Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study #24

Deepfakes and Synthetic Content: International Standards for Labeling, Origin, and Victim Protection

Legal basis, digital evidence, independent verification and response measures

Organization: Observatoire ARGA

English version

21 rue de l'Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

October 23, 2026

Contents

Research Summary....4
1. Subject of the study and limits of applicability....5
2. Terminology: synthetic content, montage and impersonation....6
3. Map of participants and distribution of responsibilities....6
4. Legal architecture of the European Union....7
5. Article 50 AI Act: Machine-readable markings....7
6. Article 50 AI Act: Disclosure of the nature of deepfake....8
7. European Guidelines and Voluntary Code of Practice 2026....8
8. Applicability of obligations to individuals and organizations....9
9. Digital Services Act and appeals to platforms....9
10. Directive (EU) 2024/1385 and intimate deepfakes....10
11. France: Image rights and criminal offences....10
12. UK: Intimate Simulations and Platform Duties....11
13. United States: TAKE IT DOWN Act....11
14. Legal regimes: comparative summary....12
15. Personal data, voice and human identifiability....12
16. Data on special categories and information on criminal prosecution....13
17. Right to privacy, expression of opinions and public interest....13
18. Copyright, related rights and performance rights....13
19. Technical model of threats: from generation to re-release....14
20. Verifiable Provenance: Claims Model and Chain of Trust....14
21. Machine-readable marking and its limits....15
22. Deepfake detectors: what they detect and what they can't....15
23. Basic statistical indicators of independent verification....16
24. Audiophorensics and voice imitation....16
25. Video, photographs and editing of original material....16
26. Origin Errors: Missing and False Cryptographic Signature....17
27. Digital Evidence Chain of Custody....17
28. Timeline, device and external context....18
29. Working with platforms and reporting illegal content....18
30. Preservation of evidence during urgent removal....19
31. Assistance to victims and prevention of secondary trauma....19
32. International cooperation: limits of competence of a private organization....19
33. The role of synthetic materials in international wanted cases....20
34. Sanctions Matches, AML, and False Corporate Statements....20
35. Migration, asylum and false visual evidence....21
36. Forged documents and visual imitations of official correspondence....21
37. Standard language for expert opinions and confidence levels....22
38. Peer review protocol and bias control....22
39. Transparency of the methodology and confidentiality of the tested material....22
40. Management of errors, refutations and the right to object....23
41. Prevention in the work of editorial offices and public organizations....23
42. Compliance of developers of generative systems....23
43. Testing Resilience to Re-Propagation....24
44. Comparison of detectors and control sets....24
45. Risk of false denial and false accusation....25
46. Institutional model of the Observatoire ARGA....25
47. Stages of implementation and quality criteria....25
48. Limitations of the study and directions for further work....26

49. Deceptively authentic recording and strategy of denial of real events....26
50. Information security when working with expert tools....27
51. Political communication, elections and public statements....27
52. Unfair use of markings and abuse of expert status....28
53. Expert errors when using visual heuristics....28
54. Limits of application of probabilistic expert opinion in court....29
55. Parameters for safe public display....29
56. Quality control of translations and transcriptions....29
57. Institutional indicators and the inadmissibility of false statistics....30
58. Financing independent audits and quality control of contracts....30
59. Register of open questions for international consultation....31
60. Conclusion....31
Appendix 1. Synthetic Material Qualification Matrix....32
Appendix 2. Separation of roles in accordance with Article 50 of the AI Act....32
Appendix 3. Matrix of dates and validity of regulatory requirements....33
Appendix 4. Initial Incident Report Card....33
Appendix 5. Minimum digital file passport....34
Appendix 6. Chain of Custody Registry....34
Appendix 7. C2PA and other origin information verification protocol....35
Appendix 8. Machine marking verification certificate....35
Appendix 9. Independent Detector Test Form....36
Appendix 10. Verification of statistics using a hypothetical example....36
Appendix 11. Audio Recording Checklist....37
Appendix 12. Video Recording Checklist....37
Appendix 13. Notification Form under Article 16 of the DSA....38
Appendix 14. Urgent instructions for intimate materials....38
Appendix 15. Comparative list of protection mechanisms in the EU, France, the USA and the UK....39
Appendix 16. Passport of lawful processing of personal data....39
Appendix 17. Control over the publication of the analytical report....40
Appendix 18. Expert Conflict of Interest Matrix....40
Appendix 19. Technical Risk Map for Financial Transactions....41
Appendix 20. Structure of the evidence dossier for international search....41
Appendix 21. Shelter Materials Evaluation Protocol....41
Appendix 22. Control of fake official letters and screenshots....42
Appendix 23. Working protocol for appeal and correction of the report....42
Appendix 24. Conditions of the expert commission and boundaries of the study....43
Appendix 25. ARGA Pilot Program Launch Criteria Matrix....43
Appendix 26. 90-Day Roadmap....44
Appendix 27. Case Study: Voice Order for Funds Transfer....44
Appendix 28. Conditional case: false confession in an international case....45
Appendix 29. Conditional case: compromising intimate imitation....45
Appendix 30. Case Study: Fabricated Official Letter....46
Appendix 31. Scale of evidentiary conclusions....46
Appendix 32. Register of risks and corrective measures....47
Appendix 33. Scientific Review Checklist....47
Appendix 34. List of prohibited methodological abbreviations....47
Appendix 35. Terms for uniform use....48
Sources and regulatory materials....48

Research Summary

Synthetic audiovisual content can no longer be viewed simply as a form of technological counterfeiting. In some circumstances, the generation of an image, voice, or video serves a clear artistic purpose; in others, it creates false evidence, facilitates unlawful persecution, imitates a public statement, or violates a person's right to privacy and intimacy. The legal outcome depends not only on the model used. It is necessary to establish who created the material, who posted it in a specific context, what information accompanied the publication, what damages arose, and what rules apply to the participants in this chain.

As of 23 October 2026, the transparency requirements of Article 50 of Regulation (EU) 2024/1689 are in force in the European Union. These include machine-readable labeling of generative systems products, a duty on professional users to disclose the nature of deepfakes, and separate rules for the publication of text on matters of public interest. The Commission's 2026 Recommendation and voluntary code of practice support the implementation of these obligations but do not replace the law. Independent victim protection regimes are being developed through European and national criminal law, internet platform regulations, personal data protection, and procedural safeguards. [01] [02] [03]

The study shows that there is no single technical feature that can prove the authenticity of any digital material. Cryptographically verifiable provenance, embedded markers, watermarks, and algorithmic detectors answer different questions. A signature can confirm the integrity of a particular publisher's statement, but not the veracity of the event itself. The absence of a signature does not prove generation; a negative detector result does not prove authenticity. Therefore, an independent audit must combine technical research with external verification of the source, distribution context, and legal assessment.

The proposed model for Observatoire ARGA is built on three secure loops. The first is designed for prevention: labeling rules, publication settings, and origin verification. The second covers rapid measures to assist victims and preserve evidence: URL identification, minimal documentation, notification to platforms, and protection of confidential materials. The third is dedicated to the independent analysis of disputed information, including cases where synthetic material ends up in dossiers related to sanctions, international investigations, migration, or financial monitoring. What's needed here is not automatic trust in the detector, but a refutable, reproducible chain of conclusions.

The report contains a legal comparative assessment, a technical model for independent review, standardized procedures, and working forms. The analytical recommendations are proposals of the ARGA Observatory and do not constitute a statement of the association's public or judicial authority. All case studies are explicitly designated as educational. Information on foreign law is current as of October 23, 2026, and requires verification before application in specific proceedings.

Key findings

1. Labeling content, establishing its origin, and establishing the truth of an event require separate evidence. 2. The European requirements of Article 50 of the AI Act are already in force, but have a precise range of addressees and established exceptions; a universal obligation for every citizen to label every image within them should not be imposed. 3. The harm from synthetic content is determined not only by the quality of the imitation but also by the context: an intimate image, a falsified recording of a public figure, and the manipulation of evidence require different measures. 4. The response of platforms and regulators must be swift, but speed alone does not justify the destruction of a single copy of evidence without securely storing the necessary information. 5. An independent organization has the right to publish a methodology and qualified assessments, but does not have the right to issue a technical opinion as a judicial recognition of forgery or establish criminal liability. 6. For ARGA, a two-stage review has the greatest practical value: first, to limit the potential harm, then to ensure independent verification and the right to object.

1. Subject of the study and limits of applicability

For research purposes, a deepfake is defined as visual or audio material artificially created or significantly altered using AI so that it can be mistaken for an authentic recording of a person, object, place, or event. This working analytical definition is broader or narrower than specific legal definitions, depending on the issue. For example, Article 3(60) of the AI Act contains its own elements of similarity to an existing object and false impression of authenticity. This definition cannot be automatically transferred to the elements of a crime in France or another country. [01]

The research focuses on four processes: the production of synthetic content; its publication and redistribution; identification, verification, and evidentiary use; and the prevention of harm and restoration of violated rights. Hybrid materials, in which only part of the original recording is altered, are considered separately: rearranged lines, generated facial expressions, substituted voiceovers, facial synthesis on a real body, and the editing of real video fragments into an inaccurate chronology. In these situations, the "real/fake" dichotomy is often insufficient.

The study does not equate every use of generative media with unlawful conduct. Feature films, dubbing, accessible communication, satire, and archival restoration may have legitimate purposes. However, the declared creative nature does not eliminate legal requirements and does not exempt the use from assessing consent, public context, and possible violations of personal rights. Each case requires a separate analysis of the creator, distributor, recipient, purpose, and nature of the alleged damage.

The ARGA framework is limited to independent analysis, legal coordination, and methodological support. The report does not advocate covert device monitoring, unauthorized acquisition of digital traces, or the collection of intimate images for public access. The objective is to establish conditions under which legally obtained information can be used, verified, and presented to a competent authority.

2. Terminology: synthetic content, montage and impersonation

Synthetic content can be entirely generated by a model or partially modified from the original. Editing goes beyond AI modification: standard video editing and compositing can create the same false context as a generative system. Impersonation refers to the use of a voice, appearance, or ID to create the impression that a message originates from a specific person. Fraud, identity substitution, and demeaning behavior can exist even without a technically sophisticated deepfake.

For reporting purposes, it is advisable to describe the material along four independent axes. First: the type of transformation, such as generation, substitution, superposition, or splicing. Second: the identifiability of a person or object. Third: the communicative function of the publication, such as a joke, advertisement, official announcement, or evidence. Fourth: the legal and actual risk: deception, invasion of privacy, discrimination, blackmail, public disinformation, and erroneous law enforcement. This framework reduces the temptation to regard the technological feature as the sole legal test.

Context is especially important. A viewer might understand that an animated character is fictitious but not recognize a voice recording mimicking an official's order. The same image, published with a prominent warning and then distributed after the warning has been cropped, creates a different informational effect. Therefore, it's important to record not only the file's bytes but also the screen, caption, title, page, time, and method of reposting.

3. Map of participants and distribution of responsibilities

There are at least five possible roles in the technology chain. The developer creates a model or software system; the supplier makes the tool available on the market; the user or operator creates and publishes the material; the platform hosts, indexes, or recommends it; and the reviewer examines the source and circumstances. Roles may overlap within a single organization, but they should not be mechanically mixed, as the applicable responsibilities depend on the actual functions.

In particular, Article 50 of the AI Act distinguishes between the provider and the deployer. The provider of the generative system is responsible for the technical feasibility of marking within established limits, while the professional user bears independent responsibilities to disclose the artificial nature of individual publications. This does not provide grounds for holding the provider responsible for any infringement by the final distributor or, conversely, exempting the final distributor from liability in the presence of a machine-generated marker. [01] [02]

An individual participant is a secondary republisher who can remove the label, change the caption, or present a satirical fragment as documentary. For them, knowledge of the nature of the material, independent editorial role, intent, and applicable law are essential. The conclusion should not record a supposed "single responsibility for deepfake," but rather the specific action of a specific subject and the legal basis for any possible liability.

4. Legal architecture of the European Union

The regulation of synthetic content in the EU is multilayered. The AI Regulation establishes specific transparency obligations regarding the generation and use of certain systems. The Digital Services Act regulates procedures for handling suspected illegal content in the field of intermediary services. The GDPR applies to the processing of personal data, including identifiable individuals, voice data, and other information, provided the criteria for personal data are met. Copyright law, civil and criminal liability complement this framework.

It's a mistake to claim that violating Article 50 labeling requirements automatically renders the work itself criminally unlawful. Similarly, transparent disclosure of synthetic origin does not render legal material that violates regulations on intimate images, privacy, or fraud. Each set of regulations regulates its own subject matter. In practice, this requires a qualification matrix that simultaneously considers transparency, data legitimacy, platform responsiveness, and individual protections.

The regulatory status must be fixed at the date of the review. The Commission's Article 50 Recommendation of 20 July 2026 provides guidance on the interpretation of the law, and the voluntary Code of Practice, published on 10 June 2026, provides practical solutions. Obligations arise from the current Regulation, not from the signing of the voluntary code itself.[01][02][03]

5. Article 50 AI Act: Machine-readable markings

Section 2 of Article 50 stipulates that providers of systems capable of generating synthetic audio, visual, video, and textual materials must ensure that the result is marked in machine-readable format and that its artificial origin or modification is detectable. The law focuses technical solutions on efficiency, compatibility, robustness, and reliability within the limits of technical feasibility, taking into account the type of content, costs, and the state of technology. This obligation is distinct from a visual inscription on a published page.[01]

Specific boundaries are established: auxiliary standard editing and minor changes to the original data or its meaning may be outside the relevant requirement; cases of legally permitted activities to detect, prevent, and investigate crimes are specifically provided for. These boundaries should be assessed strictly on a case-by-case basis, rather than being transformed into a blanket exemption for all corporate editing tools.

The general deadline for application of Article 50 is 2 August 2026. According to the Commission's consolidated submissions, a limited transitional period until 2 December 2026 is provided for certain systems placed on the market before this date, specifically with regard to technical marking and detection under Article 50(2). This does not abolish the entire set of requirements for any systems and does not imply a retroactive obligation to mark every old publication.[02][04]

In practice, the inspector should request from the supplier the exact model version, the method for displaying the marks, the conditions for preserving them during export, information on known

removal methods, and test results. A statement that the system complies with Article 50 without such information is merely a statement, not independent proof of compliance.

6. Article 50 AI Act: Disclosure of the nature of deepfake

Section 4 of Article 50 obligates professional operators who use AI systems to generate or manipulate images, audio, or video that constitute deepfakes to disclose the artificial nature of the material. This disclosure must be clear, distinguishable, and accessible to interested parties no later than the first interaction or presentation of the content. This disclosure is not limited to an invisible digital watermark.[01][02]

For obviously artistic, creative, satirical, fictional, or similar works, the law permits a special method of disclosure that does not interfere with viewing or perception of the work. This exception applies to the form of disclosure, not to a blanket waiver of privacy protections, consent, or other legal requirements. Cases of legally authorized law enforcement activity are specifically provided for. For materials that pose a risk of false attribution to a real person, a reliable approach is a prominent notice directly adjacent to the content, not just within the user agreement.

For text publications intended to inform the public on matters of public interest, special disclosure rules apply, taking into account legal editorial responsibility and human oversight. These cannot be described as a universal obligation to label every text produced using AI. An organization's internal rules may, however, voluntarily establish greater transparency without claiming that they are expressly mandated by Article 50 for each participant.

7. European Guidelines and Voluntary Code of Practice 2026

The Commission published recommendations on the scope of Article 50 on 20 July 2026, and a final voluntary code on the labeling and disclosure of synthetic content on 10 June 2026. These documents provide important technical and organizational guidance. The code has separate sections for providers responsible for machine-readable labeling and operators responsible for disclosing deepfakes and certain public texts.[02][03]

When assessing an organization's integrity, the auditor may compare the actual labeling system with recommended practices, but must distinguish between two conclusions. First, the technical solution implements specific provisions of the code. Second, the legal obligations of a specific recipient in the context of a specific publication have been fulfilled. There is no automatic equality between the two. It also cannot be claimed that adherence to the code replaces oversight by a competent authority or provides immunity from complaints.

The International Observatoire values openness of the methodology, the indication of the version of the software tool being tested, accessibility of the markings for the visually or hearing impaired, and an assessment of whether the information is retained after publication in various channels. The

report should separately identify whether the information is "technically applied," "visible to the average recipient," and "retained during normal reuse."

8. Applicability of obligations to individuals and organizations

The AI Act cannot be mechanically applied to every action a person takes with a personal phone. The Commission's materials distinguish between personal, non-professional use and use in business or professional activities. Professional status is determined by actual functions and circumstances, not just official registration. A corporate creator of an advertising campaign and an employee acting within their authority may have different roles, as can an independent contractor. [02]

This distinction does not mean that the personal creation and distribution of intimate imitation is legal. National criminal laws, civil remedies, platform rules, and personal data regulations may apply, depending on the facts. Therefore, the initial legal review should distinguish between the applicability of Article 50 and the independent question of the illegality of the harmful material.

It's helpful to organize the internal compliance process as follows: identify the organizational user; establish the purpose; document the publication channel; verify the presence of an identifiable real object; assess the risk of false impression of authenticity; verify the presence of a warning to people and a technical label; and check for specific industry legislation. The solution shouldn't consist solely of a formal "AI used" checkbox.

9. Digital Services Act and appeals to platforms

Article 16 of Regulation (EU) 2022/2065 establishes a notification mechanism for hosting providers regarding suspected unlawful content. The notification must be sufficiently precise and substantiated, include an explanation of the unlawfulness, the specific electronic location, and the required information about the complainant, with exceptions provided by law. Upon receipt of a proper notification, the platform must consider it promptly, objectively, and in good faith, and inform them of the decision and available means of appeal. [05]

The DSA does not establish a general rule that any post labeled "deepfake" must be removed within 24 or 48 hours. The response depends on the nature of the violation, the type of service, the applicable law, any special regimes, and the content of the notice. A false or insufficiently specific claim may hinder removal and affect legitimate speech. Therefore, the notice should specify the specific legal provision allegedly violated and what in the material creates the risk.

The reviewer must maintain a distinction between the fact of filing a complaint, acknowledgment of receipt, the platform's decision, the actual unavailability of the URL, and the cessation of further distribution. Removal of the original post is not the same as removal of copies and does not establish legal liability for the author. Trusted flagger status under Article 22 of the DSA is also a specially recognized status; it should not be attributed to ARGA solely on the basis of expert activity.

10. Directive (EU) 2024/1385 and intimate deepfakes

Directive (EU) 2024/1385 obliges Member States to establish criminal law measures against certain forms of violence, including the non-consensual dissemination of intimate material and certain manipulated images that create the impression of sexual activity. Article 5 sets out specific elements of the offence, including the creation or modification of the material, the subsequent making it publicly available, and the threshold of serious harm in specified cases. Not all image editing is considered to constitute such an offence.[06]

According to official information from the Commission, the deadline for transposition of the Directive is set for 14 June 2027. Therefore, as of October 2026, it should be analyzed as a valid EU act, placing obligations on states to implement the relevant measures, and not as a single, fully operational national criminal law in all Member States. The national laws of individual countries may provide earlier and broader protection.

Article 23 provides for measures to expeditiously remove or restrict access to certain illegal materials, including the possibility of mandatory orders from competent authorities. Practical implementation requires national procedures. Therefore, comparative tables must separately indicate the State's obligation, the current remedy available to the applicant, the competent authority, and the potential date of implementation.

11. France: Image rights and criminal offences

In France, Article 226-8 of the Penal Code covers the non-consensual dissemination of a specific edited image or speech of a person, when the edit is not clearly recognizable or its nature is not indicated. Following amendments in 2024, the dissemination of undisclosed algorithmically generated visual or audio content representing the image or speech of a person will be considered as legally binding. The provision contains specific conditions and increased penalties for dissemination via an online service.[07]

A separate article, Article 226-8-1, concerns sexualized montages and algorithmically generated visual or audio material of a sexual nature containing the image or words of a person without consent. This article provides for different sanctions; at the qualification stage, it cannot simply be replaced by the general provision of Article 226-8. The relationship between criminal and civil remedies should be considered on a case-by-case basis: the image, privacy, reputation, harm, and means of stopping distribution. [08]

A practical conclusion for ARGA: a complaint should not be limited to the general formula "the AI Act has been violated." It is necessary to establish consent, publication to third parties, the audience's perception of the material, the nature of the image, and the circumstances of posting. The decision on whether a criminal offense exists is made by the competent authority; an analytical organization can develop a factual chronology and identify possible legal grounds, avoiding a categorical conclusion about the guilt of an unknown person.

12. UK: Intimate Simulations and Platform Duties

In the UK, protection rules against intimate violence have undergone significant changes by 2026. According to the Crown Prosecution Service, new provisions of the Sexual Offences Act 2003 will apply from 6 February 2026, covering the making or requesting the making of certain purported intimate images of an adult without consent or reasonable belief of consent. From 29 June 2026, separate provisions will apply to the production or provision of tools to create such images.[09]

This criminal law framework does not align with European labeling regulations. Criminality may be associated with the creation or targeted use of material, even if it has not yet reached a large audience; however, the specific elements of the regulation, the person, and the date of its implementation require verification. The rules should not be described as if every unauthorized use of a photograph in an amateur collage automatically constitutes such a crime.

In the area of online security, Ofcom announced strengthened measures against illegal intimate images and recommended automated fingerprint matching or equivalent measures subject to compliance with relevant codes. When assessing compliance, it is important to distinguish between the legal obligation of a specific service, the recommendations of the code, the effective date of the relevant measure, and the technical method of implementation. The absence of a specific commercial hash comparison product does not automatically prove a violation if another equally effective system is acceptable. [10]

13. United States: TAKE IT DOWN Act

The US federal law TAKE IT DOWN Act was passed in May 2025. Effective May 19, 2026, Section 3 requires certain online platforms to provide a mechanism for reporting and, upon appropriate notice, remove covered intimate images and known identical copies within 48 hours. The law applies to the "digital forgeries" it provides, as well as real-life images. The Federal Trade Commission claims oversight of the platforms' respective obligations. [11]

This rule's scope is limited by the subject matter, the addressee, and the legal conditions. The 48-hour period does not apply to any political video, satire, synthetic voice, or publication in any jurisdiction. To use this mechanism, it is necessary to determine whether the service is a covered platform, whether the material meets the definition, whether the applicant's request is admissible, and whether valid notice has been received. Criminal law elements and platform obligations must be considered separately, not lumped together in a single procedure.

When working internationally, ARGA recommends collecting a minimum package for the platform: a link to the specific material, a description of the identifiable individual and the lack of consent, the image category, the date of discovery, a contact person for a response, and proof of representative authority. Excessive resubmission of the most intimate image increases the risk of dissemination and should not be standard practice.

14. Legal regimes: comparative summary

The comparison reveals three distinct regulatory frameworks. The European AI Act primarily establishes transparency requirements for certain providers and operators. French criminal law applies specific provisions to the distribution of manipulated and sexualized material; British law specifically provides for liability for the creation of certain imitations. The American TAKE IT DOWN Act provides a specific mechanism for removing covered intimate content on certain platforms. It is impossible to derive a general rule mandating the removal of any deepfake from a single regime.

A proper matrix includes for each state not only the title of the law and the penalty, but also the date of entry into force, the protected object, the requirement for intent or consent, the conditions of publicity, the recipient of the obligation, the mechanism of immediate relief, and the available route of review. For a cross-border case, the location of the victim, the location of the content accessed, the hosting service, and the applicable jurisdiction are separately recorded. These connections are significant, but do not in themselves determine jurisdiction or the applicable law.

15. Personal data, voice and human identifiability

The presence of a face, voice, or recognizable biographical detail in synthetic material does not independently answer the question of whether the GDPR applies. The assessment begins with whether the individual can be identified directly or indirectly, what information is processed, who determines the purposes and means of processing, and whether there is a territorial link with European law. Identifiability may be maintained even if the original name is replaced by a pseudonym: a combination of appearance, geography, position, and rare circumstances can point to a specific individual. [12]

Data collection for model creation, uploading a photo or voice to the service, publishing the result, moderating the complaint, and storing the materials for legal proceedings should be separated. Each stage may involve different controllers, processors, purposes, storage periods, and legal grounds. A person's consent to an interview or the publication of an original photo cannot be converted into consent to have their voice synthesized for advertising, compromising videos, or fictitious statements without further review.

For independent research, a data processing table for each case is useful: source, identifiers, basis for access, purpose, recipients, publication restrictions, retention period, and the person responsible for further processing. Such a register is more useful than a declaration that "everything is anonymized." Partial obscuring of the face, a preserved voice, or unique circumstances may still allow identification.

16. Data on special categories and information on criminal prosecution

A synthetic publication may report on a person's health, beliefs, intimate life, or alleged criminal conduct. When processing the relevant personal data, not only the general provisions of Articles 5 and 6 of the GDPR must be assessed, but also the specific provisions of Article 9, and for information on criminal records and offenses, Article 10. The fact that the accusation is fictitious does not eliminate the legal risk: a false statement can also process information relating to an identifiable person. [12]

In materials about INTERPOL and sanctions, the mixing of sources is particularly dangerous. A fake "interrogation tape," a synthesized confession, and a screenshot of a non-existent order create the appearance of institutional origin. Before any use, verification of the official source, details, and source of receipt is required. ARGA should not place a questionable imitation in a public record under the guise of established fact.

A restricted access regime is proposed: minimizing copies, separating text descriptions from confidential media material, tracking each viewing, and individually deciding on destruction or archiving. A separate legal basis, rather than the abstract "in the interests of the investigation," is specified for the transfer of materials to a lawyer, expert, or government agency.

17. Right to privacy, expression of opinions and public interest

Freedom of artistic expression and public criticism are protected, but they do not undermine the rights to privacy, reputation, and personal protection. Verification must consider whether the image is clearly satirical, the person it depicts, whether the publication addresses a matter of public interest, the risk of misleading, and whether there is a less harmful way to convey the message. The presence of the "parody" caption may reduce the risk of deception, but does not in itself neutralize potential humiliation or disclosure of sensitive information.

On the other hand, it's impossible to label every piece of material that's inconvenient for a public figure as a deepfake. If the only evidence of forgery is a statement from an interested party, the researcher must document the conflicting testimony and the limits of the available technical data. An independent assessment requires symmetrical standards, regardless of the political, commercial, or procedural status of the participants.

Observatoire recommends three separate categories for publications: the technical status of the material; the factual accuracy of the statements; and the legal and ethical assessment of dissemination. These categories may diverge: a synthetic video may faithfully recount an event, while an authentic recording may be misleading due to context being cut.

18. Copyright, related rights and performance rights

The use of vocal and visual images affects more than just the protection of personality. The original soundtrack, photograph, stage performance, character, and montage may be protected by different

rights. A license to use a musical fragment does not constitute permission to imitate the performer's voice; a photographer's permission to publish a portrait does not constitute the depicted person's consent to commercial impersonation.

The expert's request must identify the author of the original work, the holder of exclusive rights, the rights to the image and voice, the terms of the contracts with the model or artist, the license area and term, permitted modifications, and advertising restrictions. The use of materials for model training and the creation of a specific result are analyzed separately: a legal assessment of one stage does not automatically provide an answer to the other.

A documented chain of custody is sufficient for an audit. Without a signed agreement, it's impossible to confidently assert a violation simply because a copy is unobtainable. It's important to specify what evidence is required, who is responsible for providing it, and what remains unproven. Similarly, the label "AI-generated" does not constitute a license for someone else's voice.

19. Technical model of threats: from generation to re-release

A threat scenario should be described as a chain of actions. First, the attacker obtains the original voice or image; then creates a modification; then selects a distribution channel, impersonates a trusted sender, removes markers, and places the fragment in a plausible context. Harm can occur at any stage, and technical protection at one stage does not guarantee protection at the others.

Particularly dangerous are audio messages requesting urgent transfers, fake video messages from officials, and "scans" of documents with artificially altered details. However, an investigation must prove more than just voice and facial similarity. It's also necessary to verify the delivery route, account ownership, file version history, payment instructions, and the presence of an independent confirmation channel. Sometimes, the key element of the scheme is the compromise of a regular email account, not neural network technology.

The ARGA threat model should include three levels: false information about a client, impersonation of an organization's representative, and compromise of evidence in external proceedings. For each level, available preventative measures, response channels, and the acceptable level of public disclosure are specified.

20. Verifiable Provenance: Claims Model and Chain of Trust

The provenance of a digital file is a sequence of statements about its creation and modification, supported by evidence. A distinction must be made between a device record, an editor's statement, a publication's signature, and a subsequent expert certification. Cryptographic verification can confirm that a given statement is signed with the appropriate key and is associated with the bytes of the file. It does not automatically establish that the signatory was an eyewitness or was not mistaken. [14] [15]

C2PA offers a manifest infrastructure that includes assertions, content associations, and electronic signatures. The verifier evaluates the manifest's validity, the signer's credibility, possible discontinuities in the manifest's history, the use of external components, and the consequences of subsequent editing. The presence of a valid manifest can be strong evidence of a particular processing step, but it does not seal the image's truth.

For forensic evaluation, separate cryptographic and contextual verification protocols should be maintained. The former answers the question of what has been technically verified. The latter establishes the source of the recording, by whom and why it was distributed, whether there are eyewitnesses, and whether independent evidence can confirm the event. These procedures are not interchangeable.

21. Machine-readable marking and its limits

Machine-readable markings aid automated detection of generation, but they do not guarantee error-free algorithms. Markers can be removed during cropping, conversion, recompression, or distribution of screen recordings; sometimes they are missing because the system was created before the required time or because the tool did not support the mechanism. Therefore, the absence of a marker cannot be considered sufficient evidence of natural origin.

It's important to check which specific methods are used: a cryptographic manifest, an invisible digital watermark, a text warning, file metadata, or a platform adding its own label. These methods vary in their resistance to transformations, applicability, and ability to be independently verified. According to NIST, there is no universal solution for all formats and threats. [13]

The recommended report describes the specific transformations under which the marker remains detectable and the limits of the experimental sample. The phrase "the watermark cannot be removed" without testing, algorithm version, and threat conditions is unacceptably categorical.

22. Deepfake detectors: what they detect and what they can't

A synthetic media classifier draws its inference from image or audio signal features, or statistical properties of the generation. Its output is meaningful only in the context of the model version, training set, test distribution, and chosen threshold. A confidence score of 0.92 by itself does not mean a 92 percent probability of forgery: it is often an uncalibrated estimate from the model.

A comparative test must include the baseline counterfeit rate, error matrix, sensitivity and specificity, false positive rate, confidence intervals, image quality, language, and voice characteristics. For the open internet and specially prepared incriminating materials, these distributions can differ significantly. Laboratory dataset results cannot be transferred to forensic analysis without validation.

If the instruments don't match, the honest answer might be "insufficient data." This status indicates the need for further investigation, not the expert's failure. NIST recommends evaluating detectors in conjunction with their provenance, labeling, and procedural safeguards. [13]

23. Basic statistical indicators of independent verification

For the control sample, true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) results are recorded. Sensitivity is TP/(TP+FN), specificity is TN/(TN+FP), and positive predictive accuracy is TP/(TP+FP). The latter indicator is particularly dependent on the prevalence of synthetic materials in the sample being tested.

A hypothetical example: out of 1,000 materials, 50 are synthetic. The detector correctly identifies 45, misses 5, and incorrectly identifies 95 of the 950 genuine materials as synthetic. Sensitivity is 90%, specificity is 90%, but out of 140 positive detections, only 45 are genuine counterfeits, or approximately 32.1%. Therefore, the claim of "90% accuracy" can be misleading when making decisions about a specific person.

Verification must distinguish between detecting synthetic content and establishing malicious use. A fake humorous video and a genuine recording of illegal behavior may fall into different categories. A subsequent legal decision cannot be automatically inferred from the result of a binary classifier.

24. Audiophorensics and voice imitation

Voice substitution is assessed based on several factors: the original audio recording, speech structure, encoding parameters, editing time signatures, and external circumstances of the conversation. Acoustic similarity to a known person does not confirm authorship. Even biometric speaker identification systems can be misleading due to a modified microphone, noise, illness, strong emotional arousal, or deliberate imitation.

It is necessary to obtain the original audio file without re-encoding, establish the source channel, time, and device information, and compare known samples only with legitimate access. The declaration that "the neural network detected voice cloning" must be accompanied by a description of the method, the probability of errors, and a list of alternative explanations. If only messages from a messenger are available, the incompleteness of the source must be explicitly stated.

For fraud, a practical measure, regardless of expert verification, is to use feedback via an independent, pre-established number, confirm the financial transaction with a second party, and pause before executing an atypical instruction. This measure reduces the risk even in cases where counterfeiting technology cannot be detected.

25. Video, photographs and editing of original material

Video footage should be analyzed at the levels of its container, frame sequence, audio track, metadata, and publication description. Anomalous facial expressions, lighting, or lip movements

can be useful indicators for further verification, but the absence of anomalies does not prove authenticity, and the presence of artifacts does not rule out normal editing, compression, or camera manipulation.

It's important to distinguish between completely synthetic video, face swapping in genuine footage, editing genuine scenes, and false context without any image modifications. The latter case demonstrates the limits of a purely technical detector: a file can be genuine, but the caption about the location, time, or participant is patently false.

The video report must contain a description of the original copy, the time positions of the examined fragments, the compared versions, the recoding properties, the verification conditions, and a list of inconsistencies. Identification of a person by appearance cannot be replaced by the formula "the resemblance to a photograph is sufficiently high" without an independent method.

26. Origin Errors: Missing and False Cryptographic Signature

A signature may be missing for objective reasons: the device doesn't create one, the platform removes metadata, the material is extracted from a screenshot, or the archive doesn't support the format. In such cases, it should be stated that "cryptographic origin not established." The conclusion that "the file was definitely created by AI" would be unfounded.

The reverse situation is no less important. A file signed in good faith by the manufacturer may contain a deliberately staged image, an invalid signature, or a generation revealed by the author themselves. A compromised key, an incorrect time setting, or a dishonest signer require verification of the certificate's identity and chain of trust. C2PA does not replace the procedure for verifying the authenticity of the claim. [14] [15]

A separate protocol field, "what the signature covers," is proposed. It should record a specific object: file, version, set of assertions, known processing stage, and trusted participant. A broader conclusion is only permissible if there is independent evidence.

27. Digital Evidence Chain of Custody

For evidentiary use, not only the technical discovery is important, but also the ability to demonstrate what exactly was examined. The protocol begins with the legal acquisition of the file and the immediate calculation of a cryptographic hash. Next, the name, size, format, date of acquisition, source address, person who copied it, method of transfer, and each subsequent action are recorded.

The original is stored separately in a restricted access mode. Analysis is performed on the working copy, and any discrepancies are fixed without "repairing" the original file. Screenshots, printouts, and transcripts are considered derivative materials; they do not replace audio or video data when the original is available.

If the material is obtained only as a URL and quickly deleted, saved pages, platform messages, and timestamps may preserve some of the actual data. However, the inability to obtain the original bytes should weaken the categorical nature of the expert's conclusion. Documenting the chain of custody does not automatically make a copy admissible as evidence in any jurisdiction: the competent authority conducts the procedural assessment.

28. Timeline, device and external context

File creation time, last modification time, server publication date, the local time of the victim, and the archive time often differ. Metadata can be overwritten by software, a device with an incorrect clock, or the platform. Therefore, temporal reconstruction requires specifying the source of each time stamp and the time zone.

Comparing independent events can significantly increase or decrease the credibility of a given story. When verifying a supposed public appearance, the official event program, attendance records, other records, journalistic reports, and objective evidence of the location are used. However, the inaccessibility of some archives does not prove the event did not occur.

The formation of a "unified timeline" is permitted only if the level of reliability is specified. If the sequence remains hypothetical, this is clearly stated. This discipline is especially important in extradition, sanctions, and immigration cases, where dating errors can affect the evaluation of procedural actions.

29. Working with platforms and reporting illegal content

Before contacting a platform, it is important to determine whether the DSA, a specific national mechanism, or only the service's rules apply. For notifications under Article 16 of the DSA, a precise URL, a description of the reason for the alleged unlawfulness, accessible contact information, and a bona fide confirmation from the complainant are helpful. The platform will process notifications according to its established procedure; this does not automatically mean the removal of any controversial political or artistic content. [05]

Mechanisms for urgent removal of certain intimate images, as provided for by national law, must be implemented separately. The US 48-hour deadline is tied to the categories of material and platforms covered, not to all deepfakes worldwide. For the UK, the services' own obligations and Ofcom codes are taken into account. [10] [11]

ARGA's internal registry must record the date and method of notification, the material identifiers, the confirmation received, the platform's decision, the reasoning for the refusal, the appeal process, and the risk of re-upload. Unnecessarily forwarding original intimate files to multiple recipients violates the harm minimization principle.

30. Preservation of evidence during urgent removal

There is a seeming contradiction between the requirement to quickly stop distribution and the need to preserve evidence. This is resolved by separating public access from the secure storage of a minimally sufficient copy by an authorized person. Removing a publication from the platform does not require the destruction of all information legally preserved for legal recourse, provided there is adequate grounds and protective measures.

However, the universal rule of "always download the original file" is unsafe, especially if it contains content involving children, unknown illegal material, or technical access restrictions. In such cases, it is necessary to use official methods to report the content to the relevant authorities, avoid reproducing the image in public chats, and avoid attempting to independently distribute dangerous content "for the record." The amount of storage allowed is assessed on a case-by-case basis.

The protocol recommends three levels of retention: URL and external description only, screen recording of accessible context without dissemination, and secure retrieval of the original by the person entitled to it. The choice of level and the rationale are recorded in the file before the materials are transferred.

31. Assistance to victims and prevention of secondary trauma

Anyone who discovers an intimate or degrading imitation should not be asked to repeatedly view the image, email it to all consultants, or publish a retraction along with the original. Intervention should focus on stopping distribution, establishing a secure communication channel, documenting the necessary information, and allowing the individual to independently choose a legal remedy.

The practical algorithm distinguishes between an initial response within the next few hours, the collection of a limited set of evidence, notifications to platforms, and subsequent reputation restoration. In urgent situations involving threats, extortion, or minors, competent services and local, secure reporting channels play a primary role. An analytics organization should not promise the impossible "complete removal from the internet."

Correspondence with the victim should not turn a report of suspected forgery into an investigation into their personal life. Personal circumstances unrelated to the complaint are not included in the file; public publication requires a separate review of the necessity and legality of identification.

32. International cooperation: limits of competence of a private organization

In practice, the same file may be located on servers in several countries, while the victim, creator, and platform may be subject to different legal systems. International coordination begins not with choosing the "most stringent" criminal statute, but with documenting the locations of the actions, the availability of the material, the location of the participants, and the legal resources of each state.

The independent association can analyze evidence and prepare appeals, but it does not have the authority to demand disclosure of classified information from platforms or law enforcement agencies outside of legally established procedures. In correspondence, a request for voluntary disclosure of data should be separated from an official request from a competent authority. It is unacceptable to present expert advice as an international judicial decision.

Minimizing the cross-border transfer of personal data is especially important. Before sending media materials to third parties, it is necessary to determine the legal basis, the recipient country, applicable restrictions, and more secure alternatives, such as secure access to individual fragments rather than copying the entire archive.

33. The role of synthetic materials in international wanted cases

A potentially fabricated recording can be used as indirect evidence of a contact, confession, public statement, or threat. In the INTERPOL case, it is important to distinguish the technical authenticity of the recording from the legal issue of the correctness of the data processing. The presence of a fabricated recording does not automatically prove the entire international recording is unlawful; its significance for initiating or supporting a specific procedure must be established.

During the analysis, legally accessible materials should be requested regarding the origin of the disputed recording, its role in the national case, the time of its creation, the method of attribution, and subsequent refutations. Depending on the procedure, evidence is submitted to the competent authority or the designated data verification mechanism. The determination of "forgery established" is only possible with a sufficient body of independent information.

ARGA recommends preparing two interconnected but distinct parts of the dossier: a technical part, dedicated to the limitations and characteristics of the file's processing, and a legal part, dedicated to the significance of this material for specific data processing and the applicant's rights. This approach reduces the risk of confusing the expert assessment methods and the powers of the international organization.

34. Sanctions Matches, AML, and False Corporate Statements

A synthesized video of a company executive, a voice-activated transfer order, or a fictitious admission of affiliation with a sanctioned entity could be subject to a counterparty's commercial due diligence. A financial institution that discovers such material must assess the origin of the information, the quality of the external source, and its connection to the legal obligation being investigated, rather than automatically consider the publication as evidence.

AML procedures are characterized by risk asymmetry: hastily trusting a fake can lead to unjustified blocking or denial of service, but hastily declaring a genuine record a "deepfake" can conceal the real risk. A proper decision requires confirmation through independent channels, documentation of the reasons for the decision, and a defined review process.

The proposed form includes the identifier of the disputed material, an external link, an event, a specific risk, confirmations from official registries, the result of the origin verification, unknown facts, and the designated responsible party. The report does not include deliberately inaccessible information masquerading as "confirmed by the security service."

35. Migration, asylum and false visual evidence

Falsified images can both groundlessly discredit the complainant and be used by them to support their prosecution. The examination should not assume deception on the part of either party. The lack of original files, poor connection quality, device limitations, and the impossibility of safely obtaining additional samples must be taken into account.

When requesting additional materials, the agency or consultant must assess the risk to the applicant and third parties. A request for access to personal accounts, devices, or private images cannot be automatically justified by doubts about a single photograph. Alternative sources may include official documents, publicly available records of events, and confirmation from independent witnesses.

In the conclusion, formulas such as "the image is not technically confirmed, therefore the event did not occur" are unacceptable. A technical conclusion and an overall assessment of the reliability of the testimony remain distinct tasks. The report must specify what fact the recording is intended to confirm and what evidentiary deficiency has been identified.

36. Forged documents and visual imitations of official correspondence

Some deepfakes are not videos of faces, but rather synthesized forms, screenshots of letters, regulatory decisions, and images of seals. The file may not contain traditional face-swapping features, but still create the false impression of an official order. Therefore, the methodology must cover documents, interfaces, and synthetic text within images.

The initial verification establishes the document number, date, addressee, signature, organizational details, and conformity to the document format. Next, an independent verification channel is sought: an official registry, an authorized sender, a secure portal, or a certified copy. Spelling errors alone cannot be used to draw conclusions: genuine documents may also contain typos, and a counterfeit may reproduce the details exactly.

If the document is related to a procedural deadline, the original file and its receipt should be saved separately, not just a forwarded screenshot. For a prompt legal response, it is sufficient to mark the document as "unverified" and request an audit; accusing a specific individual of falsification requires other evidence.

37. Standard language for expert opinions and confidence levels

A standard scale of wording is useful for independent reporting. The most definitive level is "supported by the cited primary source," when the assertion being tested follows directly from the recorded material. Next comes "highly consistent," when several independent features support the conclusion, but alternatives remain. Then come "possible hypothesis," "insufficient data," and "not supported."

Terms must be defined operationally. "Not confirmed" is not the same as "refuted," and "signs of editing" do not always indicate AI generation. For each conclusion, the tools used, versions, analysis conditions, number of files examined, and reasons for doubt must be specified. The opposing party must have a reasonable opportunity to point to alternative sources or significant errors in the method.

It is organizationally important to separate expert opinion from ARGA's public position on the matter. If the organization provides paid services to an interested party, this fact is disclosed to the extent necessary to assess the independence of the research, without disclosing the client's protected data.

38. Peer review protocol and bias control

The reviewer must obtain a description of the research question and a minimally sufficient data set, without knowing, to the extent possible, the author's desired outcome. If full blinding is not possible, conflicts of interest are recorded, and disagreements with the conclusions are recorded in a written appendix. Rewriting a study simply to achieve consensus is prohibited.

The review is conducted in five areas: completeness of source accounting; accuracy of preservation procedures; validity of technical tests; distinction between established facts and assumptions; and compliance of legal conclusions with the applicable regime. If a serious defect is discovered, a re-examination of the specific section is conducted, rather than an unfounded declaration of the entire material as false.

The final protocol reflects the comment, its significance, the author's decision, the accepted and rejected changes, the reasons, the date, and the version number. This revision history is important when the study is later cited in court or in a public dispute.

39. Transparency of the methodology and confidentiality of the tested material

Independence requires the publication of general evaluation criteria, but does not require the release of confidential files, personal contacts, and all internal data. A distinction must be made between the reproducibility of calculations using an authorized control set and the uncontrolled dissemination of private evidence. Publication of the algorithm, test parameters, and anonymized statistics, with controlled access to the source materials for an authorized expert, is sometimes sufficient.

The published methodology must explain the thresholds, detector limitations, and error correction procedures. If the model or tool is unavailable for independent replication, this circumstance affects the evidentiary value of the result and must be publicly disclosed.

A two-tiered dossier is recommended: a closed evidentiary appendix and a public analytical version. Information is transferred between tiers only after a separate assessment of the necessity and appropriateness of publication.

40. Management of errors, refutations and the right to object

Publishing an expert report can cause additional harm if it falsely attributes to a person the synthesis of material, involvement in fraud, or a false statement. Therefore, the Observatoire must provide a correction mechanism to quickly flag disputed passages, preserve the original version for history, and issue a verifiable clarification.

The applicant must know where to submit an objection, what materials are acceptable to submit, who reviews the request, and how conflicts of interest are resolved. A forced correction should not be disguised by an inconspicuous file replacement: the public version must contain a history of significant changes without disclosing confidential information.

In urgent cases, a time stamp stating "conclusion disputed, re-examination in progress" is acceptable when new information potentially changes the decision. This is better than an unsubstantiated denial of error or the immediate deletion of the analytical archive.

41. Prevention in the work of editorial offices and public organizations

Editorial boards, press offices, and research organizations should implement a simple standard for verifying unusual audio and video reports. High-risk materials require independent contact with the source, verification of the official channel, documentation of the source, and approval by an editor not involved in the report's preparation.

Automatic labeling helps, but doesn't replace human verification. An image may be captioned and presented out of context; a compressed live video may not have the original available. When publishing, known technical limitations should be clearly stated rather than pretending to be 100% certain.

The response plan includes a designated person responsible for receiving reports of fakes, a secure archive, a public clarification template, and a procedure for interacting with victims. In the case of recurring incidents, it's more useful to revisit the editorial decision-making process than to search for a new detector each time.

42. Compliance of developers of generative systems

Compliance with Article 50 of the AI Act requires the provider of relevant systems to ensure machine-readable labeling and technical detectability of synthetic content within the specified

limits. Documentation must describe the scope, supported formats, exceptions, resilience test results, and residual risks. These obligations should not be confused with the separate obligation of professional users to disclose the nature of a specific deepfake. [01] [02]

The supplier verification plan should record the product version and the date of market entry. For some systems deployed before 2 August 2026, the transitional period for compliance with the obligations under Article 50(2) has been extended to 2 December 2026; this does not constitute a general deferment of all provisions of Article 50. [04]

Having a voluntary code and certificate can be a useful element of compliance, but it does not relieve the need to verify actual compliance with legally binding obligations. Contract terms may require logging of labeling policy changes, notification of critical defects, and a separate procedure for communicating with affected parties.

43. Testing Resilience to Re-Propagation

Publishing synthetic material is rarely limited to a single file. Users trim fragments, add captions, change playback speed, record the screen with a phone, or upload the video to a different format. This means that detection and labeling should be checked on typical derivative copies, not just the high-quality original.

The test set is constructed from pre-cleared control files and pre-determined transformations. For each, the original hash, operation, encoding settings, new checksum, original token, and verification result are recorded. Confidential materials of the victim may not be used in public demonstrations without legal justification.

It's best to show results by transformation class, using absolute numbers: how many out of 100 tests the method passed with cropping, overcompression, noise, and screen recording. An average percentage without breakdowns can hide a critical weakness specifically on a common platform.

44. Comparison of detectors and control sets

A fair comparison of algorithms requires an identical, pre-frozen sample. You can't just review the results, then replace the difficult files with more manageable ones and publish the new percentage as an objective assessment. It's necessary to separate training, setup, and final test data, eliminate duplicates, and document any inevitable overlaps.

The test should represent real-world usage conditions: short videos, different languages, low quality, screen recordings, traditional editing, noisy audio. Groups for which the tool makes more errors are considered separately. If the size of such a group is too small, the correct conclusion is an underestimation, not an unfounded assertion of equal accuracy.

The threshold chosen depends on the cost of error. For automatic blocking of a public post, a false positive may restrict freedom of expression; for a signal requiring manual review, the acceptable

threshold may be different. The methodology should link each threshold to specific use and the appeals process.

45. Risk of false denial and false accusation

Two opposing biases are equally destructive to institutional trust. Fake material can be accepted as authentic, while authentic recordings can be dismissed as neural network-based. In the latter case, the so-called evidence denial effect occurs: the very existence of generative tools is used as a universal excuse.

To avoid this, an organization should not declare material "fake" simply because of its political inconvenience, questionable aesthetics, or lack of metadata. A reliable assessment combines technical expertise, file provenance, external evidence, and contextual analysis. Alternative explanations must be presented, especially when making conclusions with potentially serious consequences.

An internal decision to restrict the distribution of a controversial file can be made with a lower level of certainty than a public accusation of fabrication against a specific author. Different standards of response should be defined in advance and not replaced by a single, vague threshold.

46. Institutional model of the Observatoire ARGA

It is proposed to separate the methodological group, the initial response group, and the editorial board. The first is responsible for technical protocols and the verifiability of tools; the second is responsible for the legal recording of complaints and harm mitigation; the third is responsible for publication decisions, conflicts of interest, and error correction. These functions can be performed by small teams, but decisions with significant consequences require a designated, responsible, and independent second opinion.

A separate issue is the research center's approach to ARGA's fee-based client practice. When providing a client's opinion, the nature of the assignment must be disclosed, the distinction between the attorney's legal position and independent scientific opinion must be made, and a public report must be released only after verifying the admissibility of disclosing the circumstances of the case.

The Observatoire does not claim the right to independently impose sanctions, remove data from international systems, compel platforms, or admissible materials in court. The value of this model lies in its transparent process, the qualification of conclusions, and the ability to critically examine them.

47. Stages of implementation and quality criteria

The first stage is the development of a notification intake procedure, a consent template and legal basis for processing, a competency matrix, and a secure archive. The second stage is a limited pilot using voluntarily provided or publicly available and legally processed materials, with the most

sensitive content not being used without special safeguards. The third stage is an external methodological review and publication of the aggregated results.

Quality criteria should measure not the volume of images collected, but rather the number of cases with a complete origination protocol, the proportion of timely decisions, the time to error correction, the completeness of conflict-of-interest disclosure, and the robustness of detection tools against an independent control set. Any increase in the number of complaints may indicate improved detection rather than increased harm: interpretation requires caution.

Launch is permitted only with a designated data processor, a described legal framework, trained personnel, and a clear process for refusing to review materials that cannot be stored securely. These requirements are set forth as a proposed internal standard, not as a statement of existing ARGA certification.

48. Limitations of the study and directions for further work

This report presents a normative and methodological study, not the results of independent laboratory testing of specific detectors. The numerical examples provided are provisional and intended to explain statistical errors. Claims regarding the actual performance of commercial instruments are not made without independent testing.

The legal provisions were assessed as of October 23, 2026. In this specific case, additional verification of the current version of the regulation, applicable exceptions, territory, and date of effect is necessary. Technical labeling methods, the availability of generative services, and platform practices for content removal are changing particularly rapidly.

Further work is possible in three areas: comparative testing of algorithms on the authorized dataset, independent assessment of the labeling's robustness, and research into the procedural handling of synthetic evidence in different countries. Each requires a separate protocol and monitoring of the research team's commitment.

49. Deceptively authentic recording and strategy of denial of real events

The proliferation of synthetic media creates a second type of societal risk: the ability to dismiss actual recordings without evidence, claiming that "anything can be generated these days." This logic is particularly dangerous in situations where digital video confirms violence, threats, or official misconduct. Researchers cannot assume reasonable doubt simply because simulation tools exist. The objection must have a verifiable basis, not rely on some abstract technological possibility.

The work process begins with a precise formulation of the disputed assertion. It's one thing for the author to deny a specific phrase, another to dispute the location of the filming, and yet another to claim a face swap within seconds. Different tests are applied to each hypothesis. The completeness of the original recording, independent copies, eyewitness accounts, location and time information,

publication details, and technical evidence of manipulation are assessed cumulatively. The absence of one element does not automatically invalidate the others.

Symmetry of standards is essential. If concerns about one party's material are examined using strict rules, similar material from the other party is assessed using the same methodology. An institutional report cannot label any recording unfavorable to the client as a "probable deepfake" without factual evidence and accessible transcripts. This undermines both the credibility of the expert's work and the ability to present materials to the judicial body in good faith.

50. Information security when working with expert tools

Uploading a suspicious file to a third-party online detector may result in the reprocessing of personal data, trade secrets, or criminal proceedings. Before using the service, check the storage conditions, the ability to retrain models on uploads, the processing location, the list of subcontractors, security measures, and deletion procedures. If this information is unavailable, a local tool or analysis that does not require the transfer of the entire file is preferable.

A researcher must distinguish between standalone file viewing and executing active content. A malicious link, an attachment with executable elements, or a specially crafted multimedia container can all impact a workstation. Therefore, work is conducted in a controlled environment with limited access rights, logging, and separate key storage. The risk of digital compromise can exist regardless of whether the video itself is synthetic.

The tool selection protocol includes the product name, version, legal assessment of the upload, processing purposes, deletion availability, cost, and expected response time. For particularly sensitive materials, the decision is made by a responsible specialist in advance, rather than after the file is accidentally uploaded to a public service. The tool's output is stored as a separate result with its own time and parameters.

51. Political communication, elections and public statements

A synthetic message from a candidate, minister, or government agency can mislead an audience within hours, whereas an independent technical verification takes significantly longer. This necessitates a two-speed response: first, clearly state that the message's authenticity has not yet been confirmed, then conduct a documented analysis and update the assessment. Such communication should not preemptively accuse the political opponent of fabricating the material.

The removal of context during republication poses a particular problem. A clip from a satirical show or a reconstruction of a historical speech may be distributed under the caption "hidden camera recording" or "official statement." When analyzing, it is important to preserve both versions of the publication, comparing the audience, time, and the presence of a warning. The legal assessment of the original author and the secondary distributor may differ depending on the role, knowledge, and intent of each.

It would be helpful for the Observatoire to develop a brief public information form with the following elements: the subject of the review; what has been reliably established; what is still unknown; verified primary sources; limitations of the method; and the time of the next revision. For a statement about a public figure, the information needed to understand the risk is sufficient, without publishing the disputed file in full resolution.

52. Unfair use of markings and abuse of expert status

Technical labels can be used dishonestly. A producer can knowingly label a staged recording, while a distributor can add the words "AI-generated" over a genuine video to deceive the audience into believing it is genuine. Therefore, the presence or absence of labels provides information about the publishing process, but is not a definitive test of the authenticity of the content. [14] [15]

Expert products also pose a risk of false credibility: a published chart with precise percentages may mask an unreliable test set or a lack of external validation. It is necessary to verify the funding for the evaluation, the composition of the test objects, the model version, the basis for threshold selection, and any changes to settings after the results are obtained. An independent researcher should not issue a manufacturer's certificate for their own expertise.

In public reports, the results of testing conducted by the organization itself should be distinguished from information obtained from the developer or published in the research literature. When laboratory testing is not available, it should be clearly stated that the recommendations are methodological in nature. The phrase "confirmed by our tests" is unacceptable in this case.

53. Expert errors when using visual heuristics

Training materials often recommend looking for unusual fingers, mirror reflections, lip synching, or lighting errors. These features can be a starting point, but their evidential power quickly diminishes as generative models mature. At the same time, regular video can contain the same artifacts as a result of night shooting, heavy compression, frame interpolation, digital stabilization, or file corruption.

The reviewer must offer alternative explanations for each detected feature. An incorrect number of fingers can appear in a regular photograph due to occlusion, perspective, or editing by an artist. Lip syncing can be caused by a delay in transmission. Metadata indicating an editing program does not indicate criminal fabrication: professional editing is present in regular television production.

To draw a quality conclusion, it's necessary to provide not just a list of "oddities," but also a connection between each feature and the specific hypothesis being tested. If the instrument only examines the visual component, the conclusion should not be extended to the reality of the speech or event. In doubtful cases, a limited formula stating the impossibility of confirming the origin is appropriate.

54. Limits of application of probabilistic expert opinion in court

The model's technical percentage doesn't equate to the legal degree of proof. Even a properly calibrated algorithm evaluates a specific class of data under certain conditions, whereas a court considers the totality of the circumstances of a specific case, the admissibility of evidence, and the adversarial objections. It's impossible to mechanically translate "0.9" into the conclusion "the defendant produced 90 percent of the forgery."

The research report must indicate the assumption being tested and the comparison hypothesis used. If the detector classifies a video as synthetic, it does not establish the author, motive, consent of the person depicted, method of distribution, or harm. Failure to prove the author does not invalidate the results of the technical investigation, but it does limit the scope of the legal opinion.

When preparing material for submission, it is useful to include a brief summary of the instrument, control tests, source material, acquisition procedure, expert limitations, and the dependence of results on data quality. If an independent reviewer is involved, their comments, including disagreements, should be retained. The need for additional procedural forms is determined by local law, not by the term "forensics" itself.

55. Parameters for safe public display

Publishing a technical report may require deepfake examples, but displaying real, compromising images can spread the damage many times over. For training events, specially created demonstrations with participant consent, fictional characters, or approved test cases are preferred. They should be clearly labeled as illustrations to prevent future news coverage without context.

A demonstration should explain both the capabilities and limitations of the technology. When demonstrating a successful detection, it's important to also include examples of false positives, rather than create the impression of absolute reliability. For public presentations, it's advisable to show only truncated or synthetically created training fragments that do not contain real personal data of victims.

If materials are used in international consultations, the difference between a closed evidentiary appendix and a public slide is agreed upon in advance. The presence of a file in the client's archive does not automatically grant the right to display it at a conference or upload it to the broadcast program.

56. Quality control of translations and transcriptions

International deepfake debates often involve translation of speech or text. A translation error can alter the meaning of an alleged confession, threat, or public order, regardless of the technical authenticity of the recording. Therefore, translation and synthetic verification must be conducted using separate protocols, preserving the original language and time stamps.

Automatic transcription can invent words in noisy passages. Speech recognition results should be compared with the audio, unintelligible passages should be marked, speakers should be distinguished, and edited corrections should be saved. Highlighting "confessions" in red within a machine transcript without verification by a professional translator is not sufficient confirmation of the content.

When translating into French or English, it's important to distinguish between "allegedly," "allegedly," "not confirmed," and "refuted." These words serve evidentiary purposes. At the same time, it's important to indicate who performed the translation, whether an automated system was used, and which version of the transcript served as the source.

57. Institutional indicators and the inadmissibility of false statistics

The number of reports of suspicious content does not equal the number of confirmed deepfakes. An increase in complaints may reflect improved reporting, a media campaign, or a change in trust in the organization. Statistics should distinguish between received reports, cases accepted for review, completed technical reports, clarified and retracted findings, and cases where the information was insufficient.

Internal metrics should include response times to potentially dangerous material, the percentage of requests with sufficient source verification, the percentage of procedural errors detected, and the number of timely corrections. An increase in the removal rate should not be presented as evidence of successful criminal prosecution: different stages have different objectives and outcomes.

When publishing aggregated statistics, the risk of indirect identification must be assessed, especially for rare cases with international searches or special data categories. Small groups may require combining categories or refusing to publish the figures. The summary table must include definitions and denominators for all percentages.

58. Financing independent audits and quality control of contracts

The financial model affects the credibility of an expert opinion. If the fee is contingent on the conclusion that the material is falsified, this directly threatens independence. This doesn't mean that any paid research is useless, but it requires a contract with a pre-defined scope, a fixed methodology, transparent pricing, and the expert's right to report a result unfavorable to the client.

Separate conditions govern access to source data, rights to the methodological report, publication of conclusions, destruction of working copies, and the possibility of independent review. The contract should not promise proof of a fact that has not yet been investigated or recognition of the report by a competent authority. Attention is paid to procedural restrictions on confidential correspondence and assignments.

When ARGA jointly participates in advisory advocacy and the public Observatoire, the right to represent the client's position and the status of independent research findings should be

distinguished. Claims of complete organizational independence should be based on the audited management system, not the name of the department.

59. Register of open questions for international consultation

In concluding this study, it's worth formulating questions for which international practice has not yet provided a comprehensive answer. How can we ensure the integrity of machine-generated labels during mass recompression by video platforms? How can mandatory labeling be prevented from becoming an indirect ban on anonymous creative expression? How can a public operator prove compliance with the disclosure requirement if its warnings are systematically removed by republishers?

Evidence-based questions are no less important. What detector test results are permissible for comparison across countries? How should uncertainty be defined for the court? How can the opposing party be ensured access to the methodology without sharing intimate images with third parties? Should the signed manifesto only certify the origin or also the existence of editorial control? These questions should be divided according to the purview of the standard's developers, government regulators, and procedural legislators.

The ARGA Observatory can propose methodological solutions, support public consultations, and collect anonymized examples of errors. However, the publication of recommendations should not present desired future regulations as already in effect. This distinction preserves the usefulness of the research for regulators and protects reader trust.

60. Conclusion

The problem of deepfakes isn't limited to whether an algorithm can detect facial anomalies. It affects the content of legal statements, the origin of the file, trust in digital channels, and the ability of individuals to quickly defend themselves against false publications. Technological labeling is necessary, but not sufficient; removal of harmful material is necessary in appropriate cases, but should not destroy legally preserved evidence; independent expertise is useful, but does not replace courts and regulators.

The most appropriate procedural model for the ARGA Observatory is to minimize harm, retain only necessary evidence, verify provenance through multiple independent means, clearly state the limits of technical methods, and ensure the right to object. A clear distinction between established facts, expert assessments, and legal assumptions will be key to ensuring trust in future research.

It is recommended to approve a standard content qualification matrix, a unified recording and storage protocol, a platform access form, an evidentiary confidence scale, and a review procedure. Practical examples are provided in the appendices to this report and require adaptation to the country, case, and expertise of the contractor.

Appendix 1. Synthetic Material Qualification Matrix

The form is intended for preliminary analysis, not for establishing the elements of an offense. The perpetrator completes the elements independently; empty cells are marked "unknown" rather than "none." The matrix is particularly useful when the source file information is incomplete, as it prevents the question of harm from being confused with the question of origin.

Field	Meaning and filling rule
Object of verification	URL, file, archive ID, date and method of retrieval
The nature of transformation	Full generation, face replacement, voice synthesis, standard editing, not installed
Subjects depicted	Identifiable persons and the basis for such identification; other objects
Context of publication	Official announcement, commerce, satire, private chat, supposed evidence
Notification to the audience	Type of marking, location in the interface, readability, preservation in copies
Harm and the expected norm	Privacy, fraud, security, copyright, risk of wrong decision
Output level	Fact, expert opinion, assumption, no data available

Completion of the form requires the signature of the person who completed the initial qualification and a separate note indicating the need to refer the case to a specialized lawyer. If the publication contains intimate images or data of a child, enhanced access rights apply, regardless of the technical assessment.

Appendix 2. Separation of roles in accordance with Article 50 of the AI Act

This is a checklist for identifying the recipient of the requirements. The contract title does not automatically determine the legal role: an organization can simultaneously provide the tool, use it, and distribute the result. If these roles are combined, the responsibilities are assessed separately.

Function	Legally significant issues
Provider	Who is bringing the generative system to market; is machine tagging and detection supported?
Deployer	Who uses the system under their own responsibility; is it used in professional activities?
Editor/Publisher	Who decides on publication, signature, and warning about artificial content?
Platform	What intermediary functions does the service perform; is there a DSA or other special regime in place?
Private individual	Are purely personal, non-professional activities excluded from the relevant AI Act duty?
Message recipient	What warning does a reasonably attentive recipient actually see and when does he see it?

It's important to check the exceptions for artistic works, criminal procedural contexts, and technical editing separately. Being included in an exception doesn't mean the GDPR, criminal law, or image rights don't apply.

Appendix 3. Matrix of dates and validity of regulatory requirements

The matrix captures the situation as of October 23, 2026; the dates and territorial applicability are re-checked before use in a dispute. The "Meaning" column specifically distinguishes between the commencement of the requirements, the transitional period, and the future obligations of states.

Date	Regime and legal significance
August 2, 2026	Application of the transparency obligations of Article 50 of the AI Act within the established limits
December 2, 2026	Transitional date for Article 50(2) for certain systems already placed on the market; not a general deferment
June 14, 2027	Deadline for transposition of Directive (EU) 2024/1385 for Member States
May 23, 2024	Effect of the revisions of Articles 226-8 and 226-8-1 of the French Criminal Code after the 2024 law
February 6, 2026	The commencement of the CPS offences for making or ordering the making of a false intimate image in England and Wales
June 29, 2026	The commencement of the action of the specified CPS composition on the manufacture or provision of special generators
May 19, 2026	Commencement of enforcement of platform duties under Section 3 of the TAKE IT DOWN Act
September 30, 2026	The introduction of Ofcom's hash matching measures for relevant services

The report does not allow the blanket statement "all deepfakes will be banned throughout the EU from August 2026." Such a statement would confuse labeling requirements, the future transposition of the specific directive, and national statutory liability.

Appendix 4. Initial Incident Report Card

An incoming message is registered without requiring immediate proof of the file's artificial origin. The goal is to determine urgency, maintain a secure channel, and limit unnecessary transmission of sensitive materials. A separate column indicates what should not be collected unnecessarily.

Required field	Rule of thumb
Request number	A unique internal identifier without the full name in the file name.
Communication channel	Secure contact, date, language, agreed upon method of feedback
Detection	Links, time, and description of the threat; do not request a copy of intimate material by default
Urgency	Threat to physical safety, extortion, child, active fraudulent transfer

Required field	Rule of thumb
Primary goal	Removal, refutation, technical verification, preparation of an appeal
Access	Who has the right to view the material, who is responsible for the storage period?
Direction	Lawyer, competent service, platform, technical expert, or termination of consideration

The employee states that the organization can evaluate the material and prepare documents, but does not guarantee the deletion of all copies. If emergency protection is needed, the victim will be referred to appropriate services without delaying assistance for laboratory analysis.

Appendix 5. Minimum digital file passport

A hash certificate links the analysis to a specific object. It is compiled immediately after legal acquisition and should allow another specialist to find the exact version that was analyzed. A hash mismatch requires an investigation into the cause, not a tacit replacement of the object.

Props	What to record
Designation	Source name and internal object identifier
Receipt	Legitimate source, authorized person, channel, date, time zone
Size and format	Bytes, MIME, container, audio and video codecs if available
Checksum	SHA-256 algorithm, meaning and calculation tool
Initial context	URL, page description, signature, information about the author of the application
Derivative versions	Working copies, transformations, individual hashes
Terms of access	Available roles, storage basis, review period

If only a screenshot was obtained, it is described in the passport as just that: a screenshot; this does not replace the original recording. The full hash may be included in a public report only when assessing the risk of matching with sensitive material.

Appendix 6. Chain of Custody Registry

Every action taken on the original or its copy must be recorded. This allows for clarification of whether changes were the result of legitimate processing, routine recoding, or possible tampering. Below is a minimal register; additional fields are required by specific production requirements.

Action	Evidence Record
Reception	From whom, on what basis and through what technical channel
Integrity control	Checking the hash on receipt and on transmission
Copying	Artist, purpose of copy, location of original
Analysis	Tool, version, parameters, obtained result
Transfer to an expert	Reason, protected method, receipt of receipt

Action	Evidence Record
Publication	Permitted redaction, excluded fields, person who approved disclosure
Archiving/destruction	Basis of the term, date, method and confirmation of the responsible person

Changes to the log are corrected with an additional line, preserving the original entry. The log itself should be stored separately from the publicly accessible report.

Appendix 7. C2PA and other origin information verification protocol

The protocol requires not only the presence of the manifest but also the result of a cryptographic verification. If the tool displays a warning, it is copied verbatim into the report within the permitted quotation limits, and the evaluator explains its meaning. The verification is performed on an unmodified copy.

Step	Security Question
Detection	Is the manifest embedded and is related external information available?
Signature	Does the digital signature of this manifest verify successfully?
Chain of trust	Is the signatory known and on what basis is he trusted?
Link to file	Do the stated bindings match the given object?
Story	What changes and original components are confirmed, and where are the gaps?
Context	Is it possible to independently confirm the event described, the author of the statement and the time?
Result	"Manifest Confirmed" is separate from "Event Confirmed"

If C2PA is not available, the mark "origin history not established by this method" is used. It is prohibited to use this mark as a conclusion about generation.

Appendix 8. Machine marking verification certificate

The test must cover the actual working format, not a promotional demo from the manufacturer. The contractor pre-defines the set of transformations and the acceptable detection loss rate. When working with a commercial tool, it is necessary to retain the license terms and access restrictions.

Region	The characteristic being tested
Original marker	Type, version, media, location, reading method
Audio/video	Supported containers, codecs, sampling rates
Sustainability	Compression, editing, cropping, text overlay, noise, screen recording
False positives	Results on the authentic control set
False omissions	Results on the resolved synthetic set
Bypass	Which threats were checked and which were not checked
Conclusion	The line between technical testing and legal obligation

It is preferable to publish the initial numbers of successful and unsuccessful trials and the selection method separately. "Detected in 97% of cases" without a description of the case mix is not a complete characterization.

Appendix 9. Independent Detector Test Form

The test begins with a written statement of the problem. Selecting suspicious files for manual review is one thing, while blocking a publication or generating an expert opinion is quite another. The error requirements vary.

Field	Recording
Method version	Name, Vendor, Version, Date, Control Settings
Dataset	Legal origin, quantity, classes and restrictions
Exceptions	Duplicates, close versions, content outside the test subject
Threshold	Value chosen before final test and rationale
Result	TP, FP, TN, FN with absolute numbers and fractions
Uncertainty	Confidence interval, rare subgroups, tolerability
Restriction of use	Scenarios where the result cannot be used independently

To compare two models, the winning criteria should be defined in advance and changes to the test set should be prohibited after the evaluations have been reviewed. Priority is given to a reproducible experiment, not the number of decimal places in the advertising table.

Appendix 10. Verification of statistics using a hypothetical example

In a training sample of 1,000 units, 50 are synthetic. The detector correctly identifies 45, misses 5, mislabels 95 genuine units, and correctly accepts 855 genuine units. All numbers are arbitrary and do not represent the results of ARGA equipment testing.

Metrics	Calculation and result
Sensitivity	45/(45+5) = 90%
Specificity	855/(855+95) = 90%
Accuracy of a positive forecast	45/(45+95) ≈ 32.1%
Negative prediction accuracy	855/(855+5) ≈ 99.4%
Overall percentage of correct decisions	(45+855)/1000 = 90%
Conclusion	A positive signal requires independent verification, especially if the prevalence of counterfeiting is low.

This example explains why "90% accuracy" doesn't mean nine out of ten suspected materials are truly synthetic. A true conclusion requires the base frequencies of the environment being tested and threshold validation.

Appendix 11. Audio Recording Checklist

For each audio sample examined, the message type, its origin, and the applicable access channel are determined. Before using acoustic means, the possibility of confirming the conversation through direct contact between the parties is assessed. This step is often more useful than complex spectrogram comparisons.

Examination	What should be included in the protocol?
Primary object	Original container, codec, duration, and hash
Source	Who recorded it and when, how was it transmitted, is there a source device?
Editing	Possible splicing, re-coding, insertions
Voice comparison	Legally Obtained Standards, Limitations, and the Risk of Voting Confusion
External verification	Conversation participants, independent contact, call log with legal access
Result	What is confirmed, what is controversial, and what additional materials are needed

Public disclosure of voice samples requires separate verification of necessity and legality. It is prohibited to train a new model on participants' voices without proper justification simply for the sake of demonstrating similarity.

Appendix 12. Video Recording Checklist

The video is examined with the frame sequence and audio track preserved. Particular attention is paid to whether the erroneous conclusion is the result of standard stabilization, HDR, sharpening, or heavy platform compression.

Component	Examination
Container	Parameters, hash, presence of a separate audio track
Footage	Frequency, omissions, repetitions, signs of recoding
Audiovisual communication	Sound matching with articulation, possible channel delay
Light and Geometry	Alternative natural causes of artifacts
Context	Place, date, author of publication, full and truncated versions
External material	Independent recordings of events, programs and testimonies
Conclusion	Separately, the type of installation, source, and reliability of the alleged event

Frames selected as examples are provided with timestamps; one cannot draw a complete conclusion from a few aesthetically questionable frames without a control analysis.

Appendix 13. Notification Form under Article 16 of the DSA

The form is addressed to the hosting service that falls under the applicable regime. It does not replace a specific criminal complaint and does not guarantee that the material will be deemed illegal. The legal basis and exact URLs must be formulated specifically for the specific publication.

Element	Example of content that requires adaptation
Destination	The name of the service and its public notice-and-action mechanism
URL	The full address of the disputed publication and, if necessary, the exact location in the service
Cause	Why specific information is allegedly illegal under applicable law
Context	Who is depicted, what is the lack of consent, what are the consequences and links to confirmation
Applicant	Required contact name/email address, subject to applicable exceptions
Integrity	Confirmation of the validity of the information provided to the applicant's knowledge
Request	Consider the notice, communicate the decision, maintain a legally accessible record of the processing

If the applicant does not wish to disclose their residential address, it is not included simply for the sake of credibility. If the material falls under a special protected category, the appropriate procedure and a minimum of information are used.

Appendix 14. Urgent instructions for intimate materials

This form establishes a procedure to minimize secondary distribution of the image. If a minor, threat, or extortion is involved, a special route to the competent authorities is used; the confidential file is not sent to experts for routine approval.

Stage	Expected action
Contact	Check the secure channel, urgency and authority of the representative
Fixation	Save the necessary URLs and information without redundant duplication
Service	Use the accessible form to report intimate content
Right	Check the applicability of the local removal mechanism and its conditions
Proof	It is legal to retain a minimum stake by an authorized person
Re-posting	Organize follow-up verification of links and known copies
Notification of a person	Explain that removal from one platform does not guarantee the termination of all copies

It is impossible to promise the victim a guaranteed result or demand the provision of unpublished intimate source material if the task is being solved without it.

Appendix 15. Comparative list of protection mechanisms in the EU, France, the USA and the UK

The matrix does not purport to provide an exhaustive overview of judicial remedies in each country. It does show that the regulated action, obligated party, and timeframe vary; the choice of mechanism depends on the circumstances and local law.

Mode	Main subject
AI Act, Art. 50	Transparency and labeling of specific systems and publications; not universal deletion
DSA, Art. 16	A mechanism for reporting suspected illegal content to a specific hosting service
France, Art. 226-8	Certain undisclosed montages and algorithmic simulations without consent
France, Art. 226-8-1	A separate offense for distributing sexualized material without consent
USA, TAKE IT DOWN	Special procedure for removal of covered intimate images upon valid request
England and Wales	Certain compositions for the production or request for the creation of certain intimate imitations
Ofcom / OSA	Responsibilities of relevant services to prevent the distribution of illegal intimate content

Before any appeal, the date, territory, persons involved, and evidence of the facts must be provided. An excerpt from the table does not replace a professional assessment of the case.

Appendix 16. Passport of lawful processing of personal data

Victim protection does not give organizations carte blanche to store any media. In each case, it is necessary to determine the legal source, purpose, scope of access, and deletion deadline. Particular attention is paid to images revealing intimate relationships or information related to criminal charges.

Requirement	Minimum entry
Controller	Who determines the purpose and means of processing
Base	Article 6 of the GDPR or an applicable alternative regime
Additional condition	If required by Article 9/10 of the GDPR or national law
Minimization	Why isn't there enough text, URL, or limited fragment?
Access	Rights, open log, and copy restrictions
Term	Review Event and Deletion/Archive Policy
Broadcast	To whom, to which country and on what grounds

Consent should not be arbitrarily used as a universal basis, especially when other depicted people or data from criminal prosecution authorities are involved in the case.

Appendix 17. Control over the publication of the analytical report

A public document must allow the audience to understand the conclusion without viewing a protected, confidential file. The editor checks whether the wording itself creates reputational harm through over-specification or unnecessary repetition of accusations.

Block	Pre-publication control
Headline	Does not claim proven forgery if there is only a probabilistic conclusion
Method	Tools, versions, sampling and limitations are indicated
Personal data	Information not needed for the public interest is excluded
Conflict of interest	The nature of communication with the customer has been disclosed to the extent permitted
Right to object	Channel and correction procedure available
Applications	The necessity of each image, screenshot and address has been verified
Archive	The revision ID and history of significant corrections have been preserved.

If an independent opinion is not possible without disclosing protected material, it is preferable to prepare a closed version for a competent addressee and a limited public reference.

Appendix 18. Expert Conflict of Interest Matrix

Financial or procedural interests do not always preclude the conduct of a study, but they must be known to the person assessing its evidentiary value. It is especially important to indicate whether the representative of the party in whose favor the question is being formulated is paying for the report.

Risk	Control measure
Payment by the interested party	Fixing the customer and the subject of the order; prohibiting payment for the desired withdrawal
Joint leadership	Independent reviewer and written verification of conclusions
Previous expert position	Declaration of previously published assessments on the disputed file
Access to data by only one party	Marking limitations and a list of unheard alternatives
Commercial relations with the developer	Disclosure of the provider, licenses, and access restrictions to the algorithm
Public campaign	Separation of expert assessment and advocacy communications

The absence of a conflict in the "nothing specified" form is not considered a completed audit. The declaration is signed and stored along with the expert appointment protocol.

Appendix 19. Technical Risk Map for Financial Transactions

In the event of a voice or video request for an unusual payment, technical expertise should not delay prompt protection. The organization operates according to pre-agreed authorization verification rules, without disclosing internal controls to the suspected attacker.

Signal	Action without recognizing the recording as forgery
Unusual urgency	Pause and second independent confirmation
New recipient account	Confirmation via previously known details and contact
Voice command	Verifying identity through a separate channel, not by voice similarity
Copy of corporate document	Reconciliation with the official register and an authorized employee
Suspicious media attribution	Preservation of legally accessible original and metadata
The subsequent dispute	Causal Path Protocol and Internal Control Adjustment Decision

A "proven attempt at fraud" conclusion requires independent facts and a legal assessment; suspicious audio alone is not sufficient.

Appendix 20. Structure of the evidence dossier for international search

The form is intended for careful analysis of a suspected synthetic "confession" or other image that may have been included in a national file. No column automatically constitutes grounds for deletion of the data from the international system.

Part of the dossier	Content
Origin of the material	Source, date, connection with national production
Technical object	Original, hash, derivative files and quality
Attribution	Who declared the recording authentic and on what grounds?
External confirmations	Documents, time, evidence, independent records
Procedural role	What specific conclusions did the body rely on the disputed material?
Alternatives	Unspecified circumstances, possibilities of editing without AI
Request	Specific inspection or measure of the competent authority indicating the applicable rules

If a request is submitted to an INTERPOL mechanism, it must comply with its procedure and the subject of the examination; a private technical opinion does not replace the decision of such a body.

Appendix 21. Shelter Materials Evaluation Protocol

In asylum cases, the quality of mobile media can be low, and additional samples can put the applicant or witness at risk. This affects the reasonableness of verification requirements, but does not justify abandoning the analysis of existing sources.

Question	Principle of evaluation
File source	How it was obtained, why the original may not be available
Context	What exactly does the recording confirm and where are its limits?
Doubts	Technical alternatives and their probabilities are named without categoricality
Additional materials	Is it possible to obtain them safely without putting pressure on third parties?
Conclusion	Don't substitute the evaluation of the entire story for the review of a single media file.
Access	Confidentiality of the applicant and persons in the country of origin

If the analysis only reveals a lack of quality, this is clearly stated. The veracity of the personal narrative and the technical characteristics of the file should be considered separately.

Appendix 22. Control of fake official letters and screenshots

A digital document can be synthesized without the use of a face or voice. For such objects, both the content and the transmission path must be verified. It's important to avoid hasty attribution of a violation to the author of the genuine form.

Field	Examination
Identifiers	Number, date, institution details, links to the procedure
Channel	Domain, electronic signature, official portal, transfer route
Visual signs	Fonts and seals are considered only as auxiliary signals
Direct confirmation	Independent appeal to the competent authority
Deadlines	You cannot miss the procedural deadline while waiting for a technical examination
Conclusion	"Authenticity not confirmed" is separate from "proven fabrication"

The original file is retained until technical editing. Redrawing signatures, correcting fields, or adding details "for readability" is prohibited on the copy being examined.

Appendix 23. Working protocol for appeal and correction of the report

The procedure must be implemented regardless of who commissioned the study. The submitter has the right to report an incorrect file, a new source, a tool error, or a misrepresentation of its position. The responsible party records the incoming message and the extent to which the comment impacts the original conclusion.

Stage	Recording
Registration	Complaint ID, date, associated report version
Base	Error of fact, error of analysis, breach of confidentiality, other

Stage	Recording
Independent viewing	Assigned reviewer, reviewed materials
Temporary measure	If necessary, mark the contested conclusion
Solution	Leave, clarify, correct, or withdraw a specific formulation
Publication	Visible change history without unnecessary personal data
Conclusion	Motivation and direction of response to the applicant

When a significant error occurs, correcting it correctly is more important than maintaining the outward appearance of correctness. Removing criticism without justification is unacceptable.

Appendix 24. Conditions of the expert commission and boundaries of the study

The contract or task statement should not require a desired outcome in advance. If it is impossible to legally obtain the initial data, the expert is obligated to indicate this deficiency; the permissible scope of the conclusion should be reduced accordingly.

Condition	Recommended wording
Item	Research of the specified files and specific hypotheses
Method	The tools used, the possibility of change upon justification
Independence	Prohibition of dependence of reward on the direction of withdrawal
Data	Legal origin, access mode and transfer
Result	A report with facts, alternatives, limitations and appendices
Review	Admissibility of independent methodological verification
Publication	A separate written decision taking into account the rights of third parties

You should not include a promise that the conclusion will be accepted as legal evidence, accepted by an international organization, or will serve as binding grounds for deletion of the information.

Appendix 25. ARGA Pilot Program Launch Criteria Matrix

Before piloting, what's needed isn't a promotional toolkit, but a functioning minimum of governance: who has the right to receive materials, where they are stored, and how unauthorized collection is prevented. Effectiveness is assessed after a small, controlled sample and independent verification.

Criterion	Admission conditions
Mandate	Expert functions are outlined in writing, and claims of authority are excluded
Legality	The grounds for processing and special categories of data are defined
Safety	Access control, encryption and copy restrictions
Methodology	Versioned protocols, thresholds, revision procedure

Criterion	Admission conditions
Footage	Designated staff, training on confidentiality and tool limitations
Publicity	Publication and Error Correction Policy
Audit	External review before public statements about effectiveness

A separate stopping criterion is provided: detection of illegal collection, leakage of sensitive material, or the inability to reliably determine the origin of the control kit.

Appendix 26. 90-Day Roadmap

The timeline is a proposed internal plan and does not confirm that the corresponding activities are already underway. The work schedule prioritizes secure message reception over scaling analysis.

Period	Control result
Days 1–15	The mandate, roles, acceptance and storage policy, and legal basis matrix have been approved
Days 16–30	A secure channel, evidence log, and patch request log have been configured.
Days 31–45	Procedures for controlled storage, labeling and verification of origin have been tested
Days 46–60	A review of anonymized training cases was conducted
Days 61–75	Public guidelines, notification forms, and response scenarios have been prepared.
Days 76–90	Independent risk assessment, decision on pilot scale and limitations

The decision to open a new department must address not only the objectives achieved but also any remaining legal restrictions. Any expedited approach does not justify working with illegally obtained files.

Appendix 27. Case Study: Voice Order for Funds Transfer

Case study: A finance employee receives an audio message with a voice similar to that of a manager; the message instructs a transfer of funds to a new counterparty. The recording quality is average, the source account is unknown, and the attachment is sent multiple times. This is a fictitious situation and does not describe a real ARGA client.

Question	Correct working answer
Is translation necessary?	No, until independent confirmation of powers and details
What to save	Legally accessible file, route of entry, metadata and hash
What to check	Callback to a known contact, contract, new account, list of powers
What can the detector detect?	Only signs of synthetic processing with the indicated restrictions
What is the result?	A separate conclusion on payment risk and a separate conclusion on the origin of the audio

Even if technical analysis fails to identify the generation, internal controls can prevent payment risk. This is what distinguishes a practically useful reaction from an attempt to first prove a specific neural network technology.

Appendix 28. Conditional case: false confession in an international case

Case study: A video appears in the media featuring a person allegedly admitting to involvement in a crime. A government agency cites it in a public comment, but a copy of the original source is unavailable. An interested party claims the footage is a deepfake. The case is fictitious; no conclusions are drawn about the identity of any actual wanted person.

Question	Necessary research
Identity	On what basis is the face in the video identified as a human being?
Source	Who first published this? Is there a full video or media available?
Technology	What signs of change have been identified, what alternatives are possible?
Meaning	Did the recording influence a specific legal decision and how exactly?
Procedure	Which body is authorized to consider the objection and what documents are required?

The conclusion that "the application appears synthetic" does not automatically terminate the international procedure. The technical assessment must be integrated into the legal argumentation based on specific data and evidence.

Appendix 29. Conditional case: compromising intimate imitation

Case study: An adult complainant discovers a synthetic intimate image of herself that has been distributed without consent. Multiple URLs are available, but the original source is unknown. The situation is fictitious and not based on real, private material.

Question	Priority
Safety	Limit contact with the distributor and assess threats or blackmail
Fixation	URL and required context without re-posting the image
Removal	Use the platform's dedicated mechanism and applicable national law
Legality of storage	Set a narrow purpose, access, time limit, and acceptable copy
Additional protection	Consider contacting the competent authorities and supporting the victim
Communication	Do not publish the source in a refutation and do not promise to destroy all copies

Publication of a report on an incident is only possible if the individual independently determines its necessity and confidentiality. Maintaining psychological and physical safety is more important than the media credibility of the report.

Appendix 30. Case Study: Fabricated Official Letter

Case study: An organization receives an image of a letter purportedly from a foreign regulator, demanding the urgent transfer of client documents. Visually, the document matches company letterhead, but the original file and channel confirmation are missing. This example is fictitious.

Question	Action
Assignment	Do not transfer documents until authorization has been reliably verified.
Channel	Independent verification of official contact and register
Expertise	Capture image, metadata and path
Assumption of forgery	Don't make a categorical conclusion based only on the image of the seal
Appeal	If necessary, check the status directly with the competent authority.

The complexity of this case lies not in the presence of a synthetic face, but in the false institutional attribution. Therefore, deepfake analysis methods must encompass documents and interfaces.

Appendix 31. Scale of evidentiary conclusions

The table is intended to standardize language, not to translate technical probability into a legal standard without justification. The categories should be used consistently in both public and private versions.

Category	Meaning
Established technical fact	Directly confirmed by an accessible and stored object
Strong feature agreement	Several independent features support the hypothesis; alternatives are described
Limited support	Some of the data are consistent with the hypothesis, but the sample is incomplete.
Not enough data	The conclusion about origin cannot be made in good faith
Not confirmed	The source provided is not sufficient to make a specific statement.
Refuted	Independent confirmations show the inconsistency of the assertion being verified with the established facts

The first three categories must not be interpreted as judicial decisions on liability. In each case, the subject of the decision and its addressee must be stated.

Appendix 32. Register of risks and corrective measures

The register is updated based on actual incidents and audits. It should help assign risk ownership, not create the appearance of rigorous quantitative science based on subjective assessments. Scoring scales are permitted only under certain criteria.

Risk	Control measure
False accusation of generation	Independent review and mandatory description of alternatives
Confidential image leak	Access control, storage protection, logging
Loss of original file	Immutable copy, hash and legal reservation
Incorrect detector statistics	Independent sample, error matrix, base frequency
Conflict of interest	Declaration of funding and separate editorial decision
Redistribution upon refutation	Publishing controversial media material without reposting
Substitution of the regulator's role	Clear description of mandate and limitations of competence

For each risk, the responsible person, review date, and residual acceptance criteria are added to the working copy. The public version should not disclose internal security measures in a way that facilitates circumvention.

Appendix 33. Scientific Review Checklist

An independent reader confirms or challenges not the overall impression of the report, but rather specific conclusions and evidence. With limited access to the original data, the reviewer indicates which elements they were able to verify and which were derived only from the author's retelling.

Quality Controlled	Reviewer's Result
Sources	Primary regulatory texts and exact dates of implementation
Data collection	Legality, source and reproducible chain of custody
Detectors	Version, validation and transfer limits of results
Statistics	Correctness of absolute numbers, proportions, intervals and samples
Conflict	Verified Disclosure of Interest
Report	The difference between a fact, a technical assessment and a legal proposal
Rights of individuals	Mechanism of objection, correction and publication minimization

The reviewer's comments are retained along with the author's reasoned responses. Disagreements regarding interpretation may be published if they do not disclose protected data.

Appendix 34. List of prohibited methodological abbreviations

Some formulas appear convincing, but fail logical scrutiny. This sheet is intended for the editor and expert as a quality control check before signing off on the report.

Incorrect formula	Correct replacement
"No C2PA means deepfake"	"Manifest of Origin Not Found"
"C2PA confirmed the event"	"The signature of certain file statements has been validated."
"The detector showed 93%, so it's proven."	"Model evaluation requires calibration and independent validation."
"AI Act requires the removal of any deepfake"	"Article 50 introduces certain transparency obligations."
"48 hours to remove in all countries"	"The special period applies within the boundaries of a specific law."
"The file is authentic because it contains no artifacts."	"Available features did not allow editing to be established."
"The file is mounted, which means the event is fictitious."	"It is necessary to evaluate the original event and the meaning of the montage separately."

The list is verified before each external publication; it reduces the risk of arbitrary extension of conclusions to situations not covered by the technical test.

Appendix 35. Terms for uniform use

It is recommended that the working documentation establish a vocabulary and reference it when creating forms. These terms do not replace the legal definitions of the relevant jurisdiction.

Term	Working value
Synthetic content	Material created or substantially transformed by generating means
Deepfake	An imitation of objects or events capable of creating a false impression of authenticity
Impersonation	Presenting content as originating from a specific person
Provenance	Verifiable information about the creation and modification of an object
C2PA manifest	The structure of statements and signatures associated with a specific object
Watermark	Built-in signal designed for subsequent detection
False positive	The genuine object was mistakenly identified as synthetic.
False negative	The synthetic object was mistakenly classified as genuine
Chain of custody	Documented sequence of receipt, storage and transfer

When translating into English or French, it is preferable to keep the card identifiers identical to avoid changes in meaning when processing an international package.

Sources and regulatory materials

[01] Regulation (EU) 2024/1689, last version as at 27 July 2026; Articles 3, 50, 99, 113.Official source

[02] European Commission. Guidelines on transparency obligations, 20 July 2026.Official source

[03] European Commission. Code of Practice on Transparency of AI-generated Content, 2026.Official source

[04] European Commission. FAQ on Article 50 and the transitional period.Official source

[05] Regulation (EU) 2022/2065, Digital Services Act, Articles 16–17 and other provisions.Official source

[06] Directive (EU) 2024/1385 on combating violence against women.Official source

[07] Code penal (France), article 226-8, edition 2024.Official source

[08] Code penal (France), article 226-8-1, edition 2024.Official source

[09] Crown Prosecution Service. Cybercrime prosecution guidance, 2026 updates.Official source

[10] Ofcom. Enforcement program on non-consensual intimate imagery, 9 September 2026.Official source

[11] Federal Trade Commission. Complying with the TAKE IT DOWN Act, 2026.Official source

[12] Regulation (EU) 2016/679 (GDPR), Articles 5, 6, 9, 10, 12, 17 and 21.Official source

[13] NIST AI 100-4. Reducing Risks Posed by Synthetic Content, 2024.Official source

[14] C2PA. Content Credentials Technical Specification, version 2.2.Official source

[15] C2PA. Content Credentials Explainer version 2.2.Official source

Note on sources. The regulatory provisions and methodological recommendations are presented in the versions available as of October 23, 2026. Illustrative scenarios and numerical calculations do not constitute ARGA field studies. In a specific case, verification of the applicable version of the law, the procedural form, and the powers of the addressee is necessary.