.

The Removal–Proof Paradox

Deepfake harm creates a direct evidential conflict: the victim needs harmful material removed quickly, but later redress often depends on proof that the material existed, where it appeared, when notice was given, what duplicates were known, and how the platform or institution responded.

That is the Removal–Proof Paradox. If content is left online, harm spreads. If content is removed without a controlled evidence packet, later proof can collapse into memory, screenshots, private platform tickets and unsupported assertion.

The strongest evidence process is not the one that collects the most material. It is the one that preserves enough proof while minimising further exposure, repeated viewing, unnecessary copying and uncontrolled redistribution.

Pressure	Risk if handled badly	Evidence-ready posture
Remove harmful material quickly	The only public proof disappears before notice, duplicate scope and response timing are evidenced.	Generate a minimal, controlled, timestamped notice-and-preservation record before or during removal.
Avoid forcing victims to re-view or re-send abuse	The evidence process becomes another exposure event.	Use role-restricted capture, minimal viewing, secure storage, evidence summaries and access logs.
Prove recurrence	One removed URL is mistaken for resolution while copies continue.	Record known identical copies, remix indicators, follow-up ticket IDs and outcome timestamps.
Support later redress	Informal screenshots cannot carry platform, school, employer, legal or safeguarding claims.	Separate content proof, identity proof, consent proof, notice proof and response proof.

Evidence basis: FTC TAKE IT DOWN enforcement guidance, FTC image-based abuse guidance, Ofcom illegal intimate-image material and Australia eSafety deepfake school/image-based abuse guidance.

Source boundary: These sources support the public evidence-pressure pattern around removal, notice, safety and platform response. They do not prove individual liability, content authenticity, unlawfulness or private evidence completeness.

The Deepfake Harm Evidence Chain

Deepfake harm is not a single-file problem. It is a chain problem. A useful response has to preserve enough evidence at each stage without turning preservation into a second act of distribution.

Stage	Evidence question	Minimum record
Capture	What exactly was seen, heard, posted, sent or received?	Content identifier, URL/account/channel, timestamp and controlled capture where lawful.
Preservation	Was enough proof preserved without increasing victim exposure?	Hash or capture log, access restriction, viewing minimisation, secure storage and retention decision.
Source	Where did it appear and under whose account, system or platform control?	Platform, account, message ID, source URL, profile identifiers and reporting route.
Identity	Whose likeness, voice, name, profile or authority was invoked?	Affected identity statement, comparison basis, role/authority context and impersonation indicators.
Consent	What record shows consent was absent, revoked, exceeded or impossible?	Consent-denial statement, prior rights/permission record and age/safeguarding status where relevant.
Notice	What was reported, to whom, when and under which policy or law?	Platform notice, regulator/police/school/employer report, ticket ID and submitted evidence summary.
Response	What did the platform, school, employer, bank or authority do?	Acknowledgement, action timestamp, escalation note, removal/denial reason and appeal trail.
Recurrence	Were duplicates, reposts, remixes or re-uploads tracked?	Duplicate log, known-identical/remix identifiers, platform follow-up reports and recurrence map.
Verification	Can a third party test the claim without trusting only the victim, platform or accused party?	Exportable evidence packet, source map, claim limits, hashes, timestamps and independent verification route.

Evidence basis: NIST synthetic-content risk material, FTC takedown guidance, FBI AI fraud warnings, eSafety school/image-based abuse guidance and EU AI Act transparency material.

Source boundary: The chain is an EviWrite evidential framework derived from cited public sources. It is not a legal test, forensic protocol, platform compliance standard or substitute for safeguarding advice.

False confidence patterns

False confidence	Why it is weak	Stronger posture
“We have screenshots.”	Screenshots detach content from source, timing, custody and spread.	Hash-linked capture, source URL/account, timestamp, preservation log and platform notice record.
“The platform has logs.”	The victim may not control, export or verify the logs.	Victim-facing case export, platform acknowledgement, duplicate-search record and escalation trail.
“The detector says it is fake.”	Detector output may not generalise and may lack reproducibility.	Original file, detector version, settings, confidence, analyst notes and limits.
“The content is labelled AI-generated.”	A label does not prove consent, legality, ownership or source identity.	Label plus generation record, consent/licence evidence, edit history and distribution context.
“The voice sounded real.”	Voice recognition is not identity verification.	Safe phrase, verified callback, channel-bound approval and call/payment metadata.
“We removed it, so the problem is solved.”	Removal can erase the public artefact before notice, duplicates, timing and response are evidenced.	Removal plus controlled preservation, ticket ID, duplicate log, response timestamp and claim boundary.

Evidence basis: NIST synthetic-content material, UK deepfake detection material, FBI AI fraud warnings, FTC voice-cloning material, EU AI Act transparency material and platform takedown guidance.

Source boundary: These sources support evidence limits and process needs. They do not prove a particular item is fake, lawful, unlawful, authorised or unauthorised.

Five findings

1. The Removal–Proof Paradox is now the centre of victim evidence.

The actionable record is no longer just “this image exists”. It is the structured packet: content location, content identity, notice time, platform receipt, duplicate-removal scope, platform decision and escalation trail.

Evidence basis: FTC TAKE IT DOWN enforcement guidance; FTC image-based abuse guidance; Ofcom accelerated illegal intimate-image measures; Australia eSafety guidance.

Source boundary: These sources support public platform-duty and victim-reporting requirements. They do not prove any specific platform failed, any item was unlawful, or any victim’s private evidence is complete.

2. Deepfake detection is not evidence unless the detector itself is evidenced.

A detector result without the preserved media, tool version, settings, confidence, transformation history and analyst limit statement is a weak record. It may be a lead. It is not enough as a defensible proof object.

Evidence basis: NIST synthetic-content report; UK deepfake detection technology report; Deepfake-Eval-2024 benchmark.

Source boundary: These sources support technical limits and benchmark pressure. They do not prove that every detector is unreliable or that any specific item is authentic or synthetic.

3. Voice cloning turns identity into a live-channel authentication problem.

The real weakness is not just that a voice can be cloned. It is that people, banks, public officials, employers and families still treat familiar sound and urgent authority as proof.

Evidence basis: FBI senior-official impersonation alert; FBI IC3 2025 Internet Crime Report; FTC Voice Cloning Challenge; European Parliament scam-calls briefing.

Source boundary: These sources support public warnings and complaint/reporting context. They do not prove technical method in every reported scam or global incidence of voice cloning.

4. A label can warn; it cannot prove consent, legality or ownership.

AI labels and machine-readable markers can support provenance. They do not establish whether the person consented, whether likeness use was authorised, whether the work was lawful, or whether the record is complete.

Evidence basis: European Commission AI Act material; Article 50 reference material; EU AI Act deepfake definition analysis; Article 50 transparency architecture research.

Source boundary: These sources support regulatory and technical pressure around transparency. They do not provide final case law on Article 50 application or prove compliance by any provider.

5. Schools, employers and youth institutions are becoming first-line evidence custodians.

Many deepfake and impersonation harms reach schools, employers, HR teams, safeguarding leads, banks or IT teams before courts or regulators. If those teams delete, forward casually, or treat the event as discipline only, the proof chain can collapse.

Evidence basis: Australia eSafety schools warning; Australia eSafety image-based abuse guidance; FBI generative AI fraud alert; FBI senior-official impersonation alert.

Source boundary: These sources support public signals around school, victim-support and organisational impersonation pressure. They do not measure all institutional incidents.

Evidence-signal scorecard

The scorecard is derived from the selected evidence-signal register in this report.

Metric	Count	Meaning
Selected public signals classified	24	Public signals reviewed for evidential relevance.
Signal sectors represented	15	Distinct sector labels across the selected signal register.
Claim categories represented	5	Claim categories represented by selected public signals.
Region groups represented	5	Region groups represented in the selected signal register.
Source families used	5	Distinct source-family labels in the source register.

Dataset basis: Derived from the selected evidence signals classified in this report.

Chart boundary: Charts and counts show EviWrite classification of selected public signals. They are not measures of global incidence, platform prevalence, legal findings, market size or telemetry.

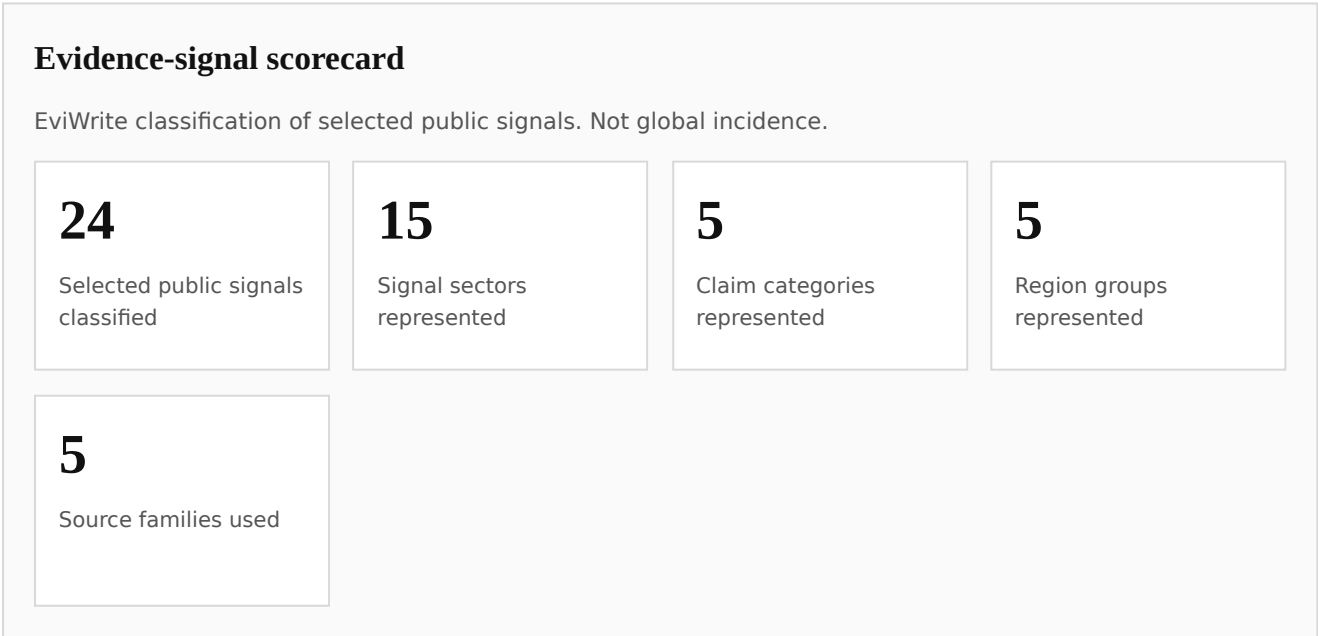
Visual chart summary

The charts declared in this report render selected public evidence signals classified by EviWrite. They show distribution by primary evidence weakness, sector, claim category, region and qualitative forward evidence-pressure trajectory.

Dataset basis: Derived from the selected evidence signals and forecast signals in the YAML register.

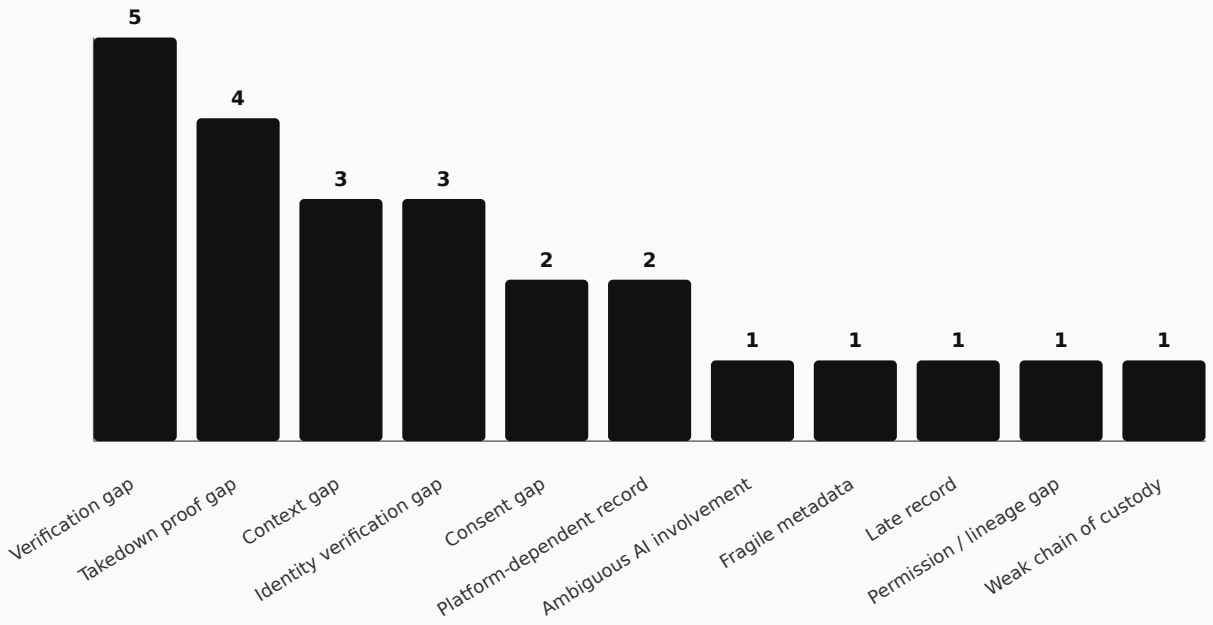
Chart boundary: These charts are not global incidence, market sizing, telemetry, legal findings or probability forecasts.

Visual chart summary



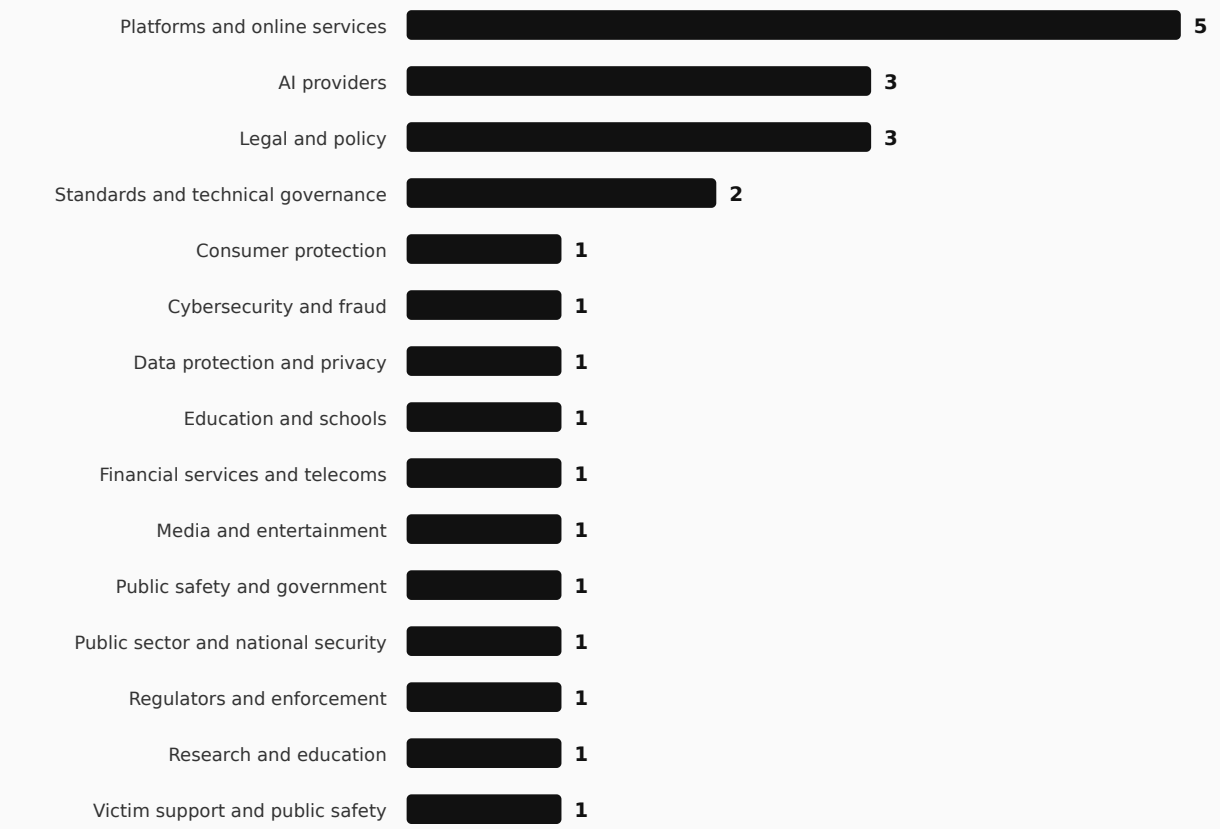
Selected signals by primary evidence weakness

EviWrite classification of selected public signals. Not global incidence.



Selected signals by sector

EviWrite classification of selected public signals. Not global incidence.



Selected signals by claim category

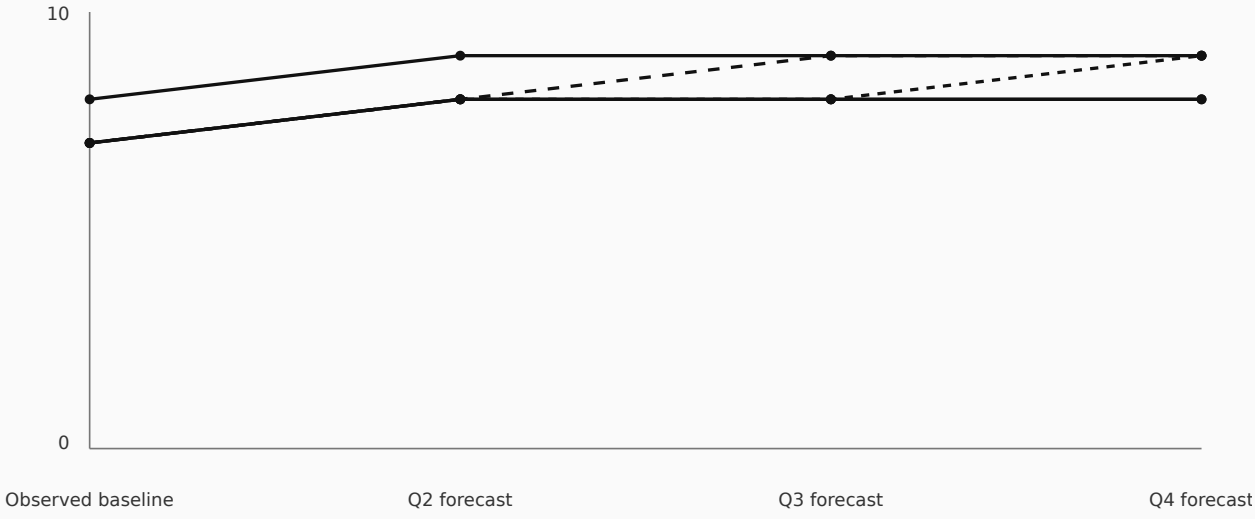
EviWrite classification of selected public signals by claim category. Not global incidence.

Selected signals by region

EviWrite classification of selected public signals by region. Not global incidence.

Forward evidence-pressure trajectories

Qualitative EviWrite forecast score. Not measured incidence, probability, market sizing, or legal prediction.



Evidence failure types by primary weakness

The selected signal set is led by verification gaps, takedown proof gaps, context gaps and identity verification gaps. That pattern matters. It says the practical failure is rarely “nobody knows deepfakes exist”. The failure is that the record cannot be made actionable fast enough.

Primary evidence weakness	Selected signals
Verification gap	5
Takedown proof gap	4
Context gap	3
Identity verification gap	3
Platform-dependent record	2
Consent gap	2
Ambiguous AI involvement	1
Fragile metadata	1
Late record	1
Weak chain of custody	1
Permission / lineage gap	1

Dataset basis: Derived from `evidenceSignals[].primaryFailureType`.

Chart boundary: Primary failure type counts are selected public-signal classifications. Secondary weaknesses are not counted in this table.

Claim-category breakdown

Identity, reputation and attribution harms dominate the selected register. Synthetic-media provenance, platform dependency, cyber incident evidence and regulatory-operational records are the supporting evidence layers.

Claim category	Selected signals
Identity, reputation, and attribution harms	10
Synthetic media and provenance	6
Platform and third-party record dependency	4
Regulatory and operational records	2
Cyber incident evidence	2

Dataset basis: Derived from `evidenceSignals[].claimCategory`.

Source boundary: Claim categories are EviWrite classifications. They do not represent global claim frequency.

Sector pressure ranking

Platforms appear most often because victims, regulators and law now demand action through platform-controlled systems. The next pressure layer is legal-policy definition, AI-provider logging and technical verification.

Sector	Selected signals
Platforms and online services	5
Legal and policy	3
AI providers	3
Standards and technical governance	2
Public safety and government	1
Regulators and enforcement	1
Cybersecurity and fraud	1
Public sector and national security	1
Research and education	1
Data protection and privacy	1
Education and schools	1
Victim support and public safety	1
Consumer protection	1
Media and entertainment	1
Financial services and telecoms	1

Dataset basis: Derived from `evidenceSignals[].sector`.

Source boundary: Sector counts reflect this report’s selected public signals only.

Regional selected-signal breakdown

Region	Selected signals
North America	7
United Kingdom	6
European Union	6
Asia-Pacific	3
International / cross-border	2

Dataset basis: Derived from `evidenceSignals[].region`.

Chart boundary: Regional counts are not prevalence. They show where this report found usable public evidence-pressure records.

Forecast / forward pressure signals

The next evidence-pressure zone is not abstract “AI safety”. It is operational: takedown packets, duplicate search, voice authentication, labelling architecture, school/workplace preservation and detector evidence.

Forward pressure signal	Direction	Q4 2026 qualitative pressure	Evidence question
Victim takedown proof and duplicate-removal evidence	increase	9 / 10	Can the victim prove what was reported, when it was reported, what copies existed, and how the platform responded?
Voice and likeness authorisation disputes	increase	9 / 10	Can the organisation prove whether a voice, face, persona, or likeness replication was authorised, limited, revoked, or misused?
Deepfake detector evidence challenges	increase	8 / 10	Can a detector conclusion be reproduced, bounded, and connected to the preserved media file?
AI Act / DSA reporting and label architecture evidence	increase	9 / 10	Can AI-generated-content labels and user notices be checked against system records rather than trusted as surface claims?
School and workplace deepfake incident preservation	increase	8 / 10	Can the institution preserve enough evidence for takedown, discipline, safeguarding, police report and victim support without amplifying the content?

Source basis: Derived from FTC, Ofcom, GOV.UK, European Commission, FBI, NIST, ICO, eSafety and specialist research sources cited in this report.

Forecast boundary: Scores are qualitative EviWrite evidence-pressure scores. They are not incidence predictions, probability forecasts, legal predictions or market sizing.

Case studies

Case study 1 — TAKE IT DOWN Act: the 48-hour removal window turns victim reporting into auditable evidence

The US removal regime creates a concrete procedural record: what the victim reported, when the platform received a valid request, whether identical copies were removed and whether the platform offered a workable process.

Source basis: FTC TAKE IT DOWN enforcement guidance; FTC Image-Based Abuse guidance; FTC stakeholder letter template.

Evidence weakness: The victim can be harmed twice if evidence preservation requires repeated viewing and repeated reporting of copies.

What stronger evidence would have required:

Weak record	Stronger evidence
Screenshot of abusive image only	URL, file hash or perceptual hash, capture timestamp, platform notice receipt, duplicate-removal status and escalation record
Victim says platform ignored the report	Submitted notice, platform acknowledgement, 48-hour deadline calculation, status messages, removal/refusal record and FTC report if applicable

Evidence boundary: This case study describes public statutory and regulatory process signals. It does not decide whether any specific image is illegal, whether any platform violated the Act, or whether private records exist.

Case study 2 — Grok/X explicit image controversy: provider, platform and complaint evidence collapse into one chain

Public reporting around X/Grok generated explicit-image allegations and regulatory scrutiny illustrates an evidence problem: prompts, safety filters, generated media, reports, moderation actions and policy changes may sit inside a provider/platform environment questioned by victims and regulators.

Source basis: Ofcom investigation notice; Reuters and Guardian public reporting; Guardian Australia reporting; UK official policy context.

Evidence weakness: The public can see the controversy, but the most important evidence may be internal: prompt logs, generation decisions, moderation records, report handling and safety changes.

What stronger evidence would have required:

Weak record	Stronger evidence
Public posts and media reports	Prompt/output records, model/tool version, filter decision record, generated-file hashes, report tickets, moderation outcome and regulator correspondence
Provider says safeguards changed	Versioned policy change, safety configuration record, test evidence, date/time of deployment and post-change incident monitoring

Evidence boundary: The public sources identify scrutiny and investigation signals. They do not prove final regulatory findings, liability, intent or internal platform facts.

Case study 3 — School deepfake abuse: safeguarding, discipline and evidence preservation collide

Australia eSafety signals show schools becoming front-line handlers of AI-enabled image abuse. The evidence risk is that schools treat the matter as conduct or discipline before preserving content, report pathways, consent boundaries and victim-support records.

Source basis: Australia eSafety schools warning and image-based abuse guidance.

Evidence weakness: Deletion-only instincts can help reduce spread but may destroy evidence needed for takedown, police referral, safeguarding, discipline or platform escalation.

What stronger evidence would have required:

Weak record	Stronger evidence
Student or parent shows phone screenshot to school	Trauma-aware intake record, URL/source account, image hash, timestamp, limited-access preservation copy, parent/victim consent controls and referral log
School discipline note only	Separate safeguarding chronology, digital evidence inventory, platform reports, police/regulator referrals and support actions

Evidence boundary: The sources support public pressure around schools and deepfake incidents. They do not measure all school incidents or determine the facts of any individual school case.

Case study 4 — Voice impersonation of officials: the proof problem is channel control, not only audio authenticity

FBI warnings on senior official impersonation show a practical failure mode: recipients trust authority cues through text and voice before verifying the channel, identity and subsequent account access requests.

Source basis: FBI senior-official impersonation alert; FBI generative AI fraud alert; FBI IC3 2025 report.

Evidence weakness: A cloned or alleged cloned voice may not be preserved; the actionable evidence can be call/message metadata, account redirects, credential attempts, payment requests, callback failures and security notifications.

What stronger evidence would have required:

Weak record	Stronger evidence
Recipient says the voice sounded like an official	Message headers, phone metadata, platform account identifiers, audio if lawfully captured, callback attempt log, credential phishing URL and incident report
Organisation says staff were warned	Training record, verification protocol, channel-bound approval logs, exception approvals and post-incident review

Evidence boundary: The sources support warning and threat pattern. They do not attribute any specific incident or prove AI use in every impersonation attempt.

Sector, claim-category, and failure-type table

Sector	Claim category	Visible pressure	Primary evidence weakness	Stronger record
Platforms and online services	Identity, reputation, and attribution harms	Takedown duties, duplicate removal, notice quality	Takedown proof gap	Structured notice, content hash, platform acknowledgement, duplicate-search and outcome log
AI providers	Synthetic media and provenance	Labels, generation controls, safety filters, prompt/output capture	Platform-dependent record	Generation record, model/tool version, label decision, safety-filter decision, generated-output hash
Financial services and telecoms	Cyber incident evidence	Voice cloning, scam calls, urgent payment pressure	Identity verification gap	Channel-bound approval, callback log, payment hold, verified-contact record
Education and schools	Identity, reputation, and attribution harms	Peer-generated synthetic intimate abuse	Late record	Trauma-aware intake, preservation copy, URL/hash, referral log and support chronology
Legal and policy	Identity, reputation, and attribution harms	Creation/request offences, likeness rights and consent disputes	Consent gap	Consent/licence record, request/prompt evidence, identity-source record and revocation trail

Minimum Evidence Records

A Minimum Evidence Record is not a guarantee of legal success, platform removal, compensation, admissibility or regulatory action. It is the smallest practical record set needed to make a claim more defensible, portable, interpretable and independently checkable.

For deepfake harm, the evidential unit is not the image, clip or voice note. It is the controlled response packet around it.

More evidence is not automatically better. The record should be sufficient, controlled, minimised and usable. Evidence collection should not force victims to repeatedly view, forward or re-upload the abusive material.

Area	Minimum Evidence Record	Why it matters
Victim takedown request	Content URL/location, capture timestamp, file or perceptual hash, notice submission, platform acknowledgement, duplicate-removal request, outcome record	Proves what was reported, when, and whether the platform acted within the required process.
Deepfake intimate-image abuse	Consent-denial statement, limited-access preservation copy, source account/URL, spread map, platform reports, support/safeguarding record	Supports removal, safety planning, police/regulator referral and avoids repeated exposure.
Removal–Proof Paradox packet	Minimal preserved evidence, platform notice, duplicate scope, takedown outcome, recurrence log, access-control record	Prevents deletion from destroying later redress evidence.
Voice-cloning fraud	Call/message metadata, lawful audio capture if available, callback attempt, payment instruction, account/URL identifiers, transaction hold and report record	Separates emotional recognition from verifiable channel and transaction evidence.
AI-generated content label	Generation system, model/tool version, prompt/output or edit record, label decision, machine-readable marker, distribution context	Shows what the label means and what it does not prove.
Detector conclusion	Original file, transformed platform copy, hash suite, detector/tool version, settings, confidence score, analyst notes and limitation statement	Turns detector output into a reproducible, bounded evidence record.
School/workplace incident	Incident intake, victim consent controls, preserved URLs/files/hashes, account/device evidence, parent/HR/safeguarding notices, referral log	Prevents informal deletion or discipline-first handling from destroying usable evidence.
Safeguarding intake	Time of report, reporter role, affected person, immediate risk assessment and responsible lead	Shows who knew what, when, and what immediate protective action was considered.
Harmful content reference	Platform, account, URL, message ID, timestamp and non-public evidence-capture method	Preserves a usable reference without unnecessary redistribution.
Preservation boundary	Who viewed, who captured, where stored, whether content was copied and why	Prevents evidence handling from becoming re-exposure or uncontrolled distribution.
Reporting trail	Platform report ID, police/regulator report where appropriate, school/employer escalation and parent/guardian contact where appropriate	Makes institutional response reviewable after the content is gone.
Support trail	Victim support offered, referral/counselling record, safety planning and contact preferences	Shows the response was not only disciplinary or reputational.
Deletion/removal trail	Takedown request, platform response, duplicate recurrence log and outcome timestamp	Separates deletion from resolution.
Access control	Restricted evidence access list, retention period, destruction/supersession record and review owner	Makes the evidence record defensible and less harmful.

Recommendations by audience

Audience	Actions
Creators	Maintain consent/licence records for authorised voice, face, style, image or likeness replication. Preserve source files, hashes, publication timestamps and takedown correspondence if impersonated.
Businesses	Replace voice-only approval with channel-bound verification and callback logs. Keep incident records linking impersonation, payment requests, account access and decision approvals.
Legal	Ask for notice packets, duplicate-removal records, detector reproducibility and consent/permission lineage. Separate public allegations from preserved records and source boundaries.
Providers	Build victim-safe exportable case histories for deepfake reports. Record duplicate-search actions, moderation decisions and appeal outcomes.
AI teams	Design provenance and disclosure as architecture, not post-hoc labelling. Keep model/tool version, prompt/output, edit, label and distribution records where lawful and proportionate.
Public institutions	Use verified callback and channel-bound identity controls for urgent requests. Preserve impersonation messages, metadata, URLs and credential-redirection evidence.
Education and research	Create trauma-aware school incident workflows for synthetic intimate-image abuse. Preserve evidence without circulating harmful images unnecessarily.
Media and publishing	Treat deepfake claims as evidence packages: source, timing, custody, verification and limits. Avoid unsupported authenticity claims based solely on detector output.

Methodology and limitations

This report uses selected public sources and EviWrite classification. Signals were selected because they reveal an evidence question around source, timing, control, sequence, verification, consent, takedown or proof limits. The register is not intended to be complete.

Signal selection criteria:

- Public source available by or materially relevant to 27 May 2026.
- Clear evidential question around source, timing, control, sequence, verification, consent, takedown or proof limits.
- Relevance to at least one EviWrite audience category.
- Ability to produce a practical Minimum Evidence Record.
- Connection to deepfake abuse, voice cloning, identity impersonation, synthetic media, platform duties, victim remedies or AI/provider governance.
- Classifiable by claim category, source type, region, sector and evidence failure type.

Limitations:

- Selected public signals only.
- Not global incidence.
- Not cyber telemetry.

- Not legal advice.
- Not a liability finding.
- Not a forensic audit.
- Private evidence may exist.
- Absence of public evidence is not evidence of absence.
- Charts are EviWrite classifications.
- Forecast scores are qualitative pressure scores, not probabilities.
- Journalism is used only as bounded public-signal support unless it links to or quotes primary records.

Deep source appendix

Source methodology

Sources were classified by evidential reliability and use. Official and regulator sources ground obligations, timelines and public warnings. Technical standards and research support limits and capabilities. Journalism supports public-signal discovery and case-study context only where official records are unavailable, incomplete, or still developing. Where an official regulator source exists, it is treated as the primary record.

Core repeatable source spine

Source family	Role in this report
Official regulator/government sources	Duties, warnings, policy direction and victim/platform process requirements.
Technical standards and government technical reports	Detection, watermarking, provenance and synthetic-content limitation analysis.
Public safety and crime reporting	Voice cloning, impersonation, AI-enabled fraud and organisational verification pressure.
Academic and specialist research	Detector generalisation limits, transparency architecture gaps and reporting-interface evidence quality.
Public-signal journalism	Cross-border discovery and case-study context with strict limitations; not used as the primary source where an official regulator record exists.

Source group synthesis

Group	Synthesis
Platform takedown and victim reporting	The strongest pattern is procedural: content harm becomes actionable only when notice, timing, matching and outcome records are structured.
Technical detection, labels and provenance limits	Tools are useful leads but become evidence only with reproducibility, custody and limitation records.
Voice cloning and impersonation fraud	The public evidence points away from recognising a voice and toward proving a channel, callback and transaction trail.
Victim, school and image-based abuse response	Institutional response must preserve evidence without forcing victims to re-view, re-send or endlessly re-report abusive content.
Removal-proof conflict	Removal speed and proof durability now collide; the best record proves enough while copying and exposure are minimised.

Source register

The full source register is included in the YAML front matter of this report. Each source card records publisher, URL, accessed date, source type, reliability tier, evidential relevance, supported uses, non-uses and limitations.

Source limitations

No cited source is used to prove private evidence absence. No public signal is treated as a complete factual record. No selected-signal count is treated as incidence. Where sources discuss laws or obligations, this report treats them as evidence-pressure context, not legal advice.

Source-to-claim map

Claim	Support level	Main sources	Boundary
Victim takedown now depends on structured notice-and-response evidence.	Strong	FTC TAKE IT DOWN guidance; FTC image-based abuse guidance; Ofcom illegal intimate-image measures	Does not prove any specific platform failed.
Duplicate removal and hash matching make content identity and matching scope central evidence questions.	Strong	FTC stakeholder template; Ofcom hash-matching announcement	Does not measure effectiveness.
Detector outputs are insufficient unless reproducible, bounded and tied to preserved media.	Medium-high	NIST synthetic content report; UK detection report; Deepfake-Eval-2024	Does not prove all detectors fail.
Voice cloning shifts identity assurance from recognition to channel-bound verification.	Strong	FBI IC3 report; FBI impersonation alert; FTC voice-cloning material; European Parliament briefing	Does not prove technical method in every incident.
AI labels and provenance signals do not prove consent, legality, ownership or completeness.	Medium-high	EU AI Act material; NIST; Article 50 research	Does not provide final case law.
Schools and employers need trauma-aware preservation before informal handling destroys useful records.	Medium-high	eSafety school warning; eSafety abuse guidance; FBI alerts	Does not measure all institutional incidents.
Selected signal counts are not global incidence.	Strong	Report methodology and source scopes	Counts are classification outputs only.

Assurance and review status

This report is a public evidential trend report. It is not an audit, legal opinion, forensic report, regulatory finding, assurance engagement, cyber telemetry report or global incidence report.

Control	Status
Source register prepared	Yes
Source-to-claim map prepared	Yes
Selected signal register prepared	Yes
Forecast signal register prepared	Yes
Chart limitations stated	Yes
Forecast limitations stated	Yes
Legal limits stated	Yes
Global incidence boundaries stated	Yes
Body-level source support included	Yes
External review	Not commissioned for this public evidential trend report

Version and audit record

Field	Value
Report number	DIDH-2026-01
Version	1.0
Period covered	Public signals reviewed to 27 May 2026
Prepared by	EviWrite
Status	Draft / publication-ready source
Report hash	Not issued
PDF hash	Not issued
Source register hash	Not issued
Machine-readable register hash	Not applicable
Receipt	Not issued for this version

Glossary

Term	Definition
Takedown proof gap	The gap between experiencing harmful content and being able to prove notice, platform receipt, duplicate scope, response timing and outcome.
Identity verification gap	The failure created when voice, face, account name or authority cue is trusted without an independent channel or evidence record.
Deepfake detector evidence	A detector result packaged with original media, tool version, settings, confidence, analyst notes and explicit limitations.
Minimum Evidence Record	The smallest practical record set needed to make a claim more defensible, portable, interpretable and independently checkable.